Multimedia Systems 7: 32–47 (1999)

Multimedia Systems
© Springer-Verlag 1999

Interactive environments for music and multimedia

Antonio Camurri, Pasqualino Ferrentino

Laboratory of Musical Informatics, DIST – Department of Informatics, Systems and Telecommunications University of Genoa, Viale Causa 13, I-16145 Genoa, Italy; e-mail: music@dist.unige.it, linus@musart.dist.unige.it

Abstract. *Multimodal Environments* (MEs) are systems capable of establishing creative, multimodal user interaction by exhibiting real-time adaptive behaviour. In a typical scenario, one or more users are immersed in an environment allowing them to communicate by means of full-body movement, singing or playing. Users get feedback from the environment in real time in terms of sound, music, visual media, and actuators, i.e. movement of semi-autonomous mobile systems including mobile scenography, on-stage robots behaving as actors or players, possibly equipped with music and multimedia output. MEs are therefore a sort of extension of augmented reality environments. From another viewpoint, an ME can be seen as a sort of prolongation of the human mind and senses. From an artificial intelligence perspective, an ME consists of a population of physical and as software agents capable of changing their reactions and their social interaction over time. For example, a gesture of the user(s) can mean different things in different situations, and can produce changes in the agents populating the ME. The paradigm adopted for movement recognition is that of a *human observer of the dance*, where the focus of attention changes according to the evolution of the dance itself and of the music produced. MEs are therefore agents able to observe the user, extract "gesture gestalts", and change their state, including artificial emotions, over time. MEs open new niches of application, many still to be discovered, including music, dance, theatre, interactive arts, entertainment, interactive exhibitions and museal installations, information atelier, edutainment, training, industrial applications and cognitive rehabilitation (e.g. for autism). The environment can be a theatre, a museum, a discotheque, a school classroom, a rehabilitation centre for patients with a variety of sensory/motor and cognitive impairments, etc. The ME concept generalizes the bio-feedback methods which already have found widespread applications. The paper introduces MEs, then a flexible ME architecture, with a special focus on the modeling of the emotional component of the agents forming an ME. Description of four applications we recently developed, currently used in several real testbeds, conclude the paper.

Correspondence to: A. Camurri

1 Introduction

Current state of the art in multimodal environments (MEs) concerns mainly *virtual environments* (VEs) and *hyper-instruments* (Machover and Chung 1989; Mulder 1994). Hyper-instruments can be considered a particular kind of VE, based on sound instead of visual feedback. Most of the existing VEs and hyper-instruments consists of a sort of static virtual world which can only be navigated. They do not change their structure and behaviour over time and do not adapt themselves to users, they do not try to guess what the user is doing or wants. Exceptions are, for example, ALIVE (Maes et al. 1995) and *Placeholder* (Laurel et al. 1994), where the users can leave traces of their actions, and, in certain aspects, the interactive music system *Cypher* (Rowe 1993). In particular, interactive music systems are receiving a growing interest by composers and artists (Rowe 1993; Povall 1995), whose work contributes to the research on MEs, not necessarily restricted to art and music domains.

Most state-of-the-art systems fall into two main categories. The first consists of real-time systems involving low-level, direct cause-effect mechanisms; the metaphor is the *musical (hyper-) instrument*, where a fixed mapping between human movement and sound or media output is defined. The second category consists basically of tools for the computer animation of virtual characters, for film and computer game industries (Maestri 1995). Here, the goal is to track as precisely as possible the human movement to animate artificial characters in movies, advertising, and computer games. This implies a simple, direct use of movement information like in puppet animation. Further, it is usually an off-line design process, with the exception of television applications, on animation of characters by actors backstage wearing complex sensors and motiontracking systems.

In MEs, we envisage active, dynamic environments as populations of agents whose behaviour can be influenced by users. In the case of systems which map movement and gesture into sound and music, the traditional approach corresponds to static, passive environments like virtual musical instruments or *hyper-instruments* (Machover and Chung 1989). On the other hand, MEs can be modified and possibly re-created by users' behavior. Using the same metaphor,

we have an orchestra to conduct or one or more "agents" to interact with, instead of an instrument to play.

The main goal of MEs is to observe users to establish high-level communication with them. An ME must be able to change its behaviour to adapt to users, and to observe users in their *general, full-body, context-dependent movement and gesture*, in a sort of gestalt approach. To accomplish these goals, it is often not necessary to go into the local details of movement: global features can often be extracted by integrating over time intervals data obtained from a number of medium-/large-grain sensors.

The paradigm we adopted is *a human observer of the dance* (or of the full-body movement), where the focus of attention changes dynamically according to the evolution of the dance and of the music produced. MEs should be able to change their reactions, their social interaction, including their possible artificial emotions and their "rational" state over time. To this aim, MEs embed multilevel representations of different media and modalities, representations of communication metaphors and of analogies to integrate modalities. MEs should be able to decide and apply plans in response to user actions and behavior. This approach is also analogous to existing works in the literature in the field of gestalt-based music perception (see, for example, Leman 1995), which can in principle be applied to MEs.

We can define MEs as an extension of augmented reality environments integrating intelligent features. We therefore envisage an audio-visual environment which can be communicated to other humans and machines, either other actors participating in the same event (including autonomous robots) or external spectators of the action. A typical sample scenario regards an integrated system which is driven, tuned and moulded by the movements and by the sounds produced by the user(s) (actors, dancers, players), using specific metaphors for reaching, grasping, turning, pushing, navigating, playing, communicating their internal state and emoting potential, etc. Users must be free to move without any wire, device, etc., in order to let them free in their specific tasks instead of the technology.

The previous considerations put into evidence the complexity and the multifarious nature of the domain. MEs need to represent and manipulate sound and music knowledge, the knowledge of mobile robots navigating on stage, perceptual and emotional spaces, analogies and metaphors, user interaction models, sensor data processing, multimedia knowledge, etc. We deem that such complexity can be managed by adopting "hybrid" models integrating different representations. Integrated hybrid agents architectures have been demonstrated to be an effective platform for MEs, and have already been experimented with in several cases (Ferguson 1992; Rowe 1993; Riecken 1992). Our work is a contribution in this direction. As a part of a general investigation on multimodal interaction in multimedia systems (the Esprit Basic Research Project 8579 MIAMI - Multimodal Interaction for Advanced Multimedia Interfaces), our research is mainly concerned with the study and development of ME architectures. This modeling effort is synergic with a stream of software developments, previously grouped in the *HARP* (Hybrid Action Representation and Planning) project (Camurri et al. 1991, 1994, 1995) and recently evolved to the current architecture. Some of our recent results, both in the research and in practical effective system applications, are reviewed in this paper. In our system, users can create agents-MEs providing gesture tracking, input feature extraction, interactive music and multimedia performance. Agents-MEs can manage the interpretation of motion and gestures, the overall (music and movement) context in which the movements occur, the current intentions and goals of the user, correlations with past actions, etc.

We began to study and experiment with movement analysis in a dance/music framework in a simulated environment (Camurri et al. 1986): the *MANI* (Music and ANimation Interface) and *Key-music* systems are the first prototype systems we developed to "observe" a computer-generated wire-frame dancing figure, and to transform a *cmusic* score (Moore 1990) according to the movement features detected at different levels of abstraction, taking into account the current context and the past actions and situations. The research and the systems described in this paper have their origin in and inspiration from previous work on the design of systems for the study and integration of dance and music (Camurri et al. 1986; 1993; Otheru and Hashimoto 1992; Ungvary et al. 1992; Schiphorst et al. 1993), on interactive music systems (Rowe 1993; Morita et al. 1991; Vertegaal and Ungvary 1995), on particular integrated-agent architectures for music (Riecken 1992) and multimodal systems (Thorisson 1995; Wexelblat 1994).

In the first part of this paper, we present our proposal of an ME architecture, with a special focus on the modeling of the emotional components in MEs, taking as a leading example an ME based on a robotic platform navigating on stage and with audio output, recently adopted in the permanent museal exhibition "Città dei Bambini" (Genova, Porto Antico), which literally means "City of Children", where it plays the role of Cicerone and partner in interactive games.

Then, three further examples of application of our architecture that are used in public events are presented: (i) the *HARP/Vscope*, for the tracking of human movement by means of on-body, wireless sensors, gesture recognition, and real-time control of computer-generated music and animation; (ii) the *SoundCage Interactive Music Machine* (IMM), a system based on a set of spatial sensors displaced in a sort of "cage", whose design has been focused to track overall, full-body human movement without the need for any on-body device and constraint; (iii) the *DanceWeb*, based on a different proprietary human movement acquisition system using different technologies of sensor systems, which can be used both in stand-alone installations and integrated with the SoundCage IMM.

2 Two sample scenarios

Let us consider two simple but concrete examples of MEs as populations of "living agents" in on-stage interactive performance. The first concerns an ME formed by a single agent, able to extract from a human some gesture and movement features, thereby controlling the generation of sound and music. It is therefore capable of reconstructing "views", and interpret in some way movement and gesture. At the beginning, the agent is a "tabula rasa", nothing is evoked by movement; the system is observing the user. We can imagine that

the agent is trying to identify features of the "style of movement" of the dancer. If the latter starts moving with a given nervous and rhythmic gestures in roughly fixed positions in the space, therefore evoking the gestures of a percussionist, the agent, after a few seconds, initiates a *continuous* transformation toward a sort of "dynamic hyper-instrument": a set of virtual drums located in points of the space where the dancer insists with his/her movement. "Continuous" means, for example, that neutral sounds begin to emerge and transform progressively into drums, e.g. gradually reducing the attack duration and moving from a default to a specific timbre. The number of drums and their spatial position is decided by the movement of the dancer. The drums' timbral and intensity features can be associated with the interpretation of the dancer movements. The dancer is therefore now allowed to play the instrument he/she has built. Instruments not played for a certain period of time may begin to fade away and loose degrees of freedom of interpretation.

If the dancer changes his/her "style" of movement (for example, by gradually reducing the speed of a harsh movement toward smoother gesture), the agent will adapt by continuously changing its behaviour toward a different *context* (again, *continuously* and in a time interval proportional to the amount of change of the dancer's style of movement). This transformation can mean a continuous change, both in the sensitivity/interpretation of gesture as well in the focus of attention, i.e. a change of the set of movements and gestures observed by the agent, as well as a change of the associations with the sound and music output. The agent could be able to generate a music output coherently with the current gestures and movements, the past state of the performance, and the musical goals defined by the composer. In our example, the transformation might change the music output from the set of user-defined virtual drums into a (virtual) string quartet, where the movement controls the interpretation and the melodic contour of the counterpoint. The system generates the counterpoint and is subject to compositional goals (thus incorporating a deliberative behaviour).

The designer of the performance (the director/composer/choreographer) introduces into the system the musical knowledge, the compositional goals, the aspects of integration between music and gesture (including a model of interpretation of gestures), and decides the amount of (possible) degrees of freedom left to the agent as concerns the generative and compositional choices. This also implies the management of an extension of music language with gesture, movement, and possibly visual languages.

This example raises several important issues about new perspectives on the integration of music and movement languages for composers, choreographers, and performers, as, for example, in dance-driven *live electronics*. Real applications based on these ideas have been developed and are currently being experimented with by various artists. We implemented and used in public events MEs with a behaviour similar to the example described above.

Let us consider a further example, a sort of extension of the previous one. The agents involved now are the following: (i) an agent similar to the previous one, able to extract from human(s) some gesture and movement features, thereby controlling the generation of sound and music; (ii) an agent based on a robotic platform on wheels capable of navigating semi-autonomously on stage, with embedded sensors and equipment for wireless communication and high-quality audio output.

Fig. 1. The "theatrical machine" ME at work in a concert, during the performance of "Spiral", by K.Stockhausen, for trombone and radio output. The radio, audio amplifier and loudspeakers were installed on top of the robot navigating on stage, thus creating effects of "physical" spatialization of sound during the performance (trombone: Michele Lo Muto, live electronics: Giovanni Cospito and Andrea Pennese. Civica Scuola di Musica, Sezione Musica Contemporanea, Milan, June 1996). The movements of the robot can be influenced by the sound, by the movement trajectories and gestures of the performer. Furthermore, the robot's sound and music output can be part of the interaction process, i.e. influenced in real time by the movement of the performer

The former observes the zone of the stage occupied by an instrument player (the trombonist in Fig. 1) and react accordingly. The latter manages the movement of the robot and its (speech as well as non-speech) audio output. Both agents communicate with each other to control/generate a sort of (musical as well as gestural) counterpoint for trombone and "robot". Figure 1 shows a similar configuration of ME at work during a public concert.

Roughly speaking, we have an interactive performance for trombone and a robot-agent. The robot music output is sent by an audio radio link to the loudspeakers installed on top of the robot. The dialogue process between the two agents significantly contributes to the resulting music output.

Movement trajectories and particular gestures of the trombone player while performing can influence in real time the movement and the audio output of the robotic agent: the robot has on-board sensors to detect obstacles around itself, and spatial sensors to capture the performer's movements.

There is a "social interaction" process, a "dialogue" between the robotic agent and the player, which plays an important role in the composition. For example, slow movements and melodic lines characterized by long notes at a low amplitude could influence the internal (emotional) state of the robot, which could result in changes in its "style" of navigation, in a deviation of its trajectories (e.g. closer to the performer), and changes in its sound output. Sudden, nervous gestures, high-loudness tunes of the trombone with sudden peaks of energy can produce an opposite influence on the robot, which could move toward a negative "emotional state".

Since the robot "wears" amplifier and loudspeakers on board, the two previous situations can be perceived by the public also in terms of a different "physical spatialization" of

sound (i.e. a sound source navigating on stage), depending on the interaction process.

We have developed several applications similar in principle to the examples described above, where agents are based on several sensor systems we recently developed (see later sections, and Camurri 1995b) and the mobile robotic platform Pioneer 1, designed at Stanford Research Institute and Activ Media Inc. (Fig. 1). We also developed similar agents in other non-artistic applications; for example, in interactive museum exhibitions, where a robotic agent plays the role of a guide, interacting with visitors, and in theatre performances with actors. Our Theatrical Museal Machine (TMM) consists of the robot, the sensors placed in the environment used as a prolongation of its sensor system, and its model of emotions.

Another related application concerns "virtual TMMs", i.e. computer-animated characters for guiding "visitors" in a hypertext, a computer game, a virtual enviroment. We can think of it as a step toward the creation of more believable opponents (and partners) in computer games and virtual environments (Bates 1994).

In the rest of the paper, we adopt the second ME example to explain our approach, specifically to explain our model and dynamics of artificial emotions.

3 The role of artificial emotions in MEs

Let us discuss some issues concerning the motivation for introducing artificial emotions in MEs.

The first issue is about the nature of emotions in humans. Interacting with (including listening to) an agent which in its whereabouts talks about himself, is a strong motivation to consider it alive. This is therefore a motivational answer: we try to simulate emotions because, in so doing, our agent receives more attention.

The second issue concerns computational complexity. In real-world situations, an agent must consider a lot of things before taking decisions. Many such things have to be derived from its past experience, but this soon becomes unpractical. We therefore extract from a complex environment only two stimuli, "please" (or "carrot") and "pain" (or "stick"), which we use as acceleration components of a navigation process in a virtual space that, for simplicity, we call "emotion space". Basically, our model consists of a repository containing the information of the emotional state of the agent. The emotional state is a point in a space, which moves in accordance with the stimuli (carrots and sticks) from the inside and the outside of the agent.

The third issue is pragmatic. Since our agent is designed to be in contact with humans, it has to be "convincing". It certainly does not get angry, but if it says "I am angry" at such a moment when a real person could say that, we are satisfied. The ways in which it could show its angriness can be various: its style of moving around in the environment, its speech and non-speech audio output, its subsequent responses to similar stimuli in the future (this is our idea of memory: emotional memory), the expression of an artificial face, etc. Further, the agent must be "believable" (Bates 1994). For example, in cases where it has not enough knowledge to produce a correct answer, it should react most similar to a human in the same situation (e.g. tries to elude the request, becomes aggressive, etc.).

The fourth issue is different in its starting point. Let us think of a theatre play. The characters exist only for the duration of the screenplay. But their emotions seem real, and those "fictitious" ones generate our "real" ones, when we look at the performance. Therefore, a robotic agent on stage, with its plot, might generate real emotions in us, even if, only for a minute, we thought that its emotions were real. This saves human leadership in the creation: we are the only living beings that can talk about emotions and, most of all, we are the only ones that have the power to *generate* emotions from a *simulation*, either represented on a stage or in our minds (Frijda 1986). This characteristic is utilized in our work. Agents in an ME, when at work, should convey emotions to the spectators. We will judge the goodness of our model of emotions from the quality of emotions it will convey to the people exposed to the system. The goal is therefore to *stimulate* the complexity of the emotional capability of humans, *instead of imitating* it.

The fifth issue is slightly related to the previous one. It is well known that one of the main differences between natural language or sign languages on the one hand, and music or dance on the other is related to the associated semantics. In the case of natural or sign languages, it concerns the denotation of precise objects in the world and the communication of concepts, ideas, etc. In the case of music and dance, roughly speaking, there is mostly communication of emotional content, embedded in the interpretation, in the expression of intentions in the performers (see, for example, Sloboda 1985). The input modules of an ME should therefore be able to capture movement, gesture, sound features concentrating on discovering interpretation and emotional contents in performances (e.g. Sawada et al. 1995). In this direction it is also important to mention the work of Laban, with particular regard to his theory of effort, which includes studies on the relation between body movement in dance and the communication of intentions (Maletic 1987).

3.1 Machine gesture

It is important to analyse the way in which the ME "communicates" to humans by means of the movement and gesture of its physical (or robotic) agents. Robot movements are of course highly constrained, and should be considered integrated (as a component of a multimodal system) with other output channels, including audio, music, animated artificial face, etc. We want a system capable of stimulating the complexity of the emotional capability of humans. Our model of a robot's emotional space is divided into 13 zones, corresponding to 13 different characters. Nevertheless, we need (at least in the current robotic platform we have adopted) only a few basic styles of motion. Let us give some examples: a nervous movement (changing from slow and fast rapidly) with sharp curves could be the manifestation of anger. A constant velocity superposed by a fast change in the direction of movement could convey the idea of a tail-wagging, close to happiness. If the robot tries to follow a moving target, we can imagine it to be an extroverted one, if

instead it follows a wall until it reaches a corner, we easily think that it is sad and does not want to communicate.

4 ME inputs: movement and gesture detection

Our approach to the embedding of sensor technology in MEs is based on the following guidelines:

(i) to adopt robust, low-cost, wireless, as-simple-as-possible technologies;
(ii) to leave the users (both the dancers and the director/choreographer/composer) free to move and concentrate on their specific tasks (e.g. dance and music performance, design of a choreography), instead of on technology. This implies the use of small, wireless on-body sensors and of spatial sensors, not requiring any light or dress limitation, and being non-intrusive;
(iii) an ME usually needs multisensor fusion, to assure that all the required input can be gathered, to be integrated for as complex as necessary gesture and movement "style" analysis.

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

4.1 Human gesture taxonomies

Sensor systems return data on movement trajectories of (parts of) the body, acceleration, etc. Such low-level data must be analysed to extract high-level information useful for advanced interaction. An unresolved problem in the literature is the recognition and classification of gestures (Nigay and Coutaz 1993; Schomaker et al. 1995), and the identification of their semantics according to the context. A classification proposed by Nigay and Coutaz (1993) and a MIT Media Lab group (Thorisson 1995; Wexelblat 1994) 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 invisible objects or tools. 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).

References in the literature are usually specific to hand gesture taxonomies (posture, motion, hand orientation), and to handwriting and pen gestures. These approaches have been considered in our systems as a source of inspiration for 3D full-body gesture recognition and analysis.

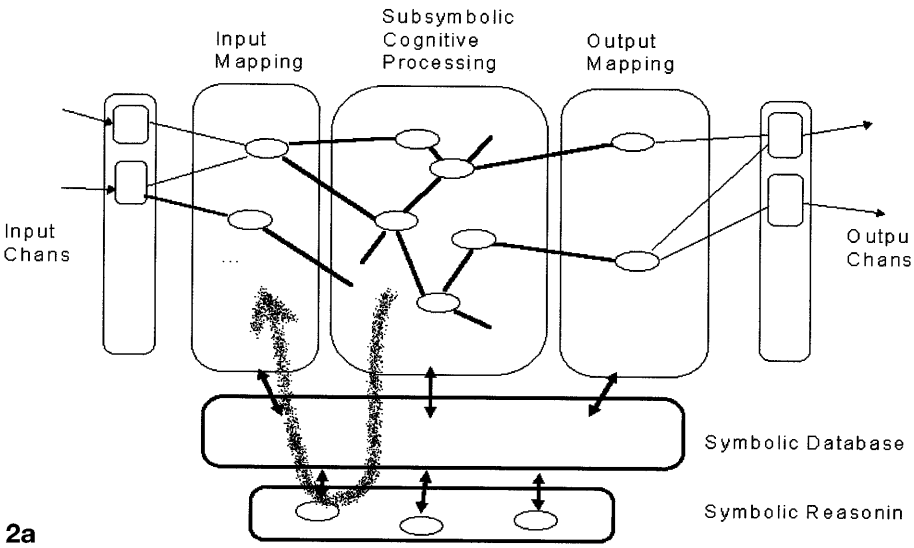
Another viewpoint in the classification of gestures concerns the physics and body response (Winkler 1995). Actions involving movement and gesture are executed by parts of the body, each having its limitations in terms of range of motion, speed, force, as well as weight and privileged directions. Furthermore, an action is characterized by features like ease and accuracy of execution, repeatability, measures of fatigue, required energy, volume range occupation. These aspects are also crucial in the design of our systems, and are included in the domain ontology (the database of system knowledge).

5 Gross anatomy of our system

The overall system architecture is a distributed network of agents. In some aspects, our system presents similarities, for example, with *Cypher* (Rowe 1993), *TouringMachines* (Ferguson 1992), and *Wolfgang* (Riecken 1992). The gross anatomy of the system, depicted in Fig. 2, can be also compared with Wexelblat (1994) and Thorisson (1995). The audio compact disc from IEEE CS that accompanied the July 1991 issue of the *IEEE Computer* magazine includes several music examples produced with an earlier version of the system designed for computer-aided composition (Camurri et al. 1991). A previous version of the system is described in (Camurri et al. 1994, 1995a).

Following the scheme depicted in Fig. 2a, we can identify the following basic building blocks:

- Input Mapping. It consists of a set of modules that receive physical signals from sensors (e.g. sounds, distances, angles, velocities) and maps them onto schemata or perceptual spaces. A schema is a carrier for structured knowledge. It can be a self-organizing neural network, or a reduced manifold representation of an input multidimensional space. The recorded signals may be mapped onto the schema in terms of movement trajectories. The recognized information is useful for subsequent cognitive processing modules (including the emotion engine). A crucial problem here concerns the data reduction from the multimodal input when mapped onto a schema. In the robotic agent example, the *input mapping* is implemented as a low-level software layer consisting of the set of the active behaviours and intentions managed by the upper level (how and why this allocation is performend is described later).
- Output Mapping. It changes the status of the external environment. It consists of a set of modules that manage the orders given by the cognitive processing modules and maps them onto the available output channels. The orders may pertain to high-level parameters. As stated before, the output may be of any kind and need not to be restricted to direct causal transformation of the input. An example of output mapping is the space of the "emotional" parameters for controlling the expression of an artificial face, the timbre space of a section of a composition, the behaviour of a robotic agent. We will see in a later section that our approach is based on an emotional space partitioned into regions. The character of the agent is a point in such space. A region prevails on the others and influences the agent output when its character enters in that particular region.
- Cognitive Processing. It is about reasoning, planning, and artificial emotions (i.e. reasoning activities are influenced by the emotional state of the agent). Activities can be subsymbolic and symbolic. An example of subsymbolic reasoning is the following: (i) the movement of a



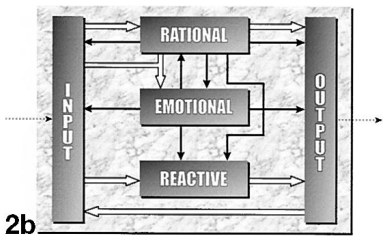


Fig. 2. a Cognitive architecture of an ME agent; **b** Software model of our agent architecture

dancer may be projected onto a topological map (input mapping), such as a neural network trained for gesture classification, (ii) the learnt gestures may correspond to particular attraction centres on the map; (iii) complex movements correspond to trajectories on the map, which may be influenced by attraction to particular gestures, thus allowing to recognize and classify high-level gestures. This example belongs to the category of subsymbolic reasoning based on metaphors, like abstract potentials, electric fields, and other dynamic systems. Also, our model of emotion space is subsymbolic. Symbolic reasoning performs symbolic inference and planning on the entities that emerge in the symbolic database from subsymbolic reasoning and recognition processes.

– Symbolic Database. It is a high-level, symbolic representation of the domain space(s) (e.g. music composition and performance spaces, movement and gesture spaces). It consists of a knowledge base management system, based on a standard representation language (a subset of first-order predicates), structured in a long-term and a working memory. The long-term memory stores ontologies related to general definitions of events, situations, facts, objects, and relations in the specific application domain. The long-term memory also stores the data describing the personality of the agent, that is, a description of the choice of regions forming the emotion space. Instances of relevant facts about events, situations, objects, and related features are stored during a work session in the (short-term) working memory. The agent updates its working memory according to its inputs and reasoning processes. Only a reduced amount of information emerges from the subsymbolic processing to the symbolic database. This seems reasonable from a cognitive point of view, and implies an improved efficiency in the symbolic reasoning.

The dynamics of the system is articulated and quite complex. Let us consider the following example on the role of the bidirectional links between the Symbolic Database, on one hand, and the Subsymbolic Cognitive Processing and the Input Mapping, on the other. Consider an example of an agent which observes the movement of a human (e.g. a performer, a dancer). The activity of the Subsymbolic Cognitive Processing modules (e.g. recognition of particular gestures, change of the style of movement) can cause modifications in the Symbolic Database in terms of new assertions of the particular situations occurred. This, in turn, can cause a change of state (or *context*, see next subsection) in the agent, including changes in its emotion state. This corresponds to changing and/or tuning the Input Mapping processes and, possibly, the Output Mapping. That is, the focus of attention in the Input Mapping drifts from one set of gestures to another (i.e., a change in the active set of observers), or the sensitivity to certain gestures (input parameters) can be tuned in the active observers. In other words, the cognitive processing can cause feedback to the perceptual processes in order to better capture the aspects of the external world of greater current interest or expectation for the agent. This important aspect of our architecture, shown by the large dotted arrow in Fig. 2a, models the system's capability to influence the development of its perception and emotion state.

Figure 2b shows another view of our agent architecture, which demonstrates the role of the artificial emotions module. In the figure, a white thick arrow represents a FIFO

buffer between modules. A black thin arrow from a component X to a component Y represents a data container upon which X has read-write access and Y read-only access. The emotional state of the agent, which is affected both by the external world and by the agent itself (through emotional stimuli respectively from Input and Rational modules), influences the working of the other four components through four parameters. The state of Rational is the rational knowledge of the agent, which is affected by and affects the external world (respectively through inputs parameters from Input and outputs parameters to Output), and influences the working of the other four components through four parameters. The Reactive module reactively interacts with the external world, while Input and Output translate between raw data of the external world and higher level information inside the agent (Input and Output Mapping). We stress that the idea of producing outputs from inputs "in parallel" through relatively fast and simple computations (Reactive) and relatively complex and intensive ones (Rational), is not new (see e.g. Ferguson 1992). Our main contribution, in fact, consists of the introduction of emotional computations and their integration with the other two. For example, the evolution of the knowledge of Rational can also depend on the current emotional state through the emotional-rational parameter. This dependence can essentially take place in the following ways: (i) the parameter encodes explicit knowledge about the emotional state (usually in the same form of the other knowledge of Rational), so that the inference engine of Rational also operates upon this "emotional knowledge" (without modifying it, since Rational has read-only access to it); (ii) the parameter affects the inference engine itself (as a kind of "emotional perturbation"), so that the same knowledge evolves differently in different emotional states. Of course, these two mechanisms are not mutually exclusive. Further details on our architecture can be found in Camurri and Coglio (1998). It is worthwhile noting that our emotional agents, besides multimedia-multimodal systems, might be very well suited to other fields as well (e.g. the kind of agents described in Bates 1994).

5.1 Contexts

Intuitively, we define a *context* as a sort of stable state of an agent in an ME, in which no relevant change occurs in its behaviour. Its emotion state, its way of observing inputs, of reasoning, and of mapping them onto outputs is roughly stable in a context. This does not mean that a context is static. In more detail, a context can be identified by the following components:

- a configuration of Input, characterized by a roughly stable "tuning" of its parameters;
- a configuration of Output, characterized by a roughly stable "tuning" of its parameters;
- an active subset of the Symbolic Database (Rational), including all short-term knowledge and the active long-term memory (situations, facts, music object classes, music composition and generative rules, etc.) which can be instanced (in the current context);
- a subset of the enabled (in the current context) reasoning processes (in Rational);
- the active emotion state in this context: a region in the emotion space corresponding to its current character;
- the goal(s) the agent tries to accomplish (in the current context).

"Roughly stable" means that possible changes in those parameters are small enough, so that they do not emerge in the symbolic database. In the introductory example, two contexts were defined: the percussion set and the string quartet. The dynamics of change between contexts is a function which characterizes the agent. It can model continuous, dynamic changes (e.g. morphing) between contexts. Such a function often corresponds in our system to navigation in potential fields or topological maps. Contexts correspond to regions in the map and changes can occur between adjacent contexts by defining meta-trajectories, driven by the dancer's movement. An example is shown in a further section, in the framework of a *HARP/Vscope* system application. Such *context change function* may also depend on the history of changes of contexts in the agent. For example, it might be useful to manage differently a context change which already occurred several times in the recent history. Such context change functions implement an ME as a *metamorphic* agent and thus also contribute to model emoting potentials (Riecken 1992).

6 The model of artificial emotions

Our model consists basically of a 2D space which can be navigated. Artificial emotion processing results in the emergence of emotion states which can affect deliberative (rational) and reactive processes of the agent.

6.1 The model

We have a 2D space. A point in such space represents the *character* of the agent. The two axes represent the degree of affection of the agent towards itself and towards others, respectively. We call these two axes "Ego" and "Nos", from the Latin words "I" and "We". A point placed in the positive x (Ego)-axis represents an agent whose character has in a good disposition towards itself. A point towards the left (negative) Ego would mean an agent fairly discouraged about itself.

The emotion space is usually partitioned into regions: for the robotic agent, we divided the entire space into 13 regions (see Fig. 3a), each of them is labeled by the kind of character the agent simulates.

We call a given partitioning of the emotion space in regions the *personality* of the agent. Each region represents one of the possible *character*s that the agent can assume: a point in a region represent the *character* of the agent, and a region around such point represent its *mood* or *disposition*.

Each region is defined by a maximum and minimum threshold for the quantities of the Ego and Nos character value. The minimum are called Apathy-Ego and Apathy-Nos, for the x- and y-axis, respectively. The maximum are called Maximum-Ego and Maximum-Nos. The character of the agent is in this example a two-component vector named Character-Ego and Character-Nos. In other agents (for music

application), the character is a vector of N reals (where N is the number of regions defining the personality), each representing the degree of membership to the corresponding region. In this latter case, we can have a "fuzzy" definition of the character.

We give here a description for 2 of the 13 regions available for the robotic agent.

- *Apathy:* (ABS(Character-Ego) $<$ Apathy-Ego) AND (ABS (Character-Nos) $<$ Apathy-Nos)) In this region, the robot is apathetic, e.g. the robotic agent has a slow, lazy style of moving and its speech and audio output is fairly sad.
- *Happiness:* (Apathy-Ego $<$ Character-Ego $<$ Maximum-Ego) AND (Apathy-Nos $<$ Character-Nos $<$ Maximum-Nos)). This is the "normality" region, that is, the robot is extrovert and happy enough to meet other people. This region represent happiness *because* the degree of liking towards itself and towards the others is medium.

The other regions follow similar guidelines and are shown in Fig. 3a.

It is important to notice that the fact of being in a given region not only defines the character, but also influences the (short-term) mood or disposition of the agent.

6.2 The dynamics

The character of an agent can change only by means of internal and external stimuli. This means that the class (in object-oriented terms) that implements the character supports only four high-level messages: "Carrots" (or "please", positive inputs) and "Sticks" (or "pain", negative ones). Carrots and Sticks can come either from the agent itself or from the external world: so we have Carrots-Ego, Carrots-Nos, Sticks-Ego, Sticks-Nos.

The motivation for having only these two types of messages is rather simple: navigation in a 2D space requires two non-aligned accelerations, and Ego and Nos are orthogonal. Possible extensions of the dimensionality of the emotion space, e.g. toward a 3D emotional space including, for example, a sort of "physical efficency" axis, only would need a third independent stimulus, still consisting of the two ambivalent high-level inputs (carrots and sticks).

Our current model is not a case of reinforcement learning: the agent is not able to learn. The four stimuli are used to explore the agent's emotion space. Nevertheless, how the character effectively moves in response to these four stimuli is in effect inspired by reinforcement learning.

The dynamics for changing the character is rather intuitive. For each region there are four styles of movements to define, for each message that can arrive from the internal modules of the agent or from the external world.

When the Mood object allocates a particular region, it adapts its interface in order to change in response to the stimuli from the lower level or from the outside, so we have actually 13 interfaces of the Mood object. We call *metamorphosis* the mechanism supported in the mood class, which allows different answers to the same stimuli in different regions. Metamorphosis is a basic mechanism in the *context change function* of an agent previously described.

As an example, let us examine how the mood class behaves when the agent is in two different regions. A detailed description of the whole model for all the 13 regions is given in Ferrentino (1997).

1. *"happiness" region*: a "happy" agent should take into consideration either positive or negative stimuli either from itself or from others. This means that the "happy-interface" is *linear*. This happens also in the other three regions, corresponding to the three quadrants, except the region around the origin (Apathy).
2. *"vanity" region*: here, the robot's Ego component is small, while its Nos is over the maximum (see Fig. 3a). This means that a vain agent will not take into consideration messages from himself, because its consideration is small (it is actually apathetic towards itself). Since it corresponds to a region already over the Maximum-Nos component, another carrot from the outside would cause no effect: it is *insensitive* to others' carrots, it already knows that the world is considering it positively. But what about others' sticks? We are in a region where the agent is used to carrots, so a single stick would be catastrophic: it is *very sensitive* to sticks from the outside here.

The algorithm for the navigation of the character point in the emotion space is rather simple. It is a variation of the well-known Metropolis algorithm, also used in simulated annealing techniques. In our algorithm, we have defined three possible maximum changes of the character's components, plus the zero-change in the case of insensibility. This means that each region has its own rectangle search. The step of change of the point representing the current character in the emotion space is then a random search in the proper rectangle. There is a "memory" of the previous character; instead of the typical Metropolis test to decide if to accept or not accept the new point, we move all the new points towards their old ones with a ratio determined by the "Temperature" of the mood, that is, the degree of "moodiness" of the agent. The new position is then classified into a region in the emotion space. As soon as the character moves from one region to another, the object "Agent Character" immediately switches off the current class and allocates an instance of the one pertinent to the new region in which the character point has moved. Then, the Character class asks the new current class to take control of the agent emotional component.

Here follows a description of the main algorithm.

Parameters: the old position of the character (OldChar); the current position (Char); the current stimulus with its direction (positive or negative); Mood and Temperature (two real numbers between 0 and 1).

Let us consider a positive stimulus in the x-axis (Carrot Ego).

1. The stimulus determines the width of the mood rectangle. The rectangle is not symmetric with respect to the starting point (the current character). This means that, with a Carrot Ego, the probability to move toward the negative abscissa is zero. The width of the mood rectangle, CarrotVal, is a constant value for each region.
2. The height of the rectangle is computed as follows: given the difference between the Y coords of the two previous

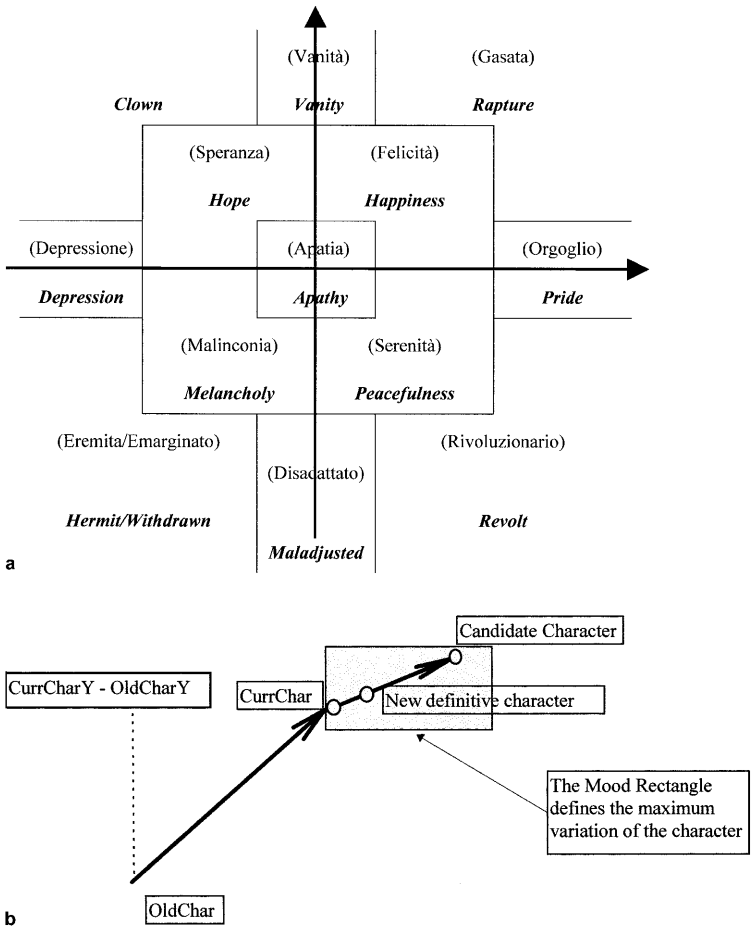


Fig. 3a,b. The example of emotion space defined for the robotic agent. **a** A proposal of segmentation of the ego-nos Emotion Space (the "personality" of the agent) **b** The dynamics of the character in the ego-nos space. The *grey rectangle* represents the current mood, which also defines the maximum variation of the character in the near future

characters, we consider a percentage given by Mood. The following formulas describe the mechanism:
rangey = $|\text{CurrCharY} - \text{OldCharY}| * \text{Mood}$
Ymin = CurrCharY $- 0.25 *$ rangeY
Ymax = CurrCharY + $0.75 *$ rangeY
Xmin = CurrCharX
Xmax = CurrCharX + CarrotVal
At this point, the rectangle is defined.

3. Choose a random point inside the rectangle, which is the candidate for the new character;
4. Define the segment between the current character and the candidate character.
5. Move the candidate character to the current character of a quantity proportional to the temperature;
6. OldChar = Char; CandidateChar = Char;
7. If the previous steps caused a change of region, apply the class-metamorphosis mechanism to adapt Mood to the new region.

The algorithm is somewhat similar to a sum of vectors, as can be seen in Fig. 3b. The algorithm and the formulas to compute the mood rectangle vary with the other cases of input stimuli.

6.3 The software layers

A layered model is necessary since we have a flow of information of different nature: sticks and carrots messages at the highest level, sensor and motor data at the lowest level.

Our model of artificial emotions is structured into three layers: the physical layer, the region layer, the mood (or disposition) layer. The first level is strictly integrated with the Saphira agent model (Konolige 1995) and is composed by *control classes*, one for each "behaviour" or "intention" (following the meaning used in Saphira). Control classes have a common parent: an abstract control class that has pure virtual member functions in common to activate or deactivate behaviours, a service function called by the behaviour, and a function that can be called from the region class currently active.

The purpose of this first level is to transform the physical quantities extracted from the behaviours into something more meaningful (i.e. higher level). For example, a ConstantVelocity behaviour can tell only if the robot it is controlling goes fast or slow relatively to the target velocity set by the user (or by the next layer, as we shall see). But this information could be integrated over some period of time to tell whether the robot is in a crowded environment or not. This is the purpose of the ControlCostantVel Object. In this and all the other behaviour control classes, we encapsulate all the Saphira-dependent implementation (variables, pointers...). If one wants to move the robot at a costant velocity, it is only necessary to allocate a ControlCostantVel object. The second layer has the responsibility of allocating the behaviour control classes.

Simplifying, the region layer is a collection of behaviours. When the agent's mood is in a certain region, the mood layer allocates the corresponding region class, which,

in turn, activates all the behaviours and intentions in the physical layer necessary for that region. Then the region starts to listen to messages from the physical layer. Some of those messages need no processing, or are not influenced by changes of mood. Instead, when a significant message is received by the region class, it generates a stimulus as a synthesis of all the current states of the agent, therefore related to the Ego component. The region passes the generated stimulus to the mood layer. But how does the region know whether to generate a positive or negative stimulus? The answer is that each region class is programmed to make the agent behave in a certain manner.

Therefore, each region class must not worry about the change of the agent mood. This is the role of the mood level. For example, when the character of the robot is in the "happiness" region, the mood allocates an object of type *Happiness* (subclass of *Region*), and activates it. This object takes control of the agent, allocates the necessary behaviour and intention classes in the physical layer, estabilishes the connection to the lower (physical) and the upper (mood) level and waits for messages. Whatever arrives from the lower level to the service function, the object *must* consider it as if it were always happy. No matter how bad the news, the *Happiness* class makes the agent always happy. If there is something that *is* bad (and this decision is made within each region class, what is bad in a region could be something good in another), the object communicates it to the mood class, but never waits for an answer.

The mood class changes the character in response to the stimulus from the lower level and, if the character happens to be in another region, it simply deactivates the Happiness class (which, in turns, deactivates what it had activated previously) and allocates an object of another region and gives it control of the agent.

This is a general guide for re-use. If, for example, we use the mood class to control an artificial composer we have to define N different "composers", each with its own character; each different composer should be enclosed in one of the N region classes. It should communicate with the lower level (which has to be totally re-defined, since we do not have motors and sonars anymore, but, for example, a database of music objects and generative rules), and extract from the music it is generating (or in general from the outworld, including other composers cooperating, for example, in the generation of a counterpoint), the stimuli for the mood class. Here, the model is slightly different, since there would be N composers concurrently active and cooperating.

6.4 Metamorphosis

Methamorphosis is a well-known object-oriented design pattern which models a class as a dynamic object, responding differently according to its internal state. It is a simple extension of the familiar concept of polymorphism. We have a metamorphic primitive (of a class) when the same name effectively refers to different bodies of the function being called at different times. In our example, we have actually several functions, but the client does not know which of them will be called, because this depends on the internal (private) state of the Region class.

7 Application environments

We used the robotic agent ME within this paper to explain our architecture. In this section, we give an overview of further ME applications we have recently developed.

7.1 HARP/V-scope

Here we describe an application for continuous human gesture tracking and its integration with sound, music and animated human models. It is an extension of the first introductory example of Sect. 2. In a few words, a dancer can control, at different levels, the music output and the expression of an artificial face by means of his/her movement. This application is composed of four main subsystems:

a) human movement acquisition – based on the V-scope sensor system (see below) – including data pre-processing and filtering;
b) input mapping, human movement and gesture recognition, including modules based on the force field metaphor, and capable of movement pattern classification by means of self-organizing neural networks;
c) Rational and Emotional, for the representation of music and gesture spaces, the management of contexts and artificial emotions;
d) output mapping, system outputs for sound, music, and computer animation (expression of an artificial face).

7.1.1 The sensor and low-level software

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 on other moving objects. 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 millisecounds per marker and the range of measuring depth can vary from 2 to 5 m. Faster sampling corresponds to a smaller depth, due to the limitations of ultrasound technology. Our experimental results show that 12–20 ms per marker is a good tradeoff between speed and stage size. We experimented with an accuracy of the V-scope hardware in the range of ± 0.5 cm, acceptable for our applications.

7.1.2 Input mapping: movement and feature extraction agents

Gesture recognition is carried out by concurrent modules in the input mapping, specialized for different kinds of recognition tasks. In our experiments, two different types of gesture/movement recognition agents were implemented: neural agents and force field agents.

The neural module adapts to gesture recognition an approach originally developed for handwriting (Morasso et al. 1995). In analogy to *graphemes* and *strokes* (handwriting primitives), we trained self-organizing nets with simple

movements and gesture primitives (*gestlets*). In the current implementation, neural modules have been developed for the classification of hands trajectories.

The concurrent raising and lowering of both hands, as well as symmetries and rhythmic patterns are observed by means of a cooperation of both the information extracted by the previous neural agents and by symbolic agents, allowing the integration of local data (e.g. the distance between hands, relative gesture speed), the output of the neural agents, and the local context information in the symbolic database. Further body gestures, e.g. raising and lowering the torso and legs movements, can be detected in the same way.

Another interesting category of feature extraction is based on the 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. An example of a bidimensional potential field, with three peaks corresponding to areas in the sensorized stage, are shown in Fig. 5 (see the video screen in the foreground). The potential field maps the stage in different areas, each characterized by a different behaviour, i.e. different movement/sound mappings. In a simple mapping, the (x,y) coordinates of the torso of the dancer can be mapped on to the field. The dance corresponds to a navigation in the map, wherein moving from an area to another means a continuous change from one context to another in the ME.

Another group of input mapping modules is delegated to extract higher-level whole-body gestures, as discussed in the previous section on our gestalt approach. This allows the composer/choreographer to model more complex music/movement correlations. Examples of this kind of information are "how *fast* the movement is", "how much *in tempo* the dancer is moving", "how he/she occupies the volume", "the *smoothness* of the movement", "measure of the coordination between arms", etc. This is obtained from integration over time of data from several sensors. Two different observation time slices, approximately 0,5–1 s and 3–5 s, are currently used. Time slices on which the agent operates can vary dynamically, e.g. on the basis of the amount of "quality" of the recognition: for example, a decreased quality in the movement recognition (e.g. different agents return conflicting data) can cause the agent to feed back the input mapping processes to vary their time granularity and/or their time slice on which they operate. This is a significant example of a feedback from Cognitive Processing to Input Mapping (discussed at the end of the previous section Gross Anatomy of the System).

7.1.3 Symbolic reasoning

In the simplest case, the recognized gestures and trajectories can be directly used to generate and control music events, in a sort of direct cause-effect mechanism. Often, it is useful to process the input data, to take into account what happened in the recent past in the dance and the music, the internal state of the system, in order to decide the most suitable output.

Symbolic resoning is designated to play an important role in this respect. It operates on symbolic representations of the movement and music and cannot directly access the subsymbolic input signals. Symbolic reasoning can be activated by the instantiation of particular assertions of a situation (e.g. a recognized gesture in a given context, or a change in the dance "style"). Such flow of new assertions in the Symbolic Database (the working memory) can trigger symbolic inferences, which, in turn, can cause further modification in the knowledge base, e.g. to force a change of the music composition/generative rules to be applied and to give a feedback to the Input Mapping to modify the roles and the parameters corresponding to changes in the sensitivity to movement patterns (a sort of guide mechanism of the focus of attention). Assertion in the Symbolic Database can also emerge from the subsymbolic level as the result of changes in the emotional state of the system, as previously described. In our sample application, symbolic reasoning manages composition rules for determining the music structure produced by the dancer, as described below.

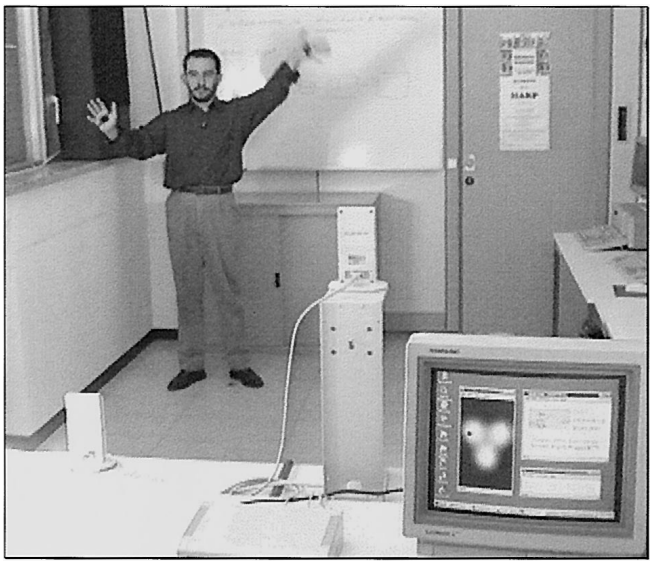


Fig. 4. HARP/V-scope at work in the DIST Laboratorio di Informatica Musicale. The position of a marker located on the user's chest corresponds to the *black dot* in the stage-agent window on the screen in the foreground. The position of hand-held markers are used by other recognition modules

7.1.4 Output agents

The system is presently designed to control sound and music in real time and particular computer animation. The MIDI and sound output low-level modules manage the messages from system's inner modules. Further output modules concern the control of computer animated human models. They reside on an SGI Indigo workstation (a computationally lighter version runs on Win32).

7.1.5 Examples of real *HARP/V-scope* applications

HARP/V-scope can be used as an "adaptive" *hyper-instrument*, able to gradually modify itself according to the position of the player on the stage. In the setup of Fig. 4, the player can navigate among three different hyper-instruments corresponding to different areas in the sensorized stage, mapped on the potential field (see the video screen in the

Fig. 5. HARP/V-scope at work in the concert "En Voyage", at the First Int. Theatre Festival "Al Confine" (25 July 1995, Ventimiglia), music by G. Palmieri, dancer N. Ragni, fagotto R. Vernizzi, sax F. Aroni Vigone, bass S. Ferrari. The system is here used for live electronics: the dancer intervenes with her movement in real time on the sound output of the acoustic instruments. One of the V-scope markers is visible on the right hand of the dancer (photo M.M. Pasqualini)

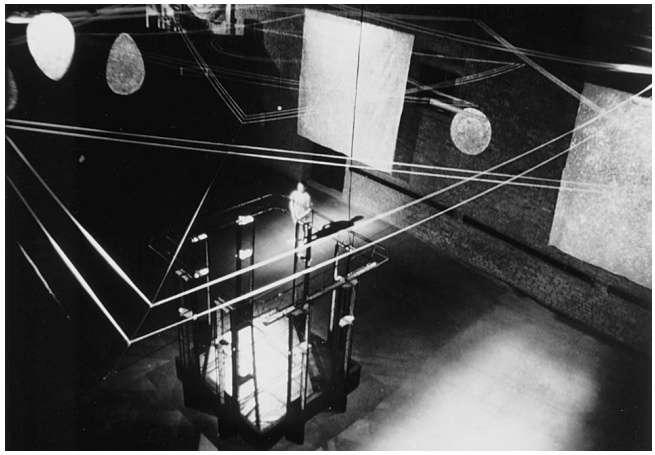


Fig. 6. A view of the ME developed for the art exhibition "Upper Space" at Theatre Fondamenta Nuove, in which the SoundCage IMM and the DanceWeb systems have been integrated in the whole space of the theatre with sculptures and artworks (patrocined by Biennale Architettura, Venice, 12 Sept. – 17 Nov. 1996)

picture of Fig. 4). The areas/hyper-instruments are contexts, and the potential field function manages the mechanism of change of context. In this simple experiment, we used three markers: one for either hand and a third for a location on the chest. The latter is used to capture (i) body position in the force field map (x and y coords), and (ii) body height (z coord), discriminating whether the dancer is standing or crouched. Hand gestures, recognized with the neural module, control the sound and music output. Opening/closing the hands allows to switch on/off the recognition process. A "morphing" effect from one instrument to the other is attained as a consequence of the movements in different environment positions, mapped onto the potential field. Moreover, with simple symbolic modules, it is possible to detect the style of the dancer, thus changing dynamically the reactivity of the hyper-instruments. A similar application concerns the control and navigation in composition/performance spaces. In this case, potential fields assume a completely new meaning: they represent composition and performance spaces which can be navigated by movement. In another application of the system, each area in the map is associated with compositional and performance rules to control the timbral, harmonic, and melodic output, allowing the dancer to navigate in the compositional/cognitive space. In this case, the three areas correspond to new contexts, and the field in the map models the rules of (continuous) change between contexts. A live performance on stage is shown in Fig. 5.

7.2 SoundCage interactive music machine

A line of work regards the use of HARP as a supervisor for the family of systems *SoundCage*™ *Interactive Music Machine (IMM)*, for the real-time acquisition, processing of the movements of a dancer, shown at work in Figs. 6 and 7. The IMM is designed to drive MIDI digital synthesizers and multimedia devices (e.g. lights, lasers, and special effects in general, including actuators).

The SoundCage hardware is composed of a set of proprietary active infrared and pressure sensor systems displaced in a sort of "cage" structure (the stage), and of special I/O boards. Pressure sensors are in the floor of the IMM. In total, there are more than 60 spatial sensors. We designed the system such that the dancer is completely free to move without any on-body sensor. The spatial sensors cover non-linearly the 3D space of a cube of less than 3 m (the cage). We assume a single dancer inside each cage. Therefore, we take into account features of the human body, the degrees of freedom, and some typical movement patterns. Certain crucial areas in the cage are more densely covered with sensors; roughly, there are four different horizontal "layers" which are more carefully observed in the cage: a floor layer, an ankle layer, an arm layer, and an over-the-head layer.

The same approach to gestalt gesture recognition is adopted here, as in all the other movement recognition systems we developed.

7.3 DanceWeb

The *DanceWeb* is based on a number of different low-cost sensor systems (ultrasound, infrared, pressure, and, recently, camera-based). The current version of the system supports up to 64 ultrasound, and 48 analogue and digital sensors, and the software includes a DLL supporting a Matrox Meteor frame grabber. All sensors can be freely distributed in the environment in which the ME operates. In default installations (Fig. 7, right), the sensors are assembled in a sort of "star" or "web" hanging from the ceiling, with pressure sensors on the floor.

A rack unit embeds a microcontroller board and simple electronics for the acquisition and preprocessing of sensor data, connected via a fast serial link to the PC or supervision workstation. Low-level measurements, e.g. distance measurements of body parts from ultrasound sensors, are integrated with other sensor data in our ME software environment. The DanceWeb can be used as a stand alone system, like in the default "star" configuration shown in Fig. 7,

Fig. 7. Another ME installation characterized by the integration of different systems: (i) the SoundCage IMM is on the left, (ii) HARP/Vscope is in the centre (the four sensors are on the floor), (iii) HARP/DanceWeb is on the right, hanging from the ceiling. This installation was presented at the European Information Technology Conference and Exhibition EITC'95 as an entertainment application in the CyberCafé area (Brussels Congress Centre, 27–29 Nov. 1995), and it was selected by CEC DG III (Industry) as one of the most innovative results from European Projects in Information Technology in 1995

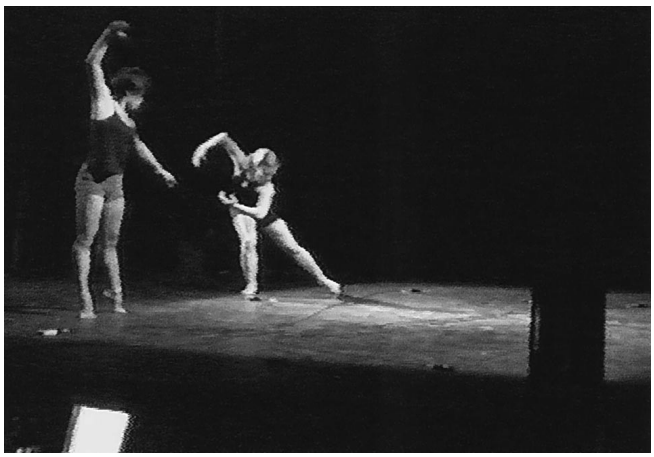


Fig. 8. The HARP/DanceWeb in a custom stage configuration at a concert with the New Music and Dance Ensemble of New York University, Pisa, July 1996. Sensors are placed on the floor and directed vertically upwards (choreography and dance: Anne Wennerstrand and Jeong Nam Park, New York University)

or in configurations freely distributed in the environment as in Fig. 8. It can also be integrated with the SoundCage IMM. Such a multisensor fusion has two basic advantages: (i) in some cases, it permits a more reliable gesture recognition by integrating the input mapping of the IMM with the one of DanceWeb (ambiguities in the recognition from one system might be solved by integration with the other), (ii) it extends the IMM with further DanceWeb observers capable of recognizing new kinds of gestures.

8 Software implementation

Our ME architecture is implemented in MS Visual C++ 5.x and Quintus Prolog 3.x, under Win32 operating systems (Windows 95 and Windows NT 4). The symbolic database and inference system is a compiled Prolog module encapsulated in a C++ class. In a recent version of the system, we re-implemented the Prolog component in C++, for better real-time performance. As for the robotic extension, the Saphira applicaton developed by Kurt Konolige for the control of the robotic platform is adopted and integrated in our environment as a separate thread (in the Input and Output modules). The standard behaviours available with the robot have been modified in order to communicate with our software environment in which they live.

The architecture of the system is rather *loosely coupled*. The behaviour-level classes in Emotion, for example, do not need to know what it is happening in the upper levels. They have only a pointer to their "father" and to that object to which they communicate what they sense in the environment. The upper level, in contrast, is a sort of agent that can control its lower layers (e.g. the robot) at a higher level. But these orders do not arrive from the outside. The same Region class is responsible for the overall robot activity and does not respond to anybody above it. The Mood class can only switch moods on and off.

9 Discussion: the role of non-speech audio in multimedia systems

In the robotic agent described in the paper, a very important aspect both during its development, testing and its use in museum exhibitions and multimedia concerts concerns its non-speech audio output: it was programmed to express its internal emotional and rational state by means of high-level music parameters. During the development and testing, this allowed us to work and understand the system behaviour without the necessity of sitting in front of a computer screen, since its audio output gave us all the required information (and at the same time we were able to look at the robot).

More generally, in real-world applications (see, for example, Fig. 9a and b), a high-level, structured non-speech audio communication channel can dramatically improve the interaction and the understanding of users, e.g. in the case of museum visitors or of performers and spectators.

10 Conclusion

In this paper, we focused on MEs, a particular family of agents capable of establishing creative, multimodal user interaction, and exhibiting dynamic, intelligent, real-time, adaptive behaviour. MEs open new niches for applications, from art to music, culture, entertainment, and a number of industrial applications, many still to be discovered. In the first part of the paper, we sketched and discussed the main guidelines and requirements for the design of MEs, including the necessity of artificial emotions. More effective, stimulating, and natural interaction with humans can be achieved by including artificial emotions as one of the main components of such complex systems. For example, artificial emotions can be a convenient way to model some aspects of full-body movement and music, which are typical languages for conveying emotional content. Furthermore, the metaphor of artificial emotions can play the role of "glue" among these

Fig. 9. a The Theatrical Museal Machine at work as a "Cicerone" during the museal exhibition *Mostra della Cultura Scientifica e Tecnologica Imparagiocando3*, Palazzo Ducale, Genoa, March–April 1996; **b** the version of the system at the permanent museum Città dei Bambini, Porto Antico, Genoa

expressive languages, as well as among different modalities (e.g. visual, acoustic, gestural).

We presented our proposal of an ME emotional agent architecture. Four recent implementations of MEs have been surveyed: the SoundCage Interactive Music Machine, the HARP/V-scope, the DanceWeb, and the Theatrical/Museal ME.

Some of the systems surveyed in this paper have been selected for presentation in live demonstrations at a stand during EITC'95 (European Information Technology Conference and Exhibition, Brussels Congress Centre, 27–29 November 1995). Our systems have been utilized for a number of recent events including the following:

- *Upper Space*, Interactive art installation, Teatro Fondamenta Nuove, patrocined by Biennale Architettura, Venezia, 12 Sept. – 17 Nov. 1996. A sophisticated ME has been developed and integrated with sculptures by Pascal Lansonneur (see Fig. 6). We designed an installation based on SoundCage IMM and DanceWeb. The idea is to have an ME, in which the reactivity and the feedback of the system (i.e. sensor density in the space) gradually grows when moving from the periphery to the centre of the ME. Light changes cause variations in the perception of the sculptures and change the overall behaviour of the ME. The ME is realized by a specific agent application, SoundCage IMM (at the centre), surrounded by a custom configuration of DanceWeb.
- *Remote Dance*, stage/workshop with the New Music and Dance Ensemble of New York University, directed by Esther Lamneck, Pisa, July 1996. We experimented with integrating HARP/V-scope and DanceWeb applications (see Fig. 8).
- *Outis*, opera by Luciano Berio, vernissage at Teatro alla Scala of Milano, October 1996. In this work, we developed a simple ME based on a floor pressure sensor system.
- *Città dei Bambini*, a 3000-m^2 interactive science museum permanent exhibition for children, recently opened at Porto Antico in Genoa (Italy). It is made up of two main modules, one developed by La Villette (Paris) and the other by Imparagiocando (formed by University of Genoa, National Institute for the Physics of Matter, National Institute for Cancer Research and Advanced Biotechnology Center, Arciragazzi, and Movement International Loisirs Science et Technologie). The Music Atelier, which is part of the latter module, has been conceived and realized by the Laboratory of Musical Informatics (of which the first author is the director and the second author a staff member) of DIST, University of Genoa. The Music Atelier consists of five agents/games characterized by multimedia-multimodal interaction involving music, movement, dance, and computer animation. One of such games is a Cicerone robot which guides visitors in the Atelier (Camurri and Coglio 1998) (see Fig. 9).

The research on MEs still requires much work. On one hand, significant results are expected in the near future in motion-tracking hardware, given the growing interests from games and movies industries. On the other hand, hot issues include theoretical and experimental work on MEs from both artists, musicologists and other scientists. Our research is currently directed at the refinement of the agent model and at the study of the integration of dance/gesture and music languages. The experimental study of the perception of gesture and sound "gestalts", the classification of full-body movements, and the

further refinement of representation and reasoning on contexts, including artificial emotions, are other crucial research topics.

Acknowledgements. The research described in this paper has been partially supported by the Esprit Project 8579 - MIAMI (Multimodal Interaction for Advanced Multimedia Interfaces), by a special project of the Italian National Research Council (CNR) on interactive multimedia systems for art and entertainment, and by MURST projects. We wish to thank the staff of the Laboratorio di Informatica Musicale at DIST: particularly, Alessandro Coglio, Alberto Massari, Claudio Massucco, Riccardo Rossi, Roberto Chiarvetto, and the composers Riccardo Dapelo and Giuliano Palmieri. The SoundCage IMM has been developed by SoundCage Srl. A special thank to Alvise Vidolin for useful discussions.

References

1. Bates J (1994) The role of emotions in believable agents. Commun ACM 37(3): 122–125
2. Camurri A, Morasso P, Tagliasco V, Zaccaria R (1986) Dance and movement notation. In: Morasso P, Tagliasco V (eds) Human Movement Understanding. North Holland, pp 85–124
3. Camurri A, Canepa C, Frixione M, Zaccaria R (1991) HARP: A System for Intelligent Composer's Assistance. IEEE COMPUTER 24(7): 64–67, July 1991. Revised extended version in: Baggi D (ed) Readings in Computer Generated Music, pp 95-115, IEEE CS Press, 1992
4. Camurri A, Frixione M, Innocenti C (1994) A Cognitive Model and a Knowledge Representation Architecture for Music and Multimedia. Journal on New Music Research - Interface 23(4), Swetz&Zeitlinger, Lisse, The Netherlands, pp 317–347
5. Camurri A, Catorcini A, Innocenti C, Massari A (1995a) Music and Multimedia Knowledge Representation and Reasoning: the HARP System. Computer Music Journal 19(2):34–58
6. Camurri A (1995b) Interactive Dance/Music Systems. In: Proc. Int. Computer Music Conference ICMC'95, Banff, Alberta, Canada. ICMA Press
7. Camurri A, Coglio A (1998) An Architecture for Emotional Agents. IEEE Multimedia, Oct.–Dec. 1998, IEEE CS Press (in press)
8. Ferguson IA, (1992) TouringMachines: Autonomous Agents with Attitudes. IEEE Computer 25(5)
9. Ferrentino P (1997) La Macchina Museale Teatrale: modello e implementazione di un sistema ad agenti con stati emotivi. Master Thesis, DIST, Faculty of Engineering, University of Genova
10. Frijda N (1986) Emotions. Il Mulino, Bologna (Italian Edition)
11. Konolige K (1995) Saphira software documentation and demos. http://www.ai.sri.com/∼konolige/saphira/#publications
12. Laurel B, Strickland R, Tow R (1994) Placeholder: Landscape and Narrative in Virtual Environments. Computer Graphics 28(2): 118–126
13. Leman M (1995) Music and Schema Theory – Cognitive Foundations of Systematic Musicology, Springer Series in Information Sciences. Springer, Berlin Heidelberg New York
14. Machover T, Chung J (1989) Hyperinstruments: Musically intelligent and interactive performance and creativity systems. In: Proc. Intl. Computer Music Conference, Columbus, Ohio. ICMA Press
15. Maes P, Blumberg B, Darrel T, Pentland A, Wexelblat A (1995) Modeling Interactive Agents in ALIVE. In: Proc Int. Joint Conf. On Artificial Intelligence IJCAI-95, Montreal, Canada
16. Maestri G (1995) Capturing Motion. Computer Graphics World 18(12)
17. Maletic V (1987) Body Space Expression. Mouton de Gruyter.
18. Moore FR (1990) Elements of Computer Music. New York, Prentice Hall
19. Morasso P, Limoncelli M, Morchio M (1995) Incremental learning experiments with Scriptor: an engine for on-line recognition of cursive handwriting. Machine Vision and Applications 8: 206–214
20. Morita H, Hashimoto S, Otheru S (1991) A Computer Music System that Follows a Human Conductor. IEEE COMPUTER 24(7): 44–53
21. Mulder A (1994) Virtual Musical Instruments: Accessing the Sound Synthesis Universe as a Performer. In: Proc. First Brazilian Symp. on Comp. Music, 14th Annual Congr. Brazilian Comp. Soc., Caxambu, Minas Geiras
22. Nigay L, Coutaz J (1993) A design space for multimodal systems – concurrent processing and data fusion. In: Proc. INTERCHI-93 – Conf. on Human Factors in Computing Systems, Amsterdam, The Netherlands. Addison Wesley, Reading, Mass., pp 172–178
23. Otheru S, Hashimoto S (1992) A New Approach to Music through Vision. In: Balaban M, Ebcioglu K, Laske O (eds) Understanding Music with AI. AAAI-MIT Press, Cambridge, Mass., pp 402–410
24. Povall R (1995) Composition in Interactive Environments. J New Music Res 24(2): 109–120
25. Riecken D (1992) Wolfgang – A system using emoting potentials to manage musical design Understanding Music with AI: Perspectives on Music Cognition, In: Balaban M, Ebcioglu K, Laske O (eds) Understanding Music with AI. AAAI-MIT Press, Cambridge, Mass., pp 207–236
26. Rowe R (1993) Interactive music systems. The MIT Press, Cambridge, Mass.
27. Sawada H, Ohkura S, Hashimoto S (1995) Gesture Analysis Using 3D Acceleration Sensor for Music Control, In: Proc. Intl. Computer Music Conference ICMC-95, Banff, Alberta, Canada. ICMA Press, pp 257–260
28. Schiphorst T, Crawford J, Gotfrit M, Demers LP (1993) The Shadow Project: An Exploratory Workshop in Performance and Technology. In: Proc. Fourth Biennal Arts & Technology Symposium, Center for Arts & Technology at Connecticut College, New London, pp 132–142
29. Sloboda J (1985) The Musical Mind, Oxford University Press
30. Schomaker L, Nijtmans J, Camurri A, Lavagetto F, Morasso P, Benoit C, Guiard-Marigny T, Le Goff B, Robert-Ribes J, Adjioudani A, Defee I, Munch S, Hartung K, Blauert J (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)
31. Thorisson KR (1995) Multimodal Interface Agents. PhD Thesis, MIT Media Lab., Cambridge, Mass.
32. Ungvary T, Waters S, Rajka P (1992) NUNTIUS: A computer system for the interactive composition and analysis of music and dance. Leonardo 25(1):55–68
33. Vertegaal R, Ungvary T (1995) The Sentograph: Input Devices and the Communication of Bodily Expression, In: Proc. Intl. Computer Music Conference ICMC-95, Banff, Alberta, Canada. ICMA Press, pp 253–256
34. Wexelblat AD (1994) A Feature-Based Approach to Continuous Gesture Analysis, PhD Thesis, MIT Media Lab, Cambridge, Mass.
35. Winkler T (1995) Making Motion Musical: Gesture Mapping Strategies for Interactive Computer Music. Proc. Intl. Computer Music Conference ICMC-95, Banff, Alberta, Canada. ICMA Press, pp 261–264

ANTONIO CAMURRI received his M.S. degree in electrical engineering in 1984 and his Ph.D. in computer engineering in 1989 from University of Genova. He currently is associate professor at DIST (Dept. of Informatics, Systems and Telecommunications of the University of Genova), where he teaches software engineering. He is the founder and scientific director of the *Laboratorio DIST di Informatica Musicale* . He is president of AIMI (Italian Association for Musical Informatics), member of the ExCom of the IEEE CS Technical Committee on Computer-Generated Music, associate editor of the international journal *Interface - Journal of New Music Research*, and served in the scientific committees of international conferences. His research interests include computer music, interactive systems, agent architectures, artificial emotions, multimodal interfaces, virtual reality. He is owner of patents in computer systems and algorithms.

PASQUALINO FERRENTINO was born in Genova, Italy, in 1973. He received his M.S. degree in electrical engineering from University of Genova in 1997. His research interests include intelligent man-machine interfaces, models of artificial emotions and didactic applications. He Currently is a member of the staff of the Laboratorio di Informatica Musicale at DIST-University of Genova.