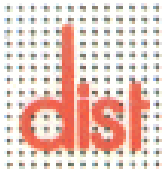
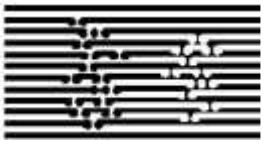
UNIVERSITY OF GENOVA
Faculty of Engineering



Department of
Communication,
Computer and
System Sciences



InfoMus Lab
Laboratory of
Musical Informatics

Ph. D. DISSERTATION

Computational models of expressive gesture in multimedia systems

by Gualtiero Volpe

Supervisor: Prof. Antonio Camurri

Preface

This work is part of a research carried out in the framework of the EU-IST Project MEGA (IST-1999-20410, www.megaproject.org). MEGA (Multisensory Expressive Gesture Applications) is centered on the modeling and communication of expressive and emotional content in non-verbal interaction by multi-sensory interfaces in shared interactive mixed reality environments. In particular, the project focuses on music performance and full-body movements as first class conveyors of expressive and emotional content. Main research issues are the analysis of expressive gesture (i.e. analysis of the expressive content conveyed through full-body movement and musical gestures), the synthesis of expressive gesture (i.e. the communication of expressive content through computer generated expressive gestures, e.g., through music performances, movement of virtual as well as real, robotic characters, expressive utilization of visual media), the strategies for mapping information coming from multi-modal analysis onto real-time generation of multimedia content.

The work has been carried out at the InfoMus Lab (Laboratorio di Informatica Musicale) at DIST - University of Genova.

Before starting the discussion I have to thank many people that directly or indirectly contributed to the ideas described in this dissertation. First of all, I wish to thank my supervisor Prof. Antonio Camurri, scientific director of the InfoMus Lab, who followed my research from its very beginning. Then, the friends and colleagues who worked with me at the InfoMus Lab along these years: Barbara Mazzarino and Riccardo Trocca on the analysis of expressive gesture in human full-body movement, the EyesWeb developers Paolo Coletta, Massimiliano Peri, and Andrea Ricci, Matteo Ricchetti for the useful discussions, and the other members of the InfoMus staff (Roberto Chiarvetto, Roberto Dillon, Alberto Massari, and Cesare Mastroianni).

I also thank the friends and colleagues who worked or are working in the MEGA project. The joint collaboration in MEGA has been of paramount importance for the development of this research. In particular, Marie Djerf, Ingrid Lagerlöf, and Erik Lindström of the staff of Prof. Alf Gabrielsson at the Department of Psychology of the University of Uppsala, Sweden; Prof. Marc Leman at IPEM – Ghent University, Belgium, the IPEM crew, and in particular Johannes Taelman, Koen Tanghe, and Guy Van Belle; Prof. Giovanni De Poli at DEI – CSC – University of Padova, Italy, together with Sergio Canazza, Carlo Drioli, Antonio Rodà, Patrick Zanon, and the colleagues who only recently joined the project; Anders Friberg and Sofia Dahl from KTH, Stockholm, Sweden, Ivar Kjellmo from Ocatga/Telenor, Norway, all the visiting researchers who worked at the InfoMus Lab in the framework of the EU TMR MOSART network (i.e., some of the ones listed above plus Matija Marolt, Joanne Mc.Elligott, Declan Murphy, Renee Timmers, and Olivier Villon), and all the ones who do not appear in this list and who I certainly forgot to thank!

Finally, thanks to my parents and my family who supported me along all these years.

Table of Contents

Introduction

While technology continuously evolves toward faster and smaller devices and broadband communication systems become available to larger groups of people, the need of novel human-machine interaction paradigms strongly emerges and researchers are faced with the challenge to provide users with effective, intuitive, powerful ways to communicate with the more and more technological environment they live in. In what now could seem a futuristic house (but we could live in one of them in a few years, and maybe someone already does) hidden computing devices placed all around and diffused into everyday objects will manage and support many aspects of people's lives.

Some research groups in the U.S. and in Japan have faced such a challenge for novel interaction paradigms by trying to introduce a further level of processing in computer systems, a level of processing dealing with emotional, affective information. This attempt produced two novel research branches in Human-Computer Interaction: Affective Computing in the U.S. (see for example Picard, 1997) and KANSEI Information Processing in Japan (see for example Hashimoto, 1997). The EU-IST MEGA project (Multisensory Expressive Gesture Applications) tries to distinguish itself from its counterparts in U.S. and Japan, by following a European route in investigating the same topics. Such route is grounded on the traditional and solid bases of European humanistic culture: theories from psychology, philosophy, performing arts, and humanities are the sources research is inspired to.

The MEGA project (and this dissertation) strongly focuses on the development of interactive multimedia systems for performing arts as a main concrete output. The choice of performing arts as application domain is due at least to two aspects: (i) performing arts widely use non-verbal and expressive communication mechanisms to convey emotional, affective information to the audience and therefore represent an ideal test-bed for computational models and algorithms dealing with this kind of information; (ii) technology can bring important concrete contributions to this field by providing tools enabling novel ways of conceiving artworks and maybe allowing the development of completely new art forms.

If from the one hand the focus on performing arts seems to be justified by the arguments above, on the other hand it constitutes a further challenge for this research. In facts, art and technology are two words that at a first glance seem to be in opposition each other, the former related to the sphere of aesthetics and humanities, the latter to the field of science and engineering. At the beginning of this dissertation it is therefore worth to ask myself (as a colleague and friend wrote) if it does make sense for technology to deal with art and vice versa. A first comment to this difficult question can be found here above, i.e., I believe that art can take advantage from technology in term of tools allowing artists to create scenarios that are not possible otherwise. Notice that these novel scenarios are not limited to employment of virtual or mixed reality techniques, as one can initially think. Of course, mixed reality techniques are of primary importance, but they are mainly related to the visualization aspects, i.e., they can display worlds in which real and virtual objects and subjects interact and overlap. Technology however can interact with art at a deeper level than visualization, i.e., at the level of the language art employs to convey

content and to provide the audience with an aesthetical experience. Interaction at this level requires technology to be able to deal with the artistic content, i.e., what the artist wants to communicate and with the communication mechanisms enabling the experience of the audience. In this perspective, research on expressive gesture as a main conveyor of information related to the emotional sphere allows a redefinition of the relationship between art and technology: from a condition in which art uses technology for accomplishing specific tasks that only technology can afford (or that computers can do better than humans) to a novel condition in which technology and art share the same expressive language and in which technology allows the artist to directly intervene on the artistic content and in the expressive communication process.

I do not know if such a deep integration will ever be possible. The current state-of-the-art is very far from such a condition and this new generation of interactive systems for performing arts is still far to be developed. Nor I forget the possible risks related to this research for example in term of partially expropriating the artist of its artistic creation or in term of repercussions in other application domains (that can be either positive, e.g., improved interaction with computers, more effective techniques for therapy and rehabilitation, or negative, e.g., intrusion in the emotional life of individuals, control of individuals' behavior through expressive communication). The studies carried out in the last ten years, the currently ongoing projects, and this dissertation should therefore considered as pioneer researches toward that "third phase of information processing" (Hashimoto, 1997) that by using the current advances in signal processing (first phase) and logic (second phase) might lead to a novel generation of computer systems able to deal with affective, emotional information. Of course, as usual it is mainly responsibility of researchers and engineers working in this area to maximize the benefits and minimize the risks of this technology.

In this framework, this dissertation focuses on the development of paradigms and techniques for the design and implementation of multimedia and multimodal interactive systems mainly in the application field of performing arts. The dissertation is divided into two main parts. In the first one, after a short review of the state-of-the-art in research fields related to this research, the focus moves on the definition of environments in which novel forms of technology-integrated artistic performance could take place: these are distributed active mixed reality environments in which information at different layers of abstraction is conveyed mainly non-verbally through expressive gestures. Expressive gesture is therefore defined and a possible internal structure of a virtual observer able to process it (and inhabiting the introduced environments) is described in a multimodal perspective. The definition of the structure of the discussed environments, of the virtual and mixed subjects inhabiting them and the techniques for expressive gesture processing constitute a source for requirements, a paradigm for design and development, and the basic bricks for implementing the interactive systems this work addresses. The second part of the dissertation introduces a concrete example of implementation of a virtual observer, i.e., a virtual subject observing human full-body movement, extracting expressive cues from it, and attempting to classify expressive gestures according to their emotional content.

The developed algorithms have been implemented as software modules for the EyesWeb open platform (see Appendix A) and constitute the core of the EyesWeb Expressive Gesture Processing Library (see Appendix B).

PART 1

EXPRESSIVE ENVIRONMENTS AND EXPRESSIVE GESTURES

1. Technologies for expressive environments

Recent developments in Human-Computer Interaction (HCI) and multimedia are leading toward the design and implementation of systems that from the one hand are widely increasing usability and user-friendliness of computers in application fields where computers are traditionally used (e.g., computer aided design, office automation), and on the other hand are introducing computer systems in areas where computers only had a marginal role or were regarded with suspicion (e.g., in humanistic studies).
Two trends in technology evolution of interest for this dissertation can be observed:

(i) Computers are more and more able to process high-level information coming from their users: they can detect and interpret user's actions and adapt their behaviour to user's needs. In this scenario a particular role is played by the ongoing research focusing on the analysis and synthesis of information related to the expressive emotional sphere. The possibility to collect, interpret, generate expressive emotional information opens novel frontiers to information processing and arises ethical concerns about possible dangers of such technologies with respect to intrusion in individuals' life.

(ii) Computers are more and more coupled with the environment in which they operate. Microchips are integrated in objects of our everyday life. Broadband networks allow a fast exchange of information. We are moving toward a scenario in which instead to have a "personal" computer to (usually) work with as it still happens in most of cases nowadays, highly miniaturized networked computers will be everywhere around us and will support us in most aspects of our everyday life. Similarly to what observed in (i), this perspective also opens novel and interesting possibilities, but it also arises an important debate about dangers related to any possible misuse of the possibilities technology provides.

In the design and development of multimedia systems for artistic applications the two tendencies sketched above are of paramount importance since they allow enriching artistic languages with elements that only technology can provide (e.g., the possibility to automatically analyse and generate expressive content, the possibility to create performance environments in which computing is embedded in the environment itself).
This Chapter will shortly review the basic concepts underlying the research fields that are mostly responsible of these evolutions in technology and that are of interest for the specific aims of this work.
Firstly, technologies for integrating computers in the environment will be introduced with particular reference to Mixed Reality, i.e., a corpus of research studies aiming at merging real (physical) world and virtual (computer generated) worlds in a single computer mediated environment. Researches on "ubiquitous" and "disappearing" computing and on "wearable" computing will also be shortly discussed as examples of

contributions to the development of a scenario in which computing is more and more distributed and embedded (e.g., in everyday clothes)[1].
Attention will then move on research on Affective Computing and KANSEI Information Processing that in the United States and in Japan respectively are trying to develop models and algorithms for analysis and synthesis of expressive emotional content. A short review of research works dealing with expressive content processing in the fields of interest (music, human movement, performing arts) will conclude the Chapter.

1.1. Mixed Reality

A main issue in the design of interactive multimedia systems for artistic performances is the combination or superimposition of computer generated sounds and visual media to the real environment in which a performance is taking place. In a broader scenario in which a performance can be distributed over the network the relationships between the involved real and virtual worlds assume a further particular importance.
The concept of Mixed Reality (MR) as a collection of technologies for creating mixed environments in between Virtual Environments (VEs) and the real world was firstly introduced in (Milgram and Kishino, 1994).
Mixed Reality is there defined as a "subclass of VR related technologies that involve the merging of real and virtual worlds". Depending on the relative weight of the two components (virtual environments and the real environment) in the merging process a continuum of possible scenarios is envisaged (what Milgram and Kishino refer as Reality-Virtuality continuum).
Such a continuum (see Figure 1.1) is bounded on the one side by the real environment and on the other side by virtual environments. It also includes as relevant intermediate cases Augmented Reality (AR) and Augmented Virtuality (AV).

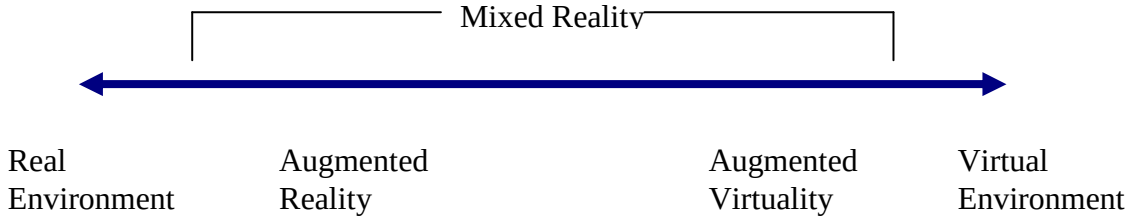


Figure 1.1: the Reality-Virtuality Continuum (Milgram et al., 1994)

The term Augmented Reality is operationally defined as referring to "any case in which an otherwise real environment is "augmented" by means of virtual (computer graphic) objects". Following the same logic, Augmented Virtuality is defined in term of

[1] It should be noticed that "ubiquitous computing" and "wearable computing" in fact start from two different points of view with respect to the role of computers in future society as it will be shortly discussed later in this Chapter.

"completely graphic display environments, either completely immersive, partially immersive or otherwise, to which video "reality" is added" (Milgram and Kishino, 1994). Ideally, the mid-point in the continuum corresponds to a situation in which reality and virtuality are so tightly mixed that it is fairly impossible to distinguish what is real from what is virtual.
If from the one hand, the work by Milgram and Kishino is very focused on visual aspects (a taxonomy of mixed reality visual displays is described in their papers), on the other hand both the dimensions of their taxonomy and some general issues about how to distinguish between real and virtual objects can be considered from a broader point of view, in the context of MR environments in which multisensory stimuli have a main role. Milgram and Kishino themselves list some augmented reality scenarios in which other modalities are involved:

- *Auditory AR*: environments in which sounds from the real world and synthetic spatialized (virtual) sounds are mixed together.
- *Haptic AR*: environments in which information related to touch and pressure is superimposed on existing haptic sensations: for example, virtual objects can be "touched" by employing special kinds of glove devices.
- *Vestibular AR*: environments in which information about acceleration of the participant's body in a virtual environment is superimposed to existing ambient gravitational forces (as, for example, in commercial and military flight simulators).

A main issue when mixing real and virtual environments is how to distinguish what is real from what is virtual. If, at a first glance, this can be thought to be quite a trivial problem, in fact it involves some subtle aspects that are worth to be shortly discussed.
Commonly, some definitions like the following ones can be assumed for real environments, virtual environments, and virtual reality:

- *Real (or Physical) Environments*: environments subjected to the usual physical laws.
- *Virtual Environments (VEs)*: computer synthesized environments that can simulate a real environment (existing or not). VEs can also go beyond the constraints of physical reality, by simulating worlds in which the usual physical laws do not hold anymore.
- *Virtual Reality (VR)*: a situation in which a participant/observer is fully immersed in a completely computer synthesized world. Such immersion is traditionally obtained by using head-mounted displays (HMD) or CAVE systems (Cruz-Neira et al., 1992).

While such definitions are usually clear and precise enough when dealing with completely real or completely virtual environments (such as in the case of VR), problems may arise in MR situations, i.e., when reality and virtuality are mixed to a certain extent.
The problem can be introduced by asking ourselves questions like: should images coming from a videocamera and then displayed on some display be considered real or virtual? Are computational representations of data coming from the real world real or virtual? Should a real object be displayed in a realistic way?
Consider, for example, an image of an environment taken by a videocamera, sent through a broadband network connection and displayed on a screen in some place at a distance of thousands of kilometres from the original location. Consider also an image of your hand taken by a videocamera and projected into a virtual environment where you can grab

synthesized objects. Which one is or should be considered "real"? Is one image more "real" than the other one?
Another example: it is straightforward to define "real" the (unprocessed) images coming from a videocamera (consider for example a videoconference situation: it is straightforward to define "real" the environment on the other side). But if the silhouettes of the participants are extracted and pasted in a virtual world (where maybe participants' actions trigger particular behaviours), the "reality" of the images is much less evident or, in any case, it seems to be a "different reality" with respect to the first situation. In any case, the real world on the other side of the network connection and the real subjects involved in the videoconference, are always the same world and the same subjects. In other words, mixing real and virtual worlds can affect our perception of reality.
Milgram and Kishino (1994) try to face such problem by proposing an objective distinction between reality and virtuality based on three aspects:

(i) A first distinction between *real and virtual objects* by means of the following operational definitions:
   a. "Real objects are any objects that have an actual objective existence" like, for example, the computer I'm using to write this document.
   b. "Virtual objects are objects that exist in essence or effect, but not formally or actually", that is, they can also be existing objects, but they do not exist here and now.

   Therefore, a real object can either be observed directly or it can be sampled and resynthesized through some display device. A virtual object, instead, cannot be directly observed since it does not exist, but it must be simulated (usually via computer graphic). To this aim, a description or a model of the object is usually needed.

(ii) A second distinction "concerns the issue of *image quality* as an aspect of reflecting reality". On the one hand, as stated above, virtual objects cannot be directly observed nor sampled: they can only be synthesized. On the other hand, technology nowadays allows synthesizing extremely realistic images. Anyway, even if an object *looks* real, this does not mean that the object actually *is* real.

(iii) A third distinction is made between *real and virtual images*. A real image is defined as "any image which has some luminosity at the location at which it appears to be located". Virtual images are conversely defined as images not having luminosity where they appear. Virtual images include holograms, mirror images, and stereoscopic displays (for which both the left and right images are real images, but not the fused image). Virtual images in MR environments are transparent, i.e., they do not occlude the objects located behind them.

In the following, while from the one hand I will try to keep the simplest distinction between real (physical) and virtual environments as far as it is possible, on the other hand when distinguishing between real and virtual objects and subjects (see Chapter 2) I will have to refer several times to the criteria mentioned above, with particular reference to the first one.
It should be noticed that criteria (ii) and (iii) as stated by Milgram and Kishino refer only to visual aspects. Anyway, they can be reformulated in a multimodal perspective for

example for distinguishing real sounds, reproductions of real sounds (with a given sound quality) and virtual (computer generated) sounds.

1.1.1. Taxonomy for Mixed Reality visual displays

Along with their definition of MR and complementary with respect to the Reality-Virtuality continuum, Milgram and Kishino (1994) developed a taxonomy for MR visual displays based on three main dimensions. As for the criteria above, such taxonomy firstly developed for visual displays, can also be reformulated and generalized to classify whole MR environments including other sensory modalities beside vision.
In this section, the taxonomy for visual displays as originally conceived by Milgram and Kishino is introduced, since even in its formulation limited to the visual channel, it anyway outlines aspects that are of particular importance for designing MR applications.
The taxonomy by Milgram and Kishino develops along three axes labelled as "Extent of world knowledge", "Reproduction Fidelity", and "Extent of Presence".
Extent of world knowledge refers to the amount of knowledge the computer system has about the world that has to be displayed. It ranges from the lack of any model (unmodelled world), such as for example in the case of direct view of a real object or of images acquired by a videocamera and directly reproduced, to a fully modelled world like in Virtual Reality where a completely virtual world can be synthesized only if a full knowledge of its objects, their locations, the point of view etc. is available to the computer system (see Figure 1.2). Interesting intermediate conditions are referred as "Where or What" and as "Where and What". In the first case, nearer to the unmodelled world side, the computer system has some information about what are the objects in the scene or about their location. In the second case, nearer to the fully modelled world, the computer system exactly knows both the essence and the location of the objects.

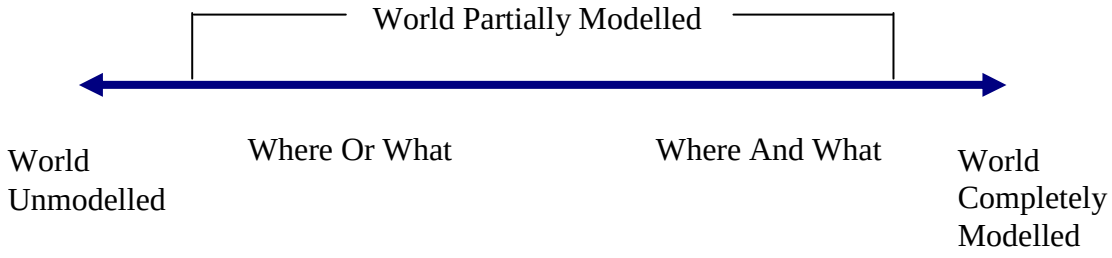


Figure 1.2: the Extent of World Knowledge dimension (Milgram et al., 1994)

Reproduction Fidelity refers to realism in Mixed Reality displays with respect to image quality. In particular, here the term Reproduction Fidelity "refers to the quality with which the synthesising display is able to reproduce the actual or intended images of the objects being displayed". Classification with respect to Reproduction Fidelity can be applied to both virtual and real objects and it is tightly related with the progression of (video) reproduction technology. The Reproduction Fidelity axis, in fact, can be

considered as a unidirectional axis showing the progression of computer graphics, modelling and rendering techniques (see Figure 1.3).

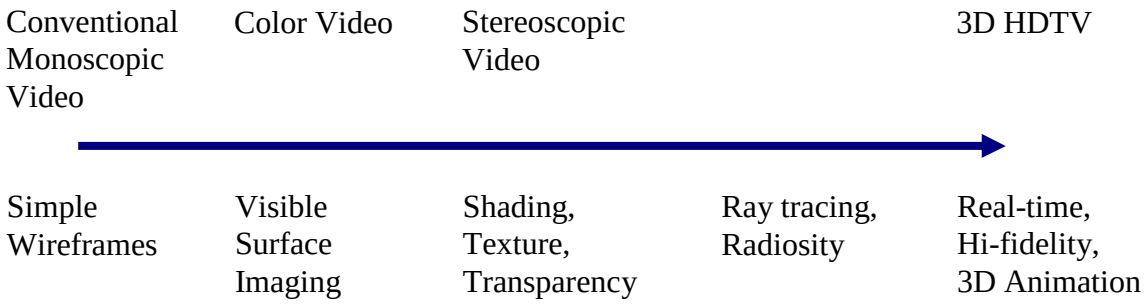


Figure 1.3: the Reproduction Fidelity dimension (Milgram et al., 1994)

Extent of Presence is related to "the extent to which the observer is intended to feel "present" within the displayed scene". While Reproduction Fidelity is more concerned with technological progress (and from a certain point of view can be considered as technology driven), this axis addresses issues more related to the paradigm of interaction/immersion (and can therefore considered as user driven). It is not perfectly orthogonal with respect to Reproduction Fidelity, but it can be distinguished from it. Like Reproduction Fidelity, Extent of Presence can also be considered unidirectional, ranging from a situation in which the user sees a virtual world through a monitor based display (something like a window on the virtual world) to "real-time imaging" (Naimark, 1991) in which ideally no differences should be noticed between the virtual/mixed world and the unmediated reality. Intermediate conditions are considered as well, mostly corresponding to the taxonomy proposed by Naimark (see Figure 1.4).

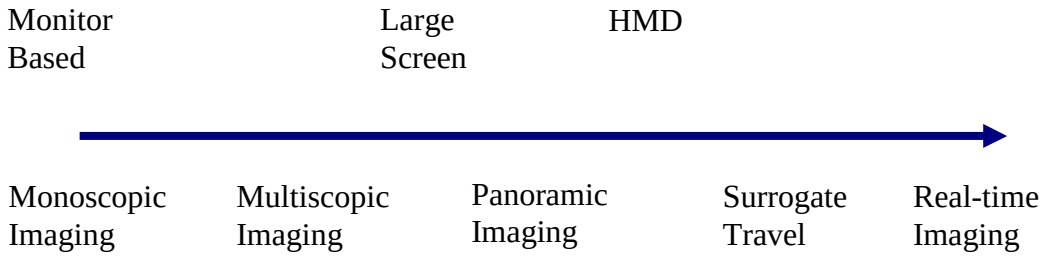


Figure 1.4: the Extent of Presence dimension (Milgram et al., 1994)

Research on issues related to the Extent of Presence dimension saw a particular grow in the interest of the scientific community in the last few years, in which projects started particularly devoted to the investigation of the mechanisms that are at the basis of presence and experience in multimedia scenarios.

1.1.2. Alternative approaches: Mixed Reality Boundaries

Benford et al. (1996, 1998) first introduced the concept of Mixed Reality Boundaries as a way to join physical and synthetic spaces.
While Milgram and Kishino's approach to MR is largely based on the idea of overlaying the virtual (synthetic) world and the real (physical) one to different extents, in the Mixed Reality Boundaries approach real and virtual worlds are kept separated by explicit (even if transparent) boundaries. In other words, while in Milgram and Kishino's approach the real and virtual worlds are overlapped, here they are "adjacent but distinct parts of a combined space". The two approaches can be complementary each other: while one can contribute in merging two specific (physical and synthetic) worlds, the other can be best suited for building larger MR structures.

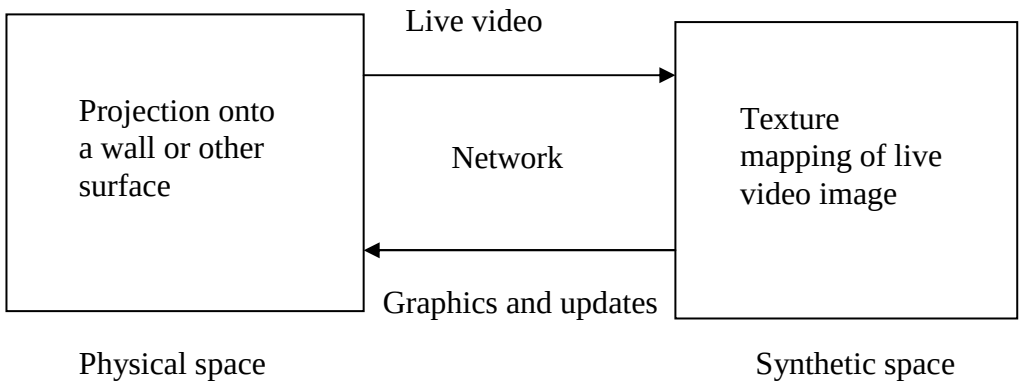


Figure 1.5: a simple Mixed Reality Boundaries scenario (Benford et al., 1998)

In its simplest instance, a Mixed Reality Boundary scenario can be described as follows (see Figure 1.5): through network connections avatars and virtual objects are projected into the physical space, whose images captured by videocameras are in turn displayed in the synthetic space by means of a dynamic texture mapping process. In such a way, the inhabitants of the physical space see the synthetic space as an extension of the physical space and vice versa. Usually an audio link is also provided to allow inhabitants of the two spaces to communicate across the boundary.
In more complex situations, the Mixed Reality Boundary approach allows to join together many physical and synthetic spaces in an integrated whole: in the scenario of interactive multimedia systems for artistic performances, this property is particularly useful when designing a system for performances distributed over the network, where several local performances taking place in several different locations (physical spaces) are joined to create a global (and virtual) distributed performance.
Benford further developed his theory of Mixed Reality Boundaries by introducing a set of properties that can be used to characterize them. Properties are grouped in three general categories of permeability, situation, and dynamics. Two meta-properties of symmetry and representation are also included.

Permeability is related to how a boundary affects the sensory information passing through it. It includes the following components: Visibility (which visual information can pass through the boundary in term both of resolution, i.e., amount of information, and field of view, i.e., the volume of space that is made visible), Audibility (which auditory information is allowed to pass through the boundary), and Solidity (the ability to traverse the boundary, e.g., metaphorically extending a limb or stepping through the boundary). Different combinations of the permeability properties generate different kinds of boundaries, ranging from analogies to common physical boundaries (like windows, curtains, walls, mirrors, lines on the ground) to completely new boundaries without a physical counterpart.
Situation "concerns the spatial relationships between the mixed reality boundary, the physical and virtual spaces that it connects and the participants and objects that these contain". It can also be divided into sub-categories: Location (how the boundary is placed in the connected spaces), Alignment (orientation of the boundary with respect to participants and objects), Mobility (whether the boundary is static or it can be moved), Segmentation (whether the boundary is made up of one or more segments), and Spatial consistency (how the spatial coordinate systems in the connected spaces are related).
Dynamics describes the temporal properties of a boundary: its Lifetime (when and how long a boundary is in existence), and its Configurability (how dynamically the boundary properties can be changed)
Symmetry is a meta-property concerning how much the properties of a boundary are similar on both its sides.
Representation is a meta-property referring to how visible are the properties of a boundary to participants and which means of representation are used.

1.1.3. Alternative approaches: tangible bits

While both the approaches by Milgram and Kishino and by Benford are focused on combining in several extents and with different methods (overlapped vs. adjacent worlds) real and virtual worlds, mainly through the visual and auditory modalities, the tangible bits approach puts strongly in evidence the physicality of the interaction in the real world as a mean to access to the virtual world (Hishii and Ullmer, 1997).
The mixing of real and virtual worlds is considered under a different perspective: rather than building an installation in which to combine real and virtual worlds and invite participants to have experience of such a combination, Hishii and Ullmer argue that nowadays everyone in his/her everyday life lives "between two realms: our physical environment and cyberspace." Nevertheless, while we commonly interact with the physical environment through haptic interaction with physical objects (like touching) and we are particularly skilled in this kind of interaction, the interaction with the cyberspace still takes place through the traditional user interfaces (like mouse, keyboard, and screen). The objective therefore becomes the design of novel user interfaces allowing to access to the cyberspace through the modalities of interaction in which we are more skilled, namely haptic interaction. As atoms are the basic components of the physical world, bits are the basic components of the virtual world (cyberspace): in this perspective, Hishii and Ullmer try to make bits tangible.

The key concepts of their approach are the following:
- *Interactive Surfaces*: each surface in the architectural space (e.g., tables, walls, windows, doors) is transformed in an active surface allowing accessing to the cyberspace.
- *Coupling of bits and atoms*: "Seamless coupling of everyday graspable objects (e.g., cards, books, models) with the digital information that pertains to them"
- *Ambient media*: sound, light, airflow, water are used as background interfaces.

In other words, "foreground bits" are made tangible by associating them to objects (and surfaces) in the real space that can be grabbed and touched; "background bits" are instead associated to peripheral ambient media (e.g., ambient sound, light).
Some system prototypes have been developed to demonstrate the tangible bits approach, see for example the metaDESK and transBOARD systems allowing users to manipulate "foreground bits" through the use of physical objects and the ambientROOM system as an example of use of ambient media for background information (Hishii and Ullmer, 1997).

1.2. Computers and environment: "ubiquitous" and "wearable" computing

Mixed Reality techniques like the ones described in the previous section allow merging physical and virtual worlds on several extents, thus obtaining environments where ideally (as a final aim) no distinction should be noticed among real and virtual objects. Such integration of reality and virtuality (i) needs the presence in the real environment of sensors (e.g., videocameras, microphones) to grab reality, (ii) leads toward a scenario where such sensors as well as processing devices are more and more integrated in objects of everyday life so that instead of having dedicated Mixed Reality installations, our whole life will take place in a Mixed Reality world (as Hishii and Ullmer already outline in their paper).
The technological and ethical concerns arisen by such a scenario have been approached from two different points of view by the research fields usually addressed as "ubiquitous computing" and "wearable computing"[2].
Ubiquitous computing can be thought to find its origin in the famous paper by Mark Weiser "The Computers for the 21st Century", where the author tries "to conceive a new way of thinking about computers in the world, one that takes into account the natural human environment and allows the computers themselves to vanish into the background" (Weiser, 1991).
Ubiquitous computing can be roughly considered as opposite to traditional Virtual Reality. In fact, while Virtual Reality creates computer-generated worlds and puts people inside them, ubiquitous computing puts computers in the real environment where people live. Anyway, in my opinion it is not opposed to Mixed Reality scenarios like the ones

[2] Notice that ubiquitous computing and wearable computing are discussed with reference to the work of researchers that originated the research fields. In fact, the interest is here more on the concepts and points of view underlying the two approaches rather than on specific (and more recent) implementations.

described by Hishii and Ullmer, rather it is useful and sometimes needed to be able to create such scenarios.

Nowadays computers need the constant attention of the user. This is frustrating for many people and often requires the work of specialized technicians. Other and older technologies are instead used in our everyday life without the need to pay attention to them or to consciously think about how to use them: see for example electric motors that are embedded in lot of devices commonly used by everyone. Ubiquitous computing can be roughly summarized as an attempt to bring computers and information technology at the same degree of disappearance[3] by providing the environment with hundreds of wireless (or wired) computer devices of all scales, ranging from little machines similar to Post-It notes, to computers of the size of a sheet of paper, to large displays similar to blackboards. Furthermore, in the ubiquitous computing approach computers have also to know where they are and the surroundings, so that they can adapt their tasks to the location in which they actually are.

The first prototypes of ubiquitous computing scenarios were developed at Xerox PARC, starting from the end of the Eighties. They mainly consisted of "tabs", "pads", and "boards". Tabs are the smallest prototypes: they are inch-scale computers similar to actual pocket calculators. For example, they can be used as active badges, allowing to track people or objects they are attached to. In this way, it is possible to envisage a world where "doors open only to the right badge wearer, rooms greet people by name, telephone calls can be automatically forwarded to wherever the recipient may be, receptionists actually know where people are, computer terminals retrieve the preferences of whoever is sitting at them, and appointment diaries write themselves" (Weiser, 1991). Of course, such a scenario (and in particular the possibility that such an amount of information could be managed by a centralized public or private institution) raises very relevant problems with respect to safeguard privacy of individuals. Weiser himself in his paper proposes some solutions ranging from the adoption of cryptographic techniques and digital pseudonyms to the possibility to build computer systems applying the same conventions as in the real world (e.g., usually information is not collected nor used against someone unless a crime is committed).

Pads are computers whose size is comparable to the size of a sheet of paper. They do not have any identity or importance, but are spread over the environment and can be grabbed and used by everyone. In fact, their use is similar to the use of the sheets of paper: they can be spread over a desk and a different task can be associated to each of them.

Boards have large displays and can be used in several environments (office, home) as video screens, blackboards, or bulletin boards. Interactions between tabs, pads, boards are also considered and possible scenarios for they joint use are envisaged. A sample day of a woman living in such a 21st century scenario as described by Weiser is a good example of how ubiquitous computers can change our everyday life.

While from the one hand, it can be thought to aim at the same objectives (i.e., seamless integrating computers in everyday life and enabling people to use them without the need of conscious attention), on the other hand "wearable computing" tries to reach its objectives by starting from a quite different point of view with respect to ubiquitous computing.

[3] In fact, research in this area is often addressed also as "disappearing computing".

In fact, instead of equipping the environment with a huge number of computers of all sizes, wearable computing proposes the development of "smart clothing", i.e., a "combination of multimedia computing, personal imaging (through the use of one or more wearable video cameras) and wireless communication" (Mann, 1996), embedded in everyday clothes and worn by users. Smart clothes embed a personal wearable multimedia computing system along with sensors (e.g., videocameras, microphones, biosensors, radar) and displays (e.g., head mounted displays). They are connected to the Internet via radio connections. They provide a broad collection of functionalities ranging from calendar (time, date), to voice communication (by replacing mobile phones), messaging, personal sound system, sound and video capture, mathematical computation, measurements.

Several application prototypes have been developed along the last twenty years and experiments have been carried out on possible concrete applications. For example, experiments in personal imaging concerned the use of the system as an extension of the visual memory, i.e., the system helps the user in remembering past situations or people by recalling images recorded in situations similar to the actual one (Mann, 1996). A straightforward application of this technique consists in recalling the names of people that the user has known.

Other experiments on personal imaging addressed the use of special "filters" generating delays, sample and hold, freeze frame effects. These demonstrated to be of particular interest since they make possible to observe aspects that are invisible to the naked eye.

A computer-assisted way-finding system has been developed able to give suggestions to the wearer about the right path to go back to a location that was already visited or to the exit (for example in a big shopping complex). The goal is achieved by taking and recording snapshots at the branch points (and if needed along the corridors); when the same path is encountered again it is recognized and, if needed, it is possible to browse back the images to come back at the starting location. A partial environment map may also be captured.

Other applications described in the same paper focused on the possibility of sharing and/or exchanging visual points of view inside a networked online community, and the possibility to get physiological measure from the smart clothes (e.g., blood pressure, heart rate, skin conductivity).

With respect to the privacy concerns regarding possible misuses of the information provided by computers systems like the ones described in the ubiquitous computing scenario, Mann arguments that in the case of wearable computing, since the computer system is strictly "personal" (i.e., it is directly worn by its user), the user is fully free to decide whether to share or not the information he/she collected. It may still happen that since the huge number of worn cameras, somebody's image is captured and spread around: Mann's conclusion is that with wearable computing "at least we'd know we had privacy when we were alone".

Ubiquitous and wearable computing open novel scenarios in which computers do not need anymore our conscious attention to accomplish their tasks. This is obtained by spreading them all around the environment and embedding them in everyday objects like our clothes. A further step toward a new generation of computer systems consists in giving them the ability to grab and process high-level information that seems to be

peculiar of human-human communication like the one related to our emotional state. Research in this direction is shortly discussed in the following Section.

1.3. The quest for expressiveness

During the last decade, lot of research effort has been spent to connect two worlds that seemed to be very distant or even antithetic: machines and emotions. Mainly in the framework of human-computer interaction an increasing interest grew up in finding ways to allow machines communicating expressive, emotional content. Such interest has been justified with the objective of an enhanced interaction between humans and machines exploiting communication channels that are typical of human-human communication and that can therefore be easier and less frustrating for users, and in particular for non technically skilled users.
Starting from the findings from psychology and neurosciences, research has been aimed at developing computational models and algorithms for analysis and synthesis of emotional content.
While from the one hand research on emotional communication found its way into more traditional fields of computer science like Artificial Intelligence, on the other hand novel fields developed explicitly focusing on such issues.
Examples are researches on Affective Computing in the United States and KANSEI Information Processing in Japan. Affective Computing and KANSEI Information Processing are shortly described with reference to the work of the two researchers that in a certain way started the two fields: Rosalind Picard and her group at MIT Media Lab for Affective Computing, and Shuji Hashimoto and his group at Waseda University, Tokyo, for KANSEI Information Processing. It has to be noticed that many other works derived from these initial studies can be found in the literature; however, here I limit myself to an overall description of the research fields since I am more concerned in describing the basic concepts and in outlining the differences in the approach rather than in writing a survey that would go in too many details of a research spread over lot of different disciplines and application scenarios. More details will be given about the applications that are nearer to the objectives of this thesis: analysis and synthesis of expressive content in performing arts, with a particular reference to music and human full-body movement.

1.3.1. Affective Computing: the American way to artificial emotions

The Affective Computing approach is mainly illustrated in the homonymous book (Picard, 1997).
In her book Picard defines Affective Computing as “computing that relates to, arises from, or deliberately influences emotions”. Affective Computing addresses the design and implementation of machines that are able (i) to recognize emotions, (ii) to express emotions, and (iii) to have emotions. These are “human-centred” machines that observe

their users and sensitively interact with them by expressing emotions depending on what they observed and on the current "emotional state" of the machine.
Computers that are able to recognize emotions are conceived as systems collecting a variety of input signals ranging from face expressions to voice, movement features (e.g., hand gestures, gait, posture), physiologic measures (e.g., respiration, electrocardiogram, blood pressure, temperature). They perform feature extraction and classification on these inputs (e.g., video analysis of movement, audio analysis of speech) and try to classify the emotion the user is communicating through a reasoning process taking into account information about "context, situations, personal goals, social display rules", and other emotion related data. Learning techniques can be employed to adapt recognition to a specific user (e.g., a personal computer can learn the habits of its master to improve its performances in the recognition task). If the computer has an emotional state, this can influence the recognition process.
Computer that are able to express emotions (either depending on instructions given by humans or as a result of an internal mechanism for generating emotions) are systems that modulate audio (e.g., synthetic voice, sound, music) and visual signals (e.g., face, posture, gait of animated creatures, colours) in a way suitable for the emotion that has to be communicated. The expressed emotion can be intentional (i.e., deliberated as a result of a reasoning process) or spontaneous (i.e., "reactively" triggered). It can directly express the "affective state" of the machine that can in turn be influenced by the expression of the emotion. Expression partially depends on social display rules.
If computers can *have* emotions is perhaps one of the most controversial issues in Affective Computing. In her book, Picard proposes to consider five components of an emotional system: a computer can be said to "have emotions" if all five components are present in it. The five components are the following:

(i) Emergent emotions and emotional behaviour, i.e., the machine is able to express an emotion through its behaviour even if it does not have any emotion. By observing the machine's behaviour, humans naturally tend to attribute an emotional state to the machine.

(ii) Fast primary emotions, i.e., mechanisms to generate a kind of hard-wired, reactive responses (especially to potentially harmful events). Fast primary emotions are what Damasio calls primary emotions (Damasio, 1994). Studies about the mechanisms triggering such emotions can be found in neurosciences: see for example (LeDoux, 1996) for a detailed description of the mechanism of fear. They are associated with the inner regions of the brain.

(iii) Cognitively generated emotions, i.e., emotions that are generated as a result of explicit reasoning. Cognitively generated emotions are slower than fast primary emotions and are usually consequence of deliberate thoughts. They are located in the brain cortex. Several cognitive models of emotion have been developed. One of the most famous is the model by Ortony, Clore, and Collins, usually referred as OCC model (Ortony, Clore, and Collins, 1988) that has been also employed in a number of concrete applications. Originally, the OCC model was not developed for building machines that could have emotions; rather it was conceived as a way for reasoning about emotions. The model develops a collection of rules associating emotions to cognitive evaluations about consequences of events, actions of agents, and aspects of objects.

(iv) Emotional experience, i.e., the system is cognitively aware of its emotional state. Emotional experience consists of cognitive awareness, physiologic awareness and subjective feelings. If it is possible to have such an emotional experience in a machine and, if yes, how it can be implemented is still an open and quite tricky issue. It relates to consciousness and requires the machine to have sensors able to measure its own "emotional state".
(v) Body-mind interactions, i.e., the emotional state can influence other processes simulating similar human physical and cognitive functions like memory, perception, decision making, learning, goals, motivations, interest, planning, etc.

Research on Affective Computing has been applied in a number of application scenarios, ranging from entertainment, to edutainment, to detection of emotional responses (e.g., frustration) in particular relevant tasks (e.g., learning, driving), to the design and implementation of devices for analysis and synthesis of emotions. Detailed descriptions of ongoing and past research projects can be found in the website of the Affective Computing group at MIT media lab (http://affect.media.mit.edu/).
With respect to the three issues mentioned above (i.e., machines recognizing, expressing, and having emotions), the work presented in this dissertation mainly address the first two aspects. That is, I'm more concerned with the design and implementation of algorithms for recognizing and communicating expressive content, rather than with machines that "have" a their own emotional state. In fact, if the goal is to open novel perspective to artistic performances by introducing new tools allowing an extension of the artistic languages by acting on the communicated expressive content through technology, what is mainly needed is (i) the possibility to classify and encode in digital format the communicated expressive content in order to process it, and (ii) the ability to produce suitable output to induce emotional reactions in spectators. In other words, in my view humans only have emotions. Machines do not need to have them, but they can give more and better support to human activities if they are able to process information not only related to the rational aspects of human behaviour, but also to the emotional ones[4].

1.3.2. The eastern approach: KANSEI Information Processing

In the same period the Affective Computing research started in the United States, another approach to understanding expressive content communication was developed in Japan: KANSEI Information Processing.
KANSEI Information Processing has been proposed as the third target of information processing (Hashimoto, 1997). In his paper Hashimoto identifies physical signals capturing data from the real world (e.g., sound, light, force) as the first target of information processing. Signal processing is the technology field that is mainly responsible of processing such kind of information. "The second phase is the semantic information processing to deal with knowledge and rule", that is the field of logic and

[4] As usual, when information is processed related to private aspects of the life of individuals (and emotions are one of the most personal and private aspects of one's life) ethical issues are concerned. These will be discussed later in this thesis.

symbolic knowledge. Artificial Intelligence is the discipline that mainly covers such aspects. The third target is KANSEI that refers to feelings, intuition, and sympathy and according to Hashimoto we are just entering in an historical period in which technology will start to deal with KANSEI, an issue that in the past was often left as a research field for only humanistic or humanistic related disciplines.
The exact meaning of the Japanese word KANSEI is something controversial for western people: it does not have a univocal correspondent in western languages and culture, but is rather associated to a collection of words related to the emotional sphere (e.g., emotion, sensibility, sensuality, sense, feeling). In his paper Hashimoto gives some examples of common uses of the word in Japanese language such as for example "Her KANSEI is excellent", "He is a man of rich KANSEI", "He has no KANSEI", "Her KANSEI seems well suited to me", etc.
It should be noticed that KANSEI refers to a dynamic process rather than to emotional labels or categories to be applied to expressive contents.
KANSEI Information Processing can be regarded as a coding and decoding process (see for example Camurri, Hashimoto, Suzuki, and Trocca, 1999). In other words, KANSEI Information Processing supposes an underlying model in which expressive content is conceived as a kind of high-level information that, in the framework of a human-human communication process, "modulates" the physical signals carrying some usually symbolic message. That is, when a (human) sender sends a message to a (human) receiver he/she encodes in the message some expressive emotional information. Such information together with the symbolic content is embedded in the physical signal carrying the message. When the receiver receives the signal he/she decodes it and extracts both the symbolic message and the additional expressive information the sender encoded into it. Notice that it is not required that the sender deliberately add the expressive information to the message: such additional expressive information can be included unconsciously and can refer to aspects such as personality traits or personal dispositions toward objects, actions, and other people.
By making a comparison with the Affective Computing approach, it can be noticed that all the three aspects of recognizing, expressing, and having emotions are included in the KANSEI process: in fact, (i) the sender *expresses* his/her emotions by encoding them in the physical signals carrying a message, (ii) the receiver *recognizes* the emotions expressed by the sender while decoding the message carried by the physical signals, and (iii) sender and receiver *have* an emotional state that can both influence the encoding/decoding process and be itself the high-level additional expressive information encoded in a message. KANSEI Information Processing seems therefore to adopt an holistic approach, broader with respect to the Affective Computing perspective (i) because it includes in the same model of encoding/decoding process all the three aspect Affective Computing separately deals with[5], and (ii) because while Affective Computing is more concerned with emotions, KANSEI rather refers to a wide collection of emotion related aspects (e.g., moods, feelings, personality traits etc.).

[5] This difference may reflect a cultural difference between western and eastern approaches to problem solving: while western people usually tend to divide a problem in sub-problems following a top-down approach and sometime losing the global perspective, eastern people often continue to keep an overall view of the problem even when they are focusing on a specific aspect of it.

In the framework of a joint collaboration between Waseda University in Tokyo and the DIST-InfoMus Lab at the University of Genova, a KANSEI Information Processing research project started also in Europe. In an attempt to translate KANSEI in western words, the term has been used to globally indicate a number of possible research targets ranging from using expressive emotional content to enhance human – computer and human – robot interaction, to understanding the communication of basic emotions to exploring the engagement of spectators exposed to musical stimuli[6] (see for example Camurri ed., 1997; Camurri et al., 2002). In his master thesis Riccardo Trocca defined a model of a KANSEI evaluation system consisting of the following components (Trocca, 2001):

(i) A *KANSEI function* mapping features of the physical signals to a space (e.g., an emotional space). This function models the interaction between the physical world and the emotional space, emulating the effects that certain physical features would have on the evoked emotional response.

(ii) An *Interpretation Function* of a point in the emotional space. For example, a function expressing the distance of a point from a set of labelled emotions in that space (e.g. in the well-known circumplex model, valence/arousal).

Such a KANSEI evaluation system has to face two main issues:

(i) The definition (or adoption) of an emotional space and the labelling of relevant points, e.g. in terms of basic emotions. Such problem has been widely faced by psychologists (e.g., see the survey in Cowie et al. 2001).

(ii) The modelling of the interpretation function. This can be based on different approaches, for example neural networks or clustering algorithms. Neural networks are often used to find non-linear relations between physical measures and the KANSEI space. An example can be found in (Suzuki and Hashimoto, 1997), focusing on sound perception, where a neural network is trained to place its output in a sort of KANSEI space.

The work presented in this dissertation has been largely influenced by the KANSEI approach since the direct participation of the author to the KANSEI research project in Genova. If from the one hand I largely agree with the encoding/decoding model delineated by the KANSEI Information Processing research, on the other hand, however, I preferred to avoid the use of the word KANSEI in this thesis because of its somewhat undefined and sometimes misused meaning.

1.3.3. Expressiveness in music and human movement

Analysis and synthesis of expressive emotional information assume a particular relevance in the context of performing arts (e.g., music, dance) whose languages are often and particularly based on and suited for conveying such information. Here I shortly

[6] I have to notice that in this attempt of "importing" KANSEI, sometimes the word has also been misused, since it has been employed as a "shortcut" to collect in just one word a huge collection of different aspects (emotion, personality, expressiveness, engagement, etc.)

discuss some main research works in this field that is the main field of interest of this dissertation.

Expression in music depends both on the structure of the composition and on the performance of the players (i.e., both composer and players contribute to expressiveness of a musical excerpt).

Eric Clarke showed the importance of the compositional structure in expression. For example in (Clarke, 1988) is stated that "expressive changes that accompany changes in performance tempo are based on structural properties of the music".

Studies on expressiveness in music performance have been carried out at the University of Padova (DEI – CSC group). Through an analysis-by-synthesis methodology a model has been derived able to synthesize an expressive performance starting from a neutral one (i.e., a performance without any expressive connotation or intention). From perceptual tests, a "Perceptual Parametric Space" has been obtained mapping expressive intentions (e.g., hard, heavy, dark, bright, light, soft) on a 2D space whose axes are related to kinetics (tempo and articulation) and energy (loudness). Given a point in the space, it is possible to calculate two sets of coefficients that applied to the neutral scores generate a performance conveying the desired expressive intention (De Poli et al., 1998; Canazza et al. 1999, 2000). The model works on scores provided as MIDI files.

In a recent work a classifier based on Bayesian Networks has been built classifying the conveyed expressive intention (soft, light, heavy, hard) on the basis of a set of measured parameters, including pitch, note number, key velocity, legato, inter-onset intervals, derived from incoming MIDI data (Cirotteau et al., 2003).

A rule-based system for generating expressive performances has been developed along many years at the Swedish Royal Institute of Technology (KTH) in Stockholm (Sundberg et al., 1991; Friberg, 1995). Rules describe how musicians deviate from the nominal score depending on their expressive intentions. They affect several aspects of the performance such as duration of tones, loudness, pitch, vibrato, crescendos and decrescendos, tempo, articulation. Each rule is also characterized by the magnitude of its effect specifying how much that rule influences the performance: for example rules can be applied in an exaggerated way.

Rules have been grouped in three different types:

- Differentiation Rules concerning the differences between scale tones (A, B, C, etc) and between note durations (quarter notes, eighth notes, etc.). Differentiation rules are related to listeners' ability to identify pitch and duration categories.
- Grouping Rules related to the ability to group together tones at several layers, ranging from tones forming melodic Gestalts, to tones belonging to the same musical phrase. The rules mark the boundaries between different groups by inserting micro-pauses and/or by lengthening the tones at the boundary.
- Ensemble Rules responsible of the synchronization of the various voices in the score by lengthening and shortening individual tones according to an overall strategy.

Rules have been implemented in the program Director Musices (Friberg et al., 2000) and performances have been synthesized to convey to listeners six different emotions (fear, anger, happiness, sadness, tenderness, and solemnity). The rules to be applied and their parameters have been selected on the basis of previous research on emotional aspects in music performance carried out by Alf Gabrielsson and Patrik Juslin at the Department of

Psychology of the University of Uppsala (e.g., see Gabrielsson, 1995; Gabrielsson and Juslin, 1996). Spectators' ratings on the synthesized performances showed that spectators were able to correctly classify the intended emotions in most cases.
Both the approaches by DEI – CSC and KTH are mainly based on an analysis-by-synthesis methodology. Attempts were also made to learn expression from examples. For example, Roberto Bresin at KTH developed neural networks able to learn KTH rules (Bresin, 1998): the input nodes of the network corresponded to the parameters present in the condition of the corresponding rule, the output nodes corresponded to the parameters that can be affected by the application of the rule.
Other relevant works are those by the research groups of Serra in Barcelona and Widmer in Vienna. The former developed a case-based system to learn expressive modifications of saxophone sounds (Arcos et al., 1998), the latter applies a collection of Artificial Intelligence techniques to study expressive music performance (see for example Dixon et al., 2002).
In the field of human full-body movement the state of the art is much less advanced than in music. Lot of researches are going on in analysis of human movement for understanding the physical mechanisms underlying it, for detecting and recognizing specific human activities (for example in video-surveillance), for analysing in details particular actions (for example gait). Very few of them are actually devoted to study expressiveness and how it is conveyed through movement. Moreover, researches on expressiveness in movement are often carried out by psychologists, with very few references to technical issues. A similar situation can be found on the synthesis side where lot of resources are spent to build characters more and more realistic and natural in their movements, but very few of them can be said to be expressive. Furthermore, both analysis and synthesis usually refer to specific actions (e.g., walking, grabbing objects) or to specific body parts (e.g., arms, hands), while full-body movement is often neglected.
A relevant exception is the work by Badler and colleagues at University of Pennsylvania. They developed EMOTE (Expressive MOTion Engine), "a 3D character animation system that allows specification of Effort and Shape parameters to modify independently defined arm and torso movements"(Chi et al., 2000). The concepts of Effort and Shape are inspired to the work of the researcher and choreographer Rudolf Laban (Laban, 1947, 1963) who developed the Theory of Effort qualitatively describing human movement in term of four main dimensions: space, time, weight, and flow. The work by Laban will be further discussed later in this dissertation since it constitutes one of the main starting points for approaching movement analysis. Laban's theories were further developed along the years by other researchers: such corpus of studies constitutes what is called "Laban Movement Analysis" (LMA) (see for example, Bartenieff and Davis, 1972). EMOTE is characterized by four main features:

- Effort and Shape parameters are independent from the geometrical definition of a movement, i.e., a gesture is specified in terms of key time and pose information and Effort and Shape parameters are applied to generate deviations with respect to the key pose information.
- Effort and Shape parameters can vary along distinct scales, e.g., each parameter can vary along a scale ranging from – 1 to + 1.
- Different Effort and Shape parameters may be specified and applied to the movement of different parts of the body (e.g., arm, torso).

- Effort and Shape parameters can be phrased across several movements (e.g., a series of coordinated movements).

While EMOTE mainly concerns synthesis of expressive body movements, a similar approach has recently been applied also to movement analysis. For example, in (Zhao, 2001) four neural networks have been trained to classify movement along the four dimensions of Laban's Theory of Effort.
Some works considering both music and movement can be found in a specific field: the analysis of the movements of a performer during the music performance. For example Eric Clarke and Jane Davidson analysed the movements of a pianist (in particular of his head) during a performance (see for example Clarke and Davidson, 1998), Sofia Dahl studied movements of a marimba player (see for example Dahl and Friberg, 2003), Marcelo Wanderley and colleagues worked on the movements of a clarinettist (see for example Wanderley, 2001).

Did the quest for expressiveness find its gold?
In my opinion, from the one hand, some positive aspects can be highlighted like, for example, the following ones:
- Research started to deal with the problem and some attempts have been done to formalize it in more precise and quantitative terms. For example, computational models of the emotional mechanisms have been developed and emotion related features have been measured.
- Some experiments demonstrated that it is possible to correctly analyse and synthesize expressive emotional content. In the field of performing arts this holds especially for music.
- Some application prototypes have been developed and used in concrete scenarios (see for example the system prototypes from MIT Affective Computing group).

On the other hand the way to expressive computers is still long and difficult:
- Research often starts from naïve hypothesis and sometimes the goals are quite ambiguous and unclear (for example, it is still not fully clear what is intended with "expressive emotional information"). A good formalization of the field (if ever possible) is still far to be reached.
- The experiments only considered very specific contexts: e.g., experiments in music are often carried out on excerpts from the classical repertoire (e.g., pieces by Mozart, Chopin, Schubert); experiments in movement often concerns very specific movements (e.g., arm and hand gestures). Their results are very far to be generalized to music and full-body movement in general.
- Application prototypes are still quite ineffective. That is, emotional machines often raise interest and curiosity (and sometimes diffidence) in people interacting with them, but in my opinion they are still not able to generate that "suspension of disbelief" that Bates (1994) considered of primary importance to make them believable.

2. Multilayered Integrated Expressive Environments

A main objective of this work is to give a scientific and technological contribution to the development of novel forms of artistic performances, where the performing action takes place in a number of physical as well as virtual connected spaces. Spectators usually become participants, since they are enabled to directly generate and modify content through their interaction. A performance can be organized on several levels of abstraction, with multiple narrative lines interleaving and interactively developing across the connected spaces. The paradigm here developed for artistic performances can further be applied to other application contexts such as, for example, museum applications.
In this chapter, a Multilayered Integrated Expressive Environment (MIEE) is envisaged in which communication mainly takes place by non-verbally conveying expressive, emotional content. Expressive gestures are addressed as first-class conveyors of such expressive information.
From a scientific point of view, attention is focused (i) on paradigms and metaphors for modeling such environments and (ii) on understanding the process of communication of expressive content, by individuating which features in an expressive gesture are responsible of such communication, and how the dynamics of these features correlates with a specific expressive content.
From a technical point of view, issues are faced on the design and development of such multilayered integrated expressive environments, in term of their hardware and software components, and with reference to possible exploitation of the technologies discussed in the previous Chapter in order to implement them.
A first step is represented by the definition of a model of such an integrated environment. The model has to take into account two main aspects[1]:

(i) The *structure* of the integrated environment, i.e., its basic components, how they are connected in the environment, and the properties of both the basic bricks and the whole environment.

(ii) The *communication process*, i.e., how information flows in the integrated environment with respect to both the interaction between environment and users, and between the basic bricks composing the environment.

In this perspective, this Chapter will discuss the structure of a Multilayered Integrated Expressive Environment and its global properties. Extended Multimodal Environments (EMEs) will be introduced as basic bricks and will be discussed in details, mainly with reference to what they contain: real and virtual objects and real and virtual subjects. EMEs will be then connected together into a network of spaces enabling geographically distributed performances. The concept of Active EMEs will be finally used to introduce MIEEs, a hierarchical structure of metaspaces each one conceived as a virtual subject collaborating in achieving the overall narrative or aesthetic goal of the performance.

[1] A third aspect, related to the dynamic along time of the envisaged environments, should also be considered. It will not directly faced in this dissertation. However, it is currently subject of ongoing research at the DIST-InfoMus Lab.

The following Chapter will deal with the second aspect: the communication processes taking place in a MIEE. A particular emphasis will be given to expressive gesture, considered as a main vehicle of information in a MIEE and as a first-class conveyor of high-level expressive, emotional content.

2.1. The basic bricks: Extended Multimodal Environments

The interactive environments discussed in this dissertation get inspiration from the Multimodal Environments (MEs) described in (Camurri and Ferrentino, 1999). MEs are conceived as "a population of physical and software agents capable of changing their reactions and their social interaction over time": the "living agents" are intended to observe the users and extract features related, for example, to motion and gesture. The extracted features are then mapped onto real-time generation of music, sound, visual media. Agents can be software agents, ranging from invisible observers to "believable characters" (Bates, 1994), as well as physical agents, namely robots moving on-stage like the Theatrical Museal Machine (Camurri and Ferrentino, 1999).
They are *multimodal* since multiple sensorial modalities (e.g., visual, auditory, haptic) are involved both with respect to perception by spectators/participants (that is, spectators/participants are exposed to stimuli activating several modalities), and with respect to analysis of inputs from spectators/participants (i.e., spectators/participants' behaviour is analysed under a multimodal perspective).
Here, the concept of ME is specified in more details from the one hand, and further extended on the other hand (i) by explicitly including humans (usually, performers and spectators/participants) in the model and (ii) by explicitly envisaging contexts in which the performance is spread over a number of distributed physical and virtual spaces together constituting a shared performing environment.
Extended Multimodal Environments (EMEs), conceived as Mixed Reality spaces containing real, virtual, and mixed objects and real, virtual, and mixed subjects, represent the basic bricks of a multilayered integrated expressive environment.
An EME can be classified in term of the Reality – Virtuality continuum (Milgram and Kishino, 1994), that is, it can be a completely real (physical) environment (as in traditional theatre performances), a completely virtual environment, or something in between: an augmented reality space, an augmented virtuality space or, ideally, a space where the user cannot distinguish what is real from what is virtual. Notice that actually the traditional distinction between physical and virtual space is here taken into account, i.e., a physical (or real) space is a space in which the usual physical laws hold, while a virtual space (environment) is one that can go beyond the constraints of physical reality, by simulating worlds in which the usual physical laws do not hold anymore.
Mixed Reality techniques like those described in the previous Chapter can be used in the design and implementation of an EME. Several approaches are possible: for example, from the point of view of Hishii and Ullmer (1997) a Mixed Reality situation already takes place when technology builds in some way a cyber-infrastructure around a single physical space. Moreover, when connecting more single Mixed Reality spaces through the network to obtain a whole integrated environment, such cyber-infrastructure is further individuated in the connection itself. Mixed Reality technologies, like methods to make

bits “tangible”, can therefore be employed both for any single EME and for building a whole integrated environment.
Here, however, a broader perspective is addressed, including the point of view by Hishii and Ullmer, but also envisaging situations in which completely virtual spaces are explicitly involved. Fully virtual spaces can be obtained by using virtual reality techniques. An EME can be an augmented reality or an augmented virtuality space; augmentation can be achieved by means of well-established augmented reality/virtuality techniques. Notice that here augmentation is intended in a multisensory perspective, that is not only concerning the visual channel, but also the auditory and possibly the haptic ones. Notice also that while from the one hand augmentation is an important aspect of EMEs and completely virtual EMEs can be considered and implemented, on the other hand, most of EMEs are real physical spaces and physicality in interaction is a main issue in the design and implementation of an EME. In my view, completely virtual environments should be used only if their use can be strongly justified: for example, because it is unpractical or dangerous to do something in a real environment, or because the designer wants to experiment a situation in which one or more physical laws do not hold anymore.
An EME can contain several kinds of entities: real objects, virtual objects, mixed objects, real subjects, virtual subjects, and mixed subjects.

2.1.1. Real, virtual, and mixed objects

An EME usually contains a number of real objects. Following the distinctions proposed by Milgram and Kishino (1994) and previously described (see Chapter 1), real objects are defined as “any objects that have an actual objective existence”. Thus, real objects are objects that effectively exist in the EME: for example, any peace of scenery can be considered as a real object, physical icons (Hishii and Ullmer, 1997) are real objects as well. Any subject actually present in a given EME can directly observe real objects and (if usable) can use them.
Conversely, virtual objects are “objects that exist in essence or effect, but not formally or actually”. This definition could be further extended since it is possible to consider virtual objects that do not correspond to any existing real object (i.e., do not exist in essence or effect), but are results of the creative imagination of the designer of a performance.
Virtual objects can be dynamically created, destroyed, used and moulded (that is, their properties can be dynamically changed over time) by subjects. Usually, they cannot be directly observed, but effects of their use can be perceived. As an example, let us consider a scenario described in (Camurri and Ferrentino, 1999): a single agent observes and interprets movements and gestures by a user (e.g., a dancer). Depending on the identified “style of movement”, a kind of “dynamic hyper-instrument” is generated and played. For example, nervous and rhythmic gestures evoking a percussionist produce a continuous transformation toward a set of virtual drums located where motion is detected. If movement evolves toward smother gestures, a continuous change takes place also in the music output: for example, virtual drums are transformed in a virtual string quartet. In the framework of the model proposed in this work, the dynamic hyper-instrument can be considered as a virtual object: it can be created in a given location, used (i.e., played), destroyed, and its properties can be dynamically changed over time.

Lot of such virtual dynamic hyper-instruments can be created in a given space. Each of them cannot be directly observed (since they are virtual), but the effects of their use (i.e., the sound produced while playing them) can be perceived.
Virtual objects are thus able to implement traditional metaphors like "hyper-instruments" (Machover, 1989), but also to go partially beyond "hyper-instruments", by enabling the dynamic behaviour previously described. Moreover, virtual objects can be employed in more complex scenarios: for example, they can implement Schaeffer's *music objects* (Schaeffer, 1977). Schaeffer's Morphology is an attempt to describe and study "concrete music": roughly speaking, in concrete music, music objects extend the traditional musical instruments with sounds coming from real life, produced by concrete objects. Virtual objects can implement Schaeffer's music objects, since they can generate sounds whose features can be changed/moulded according to Schaeffer's perceptual cues (e.g., "grain", "texture", "allure"...)[2].
The same mechanisms here described for audio can be employed also for objects whose use is perceived in visual form. Referring again to the example above, it is possible to create an object such that when the agent detects nervous and rhythmic gestures it produces a continuous transformation toward an image in which some features (e.g., colours associated to energy, sharp edges) are emphasized.
In the Milgram and Kishino's perspective, real and virtual objects can be considered as two extremes of a reality-virtuality continuum. Mixed conditions (mixed objects) are possible in between, i.e., real objects having a kind of virtual augmentation or virtual objects having a kind of physical counterpart. Moreover, objects can move along the continuum during their lifetime: thus as a consequence of an interaction moulding its properties, a real object can for example acquire and develop a virtual augmentation.
The introduction of real, virtual, and mixed subjects will definitely allow overcoming the hyper-instrument paradigm, by introducing novel interaction metaphors. By the way, notice that a "virtual subject" has already been introduced in the described example: "an agent" observes the movements of the user. Real, virtual, and mixed subjects are discussed in details in the following subsection.

2.1.2. Real, virtual, and mixed subjects

With the word "subject" is intended everything able (i) to perceive what is happening in the environment surrounding it and (ii) to act accordingly. In other words, subject is here used as synonymous of agent. An agent is in fact defined as "anything that can be viewed as perceiving its environment through sensors and acting upon the environment through effectors" (Russell and Norvig, 1995). Nevertheless, "subject" is here used since the term "agent" has been often abused in the literature in the last years.
By making for subjects a distinction similar to the one made for objects, real subjects are defined as subjects that have an actual objective existence. Humans and robots[3] are the

[2] From a certain point of view, this operation would be against the original objective of Schaeffer who aimed at extending traditional musical instruments by including sounds coming from real life. Here we would synthesize sounds having the morphological properties described by Schaeffer but that would not necessarily come from or belong to the daily experience.
[3] In fact, animals would also be real subjects on the basis of the definitions above, but it is much less common to find animals participating in artistic performances!

two kinds of subjects having an "objective existence" that are usually found in EMEs. Notice that while humans are always considered as real subjects, robots are considered as real subjects only if they are able to perceive and act, i.e., they have a certain degree of expressive autonomy. The concept of *expressive autonomy*, defined as "the amount of degrees of freedom that a director, a choreographer, a composer (or in general the author of an application including expressive content communication), leaves to the agent in order to take decisions about the appropriate expressive content in a given moment and about the way to convey it" (Camurri, Coletta, Ricchetti, and Volpe, 2000), will be further discussed later in this dissertation; at the moment, just consider an example: a small robot in a performance is used to carry a videocamera (for example, the videocamera can get images of the performers that can be deformed and projected on large screens). The robot moves strictly according to commands coming from the director. In this case, the robot does not perceive anything and its actions are just the results of the commands given by the director: it does not have any expressive autonomy and, in fact, it is a real *object*, i.e., it only is used by the director (who can be considered as a real subject). Consider instead a situation where the robot decides toward which performer the videocamera has to be pointed in a given time instant, basing its decision on the expressive gestures each performer did in the last few seconds. In this case, the robot is able to make decisions according to its perceptions, it has a certain degree of expressive autonomy, and thus it can be considered as a real subject.

Virtual subjects, instead, do not have an objective existence; they can be dynamically created and destroyed and, since they are subjects they are able to perceive and act. From the point of view of perception, virtual subjects are able to observe the environment through (real in EMEs) sensors (e.g., videocameras, microphones) and to process information in order to get an internal representation (state) of the environment. From the point of view of action, they use (real in EMEs) actuators to generate outputs (e.g., music, sound, visual media) in the environment. Similarly to what happen for virtual objects, virtual subject cannot be directly observed, but the effects of their actions can be perceived.

Again, real and virtual subjects can be considered as the extremes of a reality-virtuality continuum. Mixed subjects are envisaged in the intermediate conditions. For example, robots can be considered mixed subjects from many points of view since they have a physical existence, but virtual augmentations of their capabilities are also possible. As for objects, subjects can dynamically move along the continuum during their lifetime.

Moreover, it is also possible for objects to become subjects, i.e., one of the possibilities in dynamically changing the properties of an object is represented by allowing the object to "acquire life", that is to acquire the ability of perceiving and acting. This possibility is also related to expressive autonomy. For example, consider again the robot controlled by the director discussed above. At a certain point, e.g., as a consequence of some event, the robot could "conquer its freedom" and starting to make autonomous decision: i.e., it was a real object and it becomes a real (or mixed) subject. Conversely, a subject could also "lose its freedom" and become an object.

Subjects (mainly virtual and mixed subjects) can be classified with respect to their properties: two main criteria distinguish virtual and mixed subjects with respect to (i) the output channel that they mainly use in their actions and (ii) the predominant aspect of their behaviour, i.e., if they mainly observe, act or do both things. With respect to the second criterion, an important role is again played by expressive autonomy. In fact, if

from the one hand subjects must have a certain degree of expressive autonomy (otherwise they would be objects), on the other hand the amount of expressive autonomy strongly influences what a subject can do.
For example, with respect to the first criterion, (virtual and mixed) subjects can be distinguished in:

- *Audio subjects*, i.e., subjects that mainly use auditory output (sound and music) for their actions (or in the opposite perspective, subject the effects of whose actions are mainly perceived through the auditory channel).
- *Visual subjects*, i.e., subjects that mainly use visual output (images, lights) for their actions (or in the opposite perspective, subject the effects of whose actions are mainly perceived through the visual channel).
- *Multimodal subjects*, i.e., subjects that use both audio and visual output for their actions (or in the opposite perspective, subject the effects of whose actions are perceived through both the auditory and the visual channel).

Notice that the classification has been restricted only to audio and visual outputs since these are the outputs we are mainly concerned with. Anyway, it can be further extend if other modalities become available. If for example a subject would be able to interact through haptic effectors (e.g., devices with force feedbacks et similia), this modality can be added to the previous ones.
With respect to the second criterion, a distinction can be made between:

- *Observers*, i.e., subjects whose main role is observing a particular aspect of the environment. They extract features, interpret them, and provide other subjects with information (structured on more levels) about what they are observing. The communication process between subjects and a possible framework for analysis of features will be discussed late in the next Chapter. A particular subset of observers groups those associated to humans, that is, observers that are responsible to track and analyse the actions a human is performing. Another subset is constituted by the observers that are again associated to a human, but are responsible to observe the environment from the point of view of the human they are associated with: in a sense they are *customized observers*. Both virtual and mixed subjects and robots can play the role of observers.
- *Actors*, i.e., subjects whose main role is acting (i.e., producing music, sound, visual media) mainly depending on input received from other subjects (mainly observers). *Avatars* are an important kind of actors. An avatar is usually conceived as a representation of a human in a virtual reality environment (see for example Bahorsky, 1998). An avatar therefore acts accordingly to what the human it represents is doing: its main role is representing the human through its actions. An avatar can for example receive information from the observer that is observing the human the avatar is associated with. Both virtual and mixed subjects and robots can play the role of actors and avatars.
- *Characters*, i.e., subjects that both perceive and act. Characters often are not associated to a given human, but they can interact with humans. Characters therefore have a higher degree of expressive autonomy with respect to observers and actors. Lot of research has been carried out on characters to improve their behaviour and they believability, in a huge variety of application fields (e.g., virtual tutors, virtual assistants, characters in game environments, characters for sign language, characters

for TV applications…). Design of virtual characters is not a main objective of this work: here I just point out two issues that are particularly relevant in artistic contexts:

(i) Anthropomorphism is not a strict requirement: it is possible (and sometime preferable) to have cartoon-like characters or abstract shapes that are not anthropomorphic at all.

(ii) Despite most research on virtual characters is actually focusing on verbal communication[4], here communication mainly takes place through non-verbal channels. Characters, therefore, have to demonstrate their believability through the audio and visual output they produce.

By combining classifications according to the two criteria, two relevant cases emerges:

- *Audio clones*, that is, avatars that mainly act through audio output.
- *Visual clones*, that is, avatars that mainly act through visual output.

Clones replicate the actions of a human by translating them in auditory (audio clones) or visual (visual clones) form. The level of abstraction at which humans' actions are translated can considerably vary: for example, in a very simple scenario, some movements (or motion in a given location) can be recognized and associated with generation of audio or visual output. In more complex cases, high-level information about expressive gesture can be involved in the translation process.

Notice that a clone will need an observer to gather information about the human and an actor (an avatar) to generate audio and/or visual output. If the clone is created in the same Mixed Reality space in which the human actually is (i.e., in the same geographical location) the two aspects can be merged and the clone is in fact a character, If instead generation of output takes place in *another* Mixed Reality space (i.e., in another location), an observer will be needed in the space where the human actually is, and an actor/avatar will be needed in the space where the output has to be generated.

This last observation raises a problem that will be further discussed in the following (see Chapter 4), i.e., mapping strategies. How is the gathered information translated into actions? Mapping strategies will be discussed in details when analysing the internal structure of a virtual or mixed subject and the communication processes taking place in the discussed environments.

Once described real, virtual, and mixed subjects, two issues need to be discussed:

(i) How are real and virtual subjects involved in the design of a performance? A performance usually has some goals, that can be identified in its narrative structure and in the aesthetic concept its designer wants to communicate. How do subjects contribute to these goals? This problem is also related to the social behaviour of the subjects, that is how they relate and interact each other. Here two paradigms of interaction are addressed: *collaborative and competitive*.

(ii) How do real and virtual subjects communicate? This problem is strictly related to the previous one, since interaction is not possible without communication. The communication channels will be explored in the following Chapter. In particular, we envisage a model in which non-verbal communication takes place through expressive gestures.

[4] And related fields, such as for example automatic and believable generation of co-verbal gestures.

2.1.3. Interaction paradigms between subjects

Let's thus start considering the first issue: how do real and virtual subjects socially interact in an EME?
Here only two approaches are discussed, but further and more complex paradigms can be introduced and employed. The two approaches taken into account are quite traditional, nevertheless (i) they are easy to understand and implement and (ii) they can be usefully employed to build prototypes of EMEs. The two considered models are the following:

(i) *Collaborative models*, i.e., subjects cooperate in the fulfilment of the goals of the performance. Collaborative models have been used in a lot of application contexts, e.g., in AI and in HCI in the field of conversational agents (see for example, Guinn and Biermann, 1993; Pérez-Quinones and Sibert, 1996)

(ii) *Competitive models*, i.e., subjects compete in obtaining resources and in achieving a goal by getting the best performance or scoring. Competitive models are mostly used in games (and videogames)

Both these models have been extensively studied in several disciplines ranging from computer science (e.g., in machine learning in the field of evolutionary and genetic algorithms) to economics and social sciences. Here I avoid going into details that are beyond the scope of this work and I just put into evidence the aspects that are relevant for the context in which the models will be employed.
In the literature the term "collaborative" is often used with reference to Collaborative Virtual Environments (CVEs), intended as systems that "use VR technology to visualise a space inhabited by multiple users, usually geographically remote in the real world" (Benford et al., 1997), and provide a framework for enhancing cooperation among users finalized to a given group work (Benford et al., 1996, 1997). Taking inspiration from Benford's definition, in the context of this work "collaborative" means that subjects (e.g., performers and spectators/participants but also virtual and mixed characters) cooperate in the common group "work" consisting in generating the performance. In other words, while from the one hand the performance may remain orchestrated and supervised by its designer (composer, choreographer, director) as in more traditional contexts, on the other hand it can evolve and be moulded on the basis of joint and coordinated actions of real (performers and/or spectators), mixed and virtual subjects that can directly collaborate in generating and transforming the content. The common goal driving the participants' actions (that is implicitly implied by the term "collaborative") in the case of an artistic performance can be identified, for example, in a communicative objective of the performance as a whole, that is, in the acquired consciousness and understanding of the message the designer wants to communicate through the shared experience. Supervision by the artist/director and evolution depending on subjects' actions can be mixed at several extents: this issue is related again to the concept of autonomy (and expressive autonomy). The word "collaborative" is therefore used mainly with reference to its social meaning (i.e., bringing together people cooperating in the fulfilment of an artistic goal) rather than in its technological implications: that is, I put less emphasis on some requirements of CVEs, like the definition of "a consistent and common spatial frame of reference", or the existence of "a well established co-ordinate system in which the relative positions and orientations of different objects can be measured" (Benford, 1997).

Such approach is motivated by the fact that this work is focused on investigating new technology-based paradigms for artistic applications, rather than in reproducing with VR (more or less) real scenarios. Paradoxically, in such a context lot of interest may be raised by a VE in which physical laws are partially or completely violated, but in this situation, concepts like "well established co-ordinate system", "relative position", and "relative orientation" can lose much of their meaning.
The term "competitive" is also intended with a less specific meaning with respect to the specialized literature (e.g., in the field of genetic algorithms). Here "competitive" refers to the traditional game paradigm where players compete in achieving a goal by trying to obtain the best performance (for example in term of the best score). Competition can imply "fighting" for obtaining a limited resource, for surviving (as in most games) or, in general, for defending our own interests against the others' ones. A performance can thus be designed in a game-like perspective where subjects "fight" each against the others to get the best score and to win the game. The game paradigm can considerably raise the interest and the engagement of participants as videogames largely demonstrate. The use of well-acknowledged conventions, as in drama and games, has been demonstrated to be effective in introducing novel forms of interaction to the general public, even to novices in technology (Rinman, 2002).
The two paradigms can also be joined: for example, it is possible to have competitive environments where subjects grouped in teams collaborate in trying to win the game. In artistic scenarios this situation can often happen: for example, two actors "fighting" during a scene, in fact are contributing with their action to the overall development of the narration, i.e., with the "fighting", competitive action they collaborate to the artistic goal of the performance. The two paradigms could therefore need a redefinition in order to be employed in performing arts.

2.2. Connecting together more Extended Multimodal Environments

Up to now the discussion focused on a single EME and on what it contains (real, virtual, and mixed objects and real, virtual, and mixed subjects). An EME exists in a given geographical location. However, the interest is on a performance environment that is not limited to a specific physical and geographical location, but can be spread on several different locations. This is nowadays allowed by broadband communication technologies. Evolution in technology will also remove the need to have dedicated installations (and dedicated places) for performance, thus enabling distribution in non-traditional environments (e.g., at home). Connecting together more EMEs raises some issues about how subjects inhabiting a given environment are represented in the other ones and how subjects in a given environment can use objects belonging to another one. Such situation can be handled by using the observer and avatar subjects described above. Let's for example consider two EMEs connected through a network (see Figure 2.1). In the Figure human-like shapes[5] represent subjects (solid lines for real objects and dashed lines for virtual objects), and cubes represent objects (again, solid lines for real subjects

[5] Notice, however, that usually anthropomorphism is not needed for virtual subjects. Here they are represented as human figures only because the Figure is thus easier to understand.

and dashed lines for virtual subjects). Two observers are associated to a real subject (a human) in EME 1. The first one observes the human trying to analyse his/her actions and behaviour (e.g., what he/she is doing with a real or virtual object), similarly to the observer previously discussed in the example from (Camurri and Ferrentino, 1999).

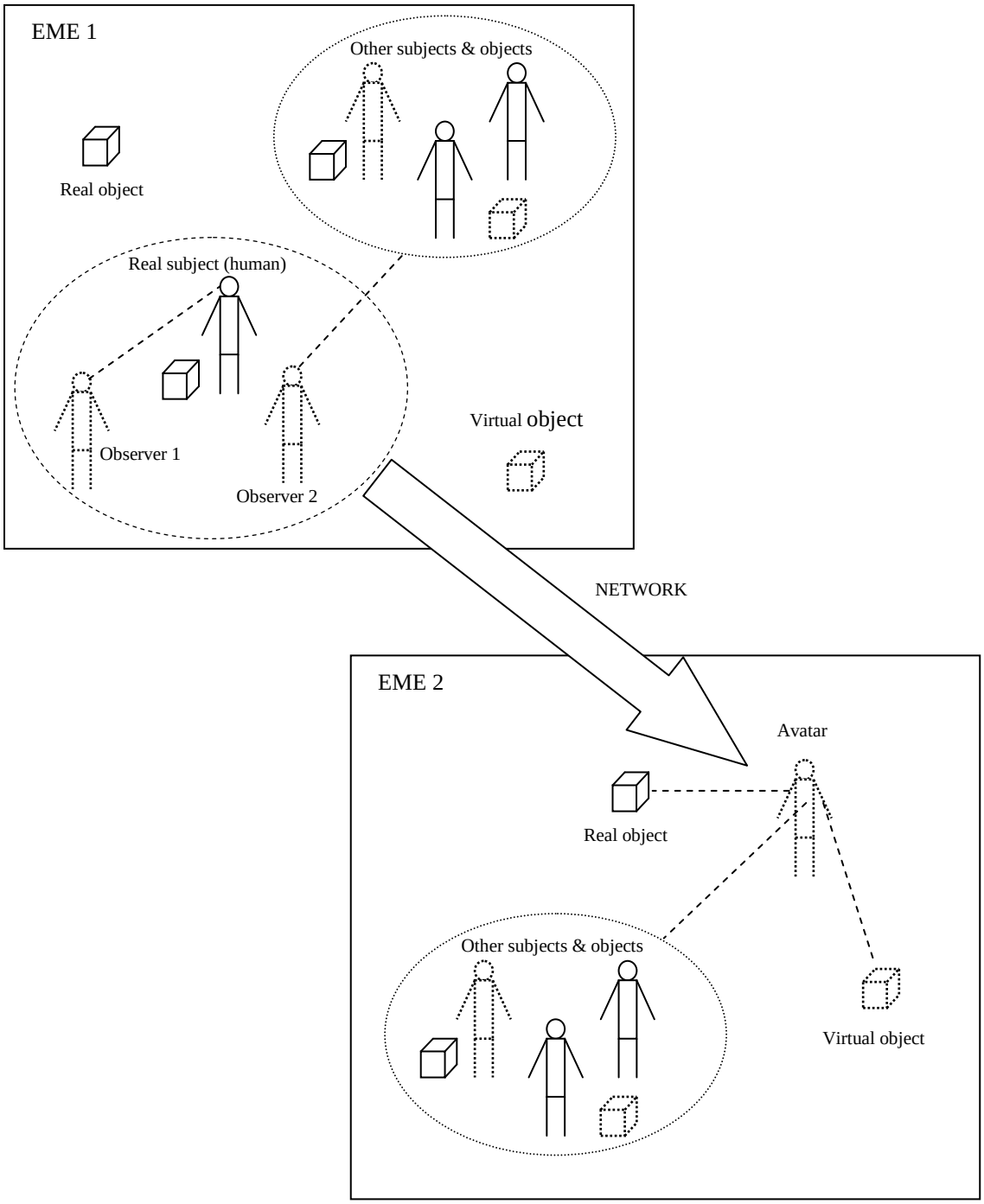


Figure 2.1: connecting two Extended Multimodal Environments.

Notice that information the observer can extract from the human ranges over multiple layers of abstraction: from simple detection of motion in given regions or of given body parts, to information about gestures the human is performing, to possible emotion the human is trying to express, to his/her engagement with respect to the performance that is taking place. Examples will be discussed later in this dissertation when, as a case study, analysis of expressive gesture in human full-body movement will be discussed.
The second observer observes what is actually happening in the Extended Multimodal Environment: for example, it observers what the other subjects (real, virtual, or mixed) are doing with real, virtual, and mixed objects. Again information over several levels of abstraction and complexity can be extracted. Notice that the second observer can be "customized" in order to observe the environment according to the preferences of the human subject it is associated with. For example, if the human subject has a particular sensitivity toward light changes or toward a given musical genre, the observer can be programmed to attribute a particular relevance to light changes and to that musical genre. The mechanisms for obtaining such a customisation will be discussed later, when the internal architecture of virtual subjects will be described in more details.
Information collect by the two observers is sent over the network to an avatar inhabiting EME 2. The avatar can thus act in EME 2 depending on what the human it represents is doing and observing in EME 1. Furthermore, the avatar can also observe what is happening in EME 2. Avatar's actions can therefore depend on (i) the actions of the human as observed by observer 1 in EME 1, (ii) what is happening in EME 1, filtered by observer 2 according to the human's preferences, (iii) what is happening in EME 2, observed and filtered by the avatar according to the human's preferences. Avatar's actions can consist in generation of audio and visual content or in suitable use (and creation/destruction, if needed) of virtual objects. The avatar could also use real or mixed objects if they can be used without the need of physically interact with them (e.g., objects that can be automatically controlled).
Conversely, information gathered by the avatar in EME 2 can be sent back to EME 1, where it can be presented to the human in several ways with increasing complexities, ranging from displays showing what is happening in EME 2 to the visual and audio feedback generated by an actor in EME 1 on the basis of data coming from EME 2.
The mechanisms here described can be replicated in order to connect together more EMEs: a network of EMEs can thus be obtained enabling distributed performances (see Figure 2.2 in the following page). Of course, complexity increases: for example, a human physically inhabiting a given EME can have avatars in each connected EME, all receiving information from the two observers associated to the human. Conversely, the human can receive feedbacks from each of his/her avatars populating the network of EMEs.
With reference to Figure 2.2, notice that a connection of each EME with any other EME is not required (i.e., the graph representing the network of EMEs may not be fully connected). However, a kind of transitive property holds on the basis of which each EME can indirectly influence what happens in any EME for which a path can be found in the graph connecting the two EMEs. Consider for example EME 5, EME 6, and EME 7 in Figure 2.2: they are not fully connected: for example, EME 5 is not directly connected with EME 7. What happens in EME 6 can depend on what is happening in EME 5 (they are connected, so there could be in EME 6 an avatar of a human living in EME 5 and acting on the basis of what it receives from EME 5). What happens in EME 7 can depend

on what is happening in EME 6 (again they are connected, and a human in EME 6 can have a avatar in EME 7 acting on the basis of what is happening in EME 6). Since what is happening in EME 7 can depend on what is happening in EME 6, but what is happening in EME 6 can depend on what is happening in EME 5, in fact EME 5 can indirectly influence what is happening in EME 7 even if a direct connection between EME 5 and EME 7 does not exist.

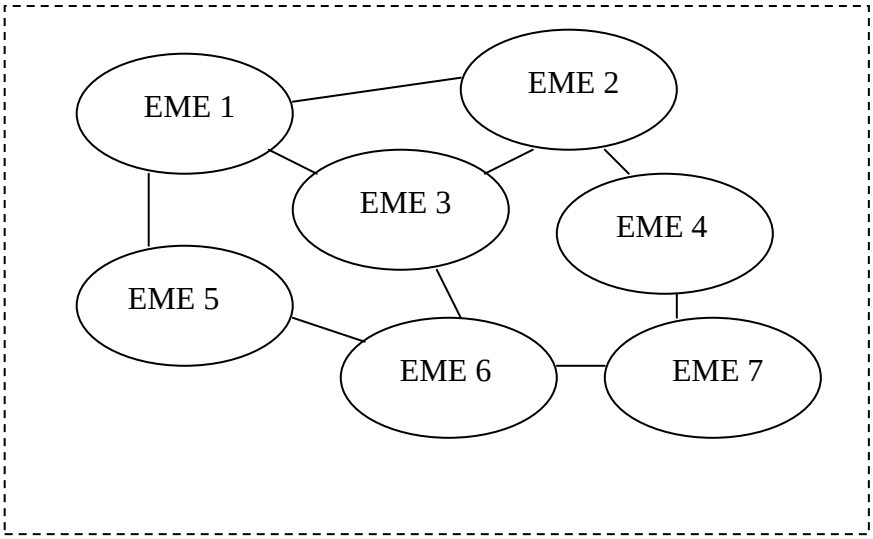


Figure 2.2: a network of Extended Multimodal Environments

Notice also that in a network of EMEs the links can dynamically change over time depending on the evolving performance. In particular, the links can be characterized by properties evolving over time (i.e., links can be parameterised) and can also be added or removed according to the needs of the performance. Moreover, in the Figure links just represent network connections among EMEs, but at a higher level a semantic can be associated with them, e.g., related to the (non-linear) narrative structure of the performance. Such aspects are not directly faced in this dissertation, but are subjects of ongoing research at the DIST-InfoMus Lab.

While techniques of augmentation such whose described by Milgram and Kishino or whose related to the tangible bits approach can be used internally to each EME composing the network, Mixed Reality Boundaries (Benford et al., 1998) can instead be a good (but not the only) choice for connection between EMEs.

Even if the complexity of an extended network of EMEs is more a theoretical condition than a practical one (in practice, usually, only few EMEs will be connected together), anyway such a complexity can make it difficult to design, organize, and coordinate a performance: the cross-influences can make it impossible to develop a narration across the EMEs and the juxtaposition of too many effects can generate situations that are both not understandable by the spectators/participants and not manageable by the director/designer. A further layer of coordination and supervising is therefore needed in multilayered integrated expressive environments.

2.3. Active Extended Multimodal Environments

An Extended Multimodal Environment can be itself equipped with sensors and effectors. Environmental sensors can be used to get an overall picture of what is happening and environmental audio and visual outputs can be generated. An EME can therefore be thought to be an *active* space, that is, it can be itself part of the performance since its environmental properties can be moulded depending on the evolution of the performance. A simple example is given by a space in which elements (e.g., lights, scenery) are dynamically changed in real-time by performers' actions. Consider for example a situation in which a concert is taking place into an EME. The EME can observe the performers and produce visual outputs (e.g., abstract shapes) depending on the played music. The same music could also be acquired through microphones, processed, and reproduced on the basis of what and how the performers play and how they move. More complex situations can also be conceived.
Active EMEs usually need to have a *state*, i.e., a corpus of information about what is actually happening and what happened in the environment.
Depending on their degree of activity, Active EMEs can be classified along a continuum, ranging from completely passive environments to highly dynamic active environments (see Figure 2.3).

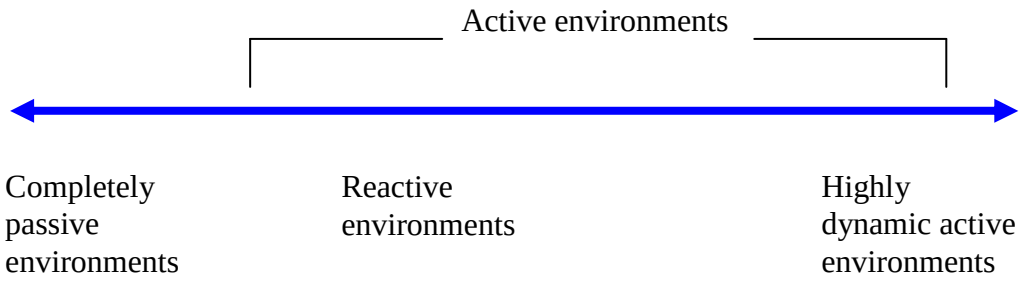


Figure 2.3: active environments can be represented along a continuum

In completely passive environments users (performers/spectators/participants) cannot influence the environment in any way. The environment constantly remains the same, or if it changes, changes are predefined. For example, this is what happens in traditional theatre scenarios, where any change in lights, scenery and so on is decided before the performance and extensively tested during rehearsals.
On the other side of the continuum, highly dynamic active environments are equipped with environmental sensors and actuators and implement complex strategies to analyse data from sensors and map them onto generation of multimedia output. Several degrees of complexity are possible for example, with respect to (i) how much memory of the past is kept and used in the mapping process and (ii) how much the mapping strategies can dynamically evolve over time.
A relevant case in between completely passive environments and highly dynamic active environments are "reactive environments" in which a collection of fixed rules is used in the mapping process.
More details will be discussed in Chapter 4 dealing with mapping strategies.

2.4. Structure of Multilayered Integrated Expressive Environments

Let's shortly reconsider the properties of an active Extended Multimodal Environment[6]: (i) it has sensors (i.e., it "perceives" what is happening inside itself through a number of environmental sensors), (ii) it has "effectors" (i.e., it is able to generate suitable multimedia content depending on what it perceived), (iii) it usually has a state (i.e., it has an internal representation of what is happening). These are the same properties that define an agent: in fact, the definition by Russel and Norvig (1995) says that an agent is "anything that can be viewed as perceiving its environment through sensors and acting upon the environment through effectors". An active EME can therefore be considered an agent whose itself is the environment and, according to the previous definitions, it can be regarded as a subject. Is it a real, virtual, or mixed subject? The problem is quite tricky.
A "real subject" is one having an "objective existence": an EME that physically exists in a given geographical location should therefore be considered as a real subject. A completely virtual environment instead should be considered as a virtual subject because it does not have an "objective existence". Anyway, as I will proceed in the discussion, the problem of understanding what in fact is real and what is virtual will become more complex, but, on a certain extent, less relevant too.

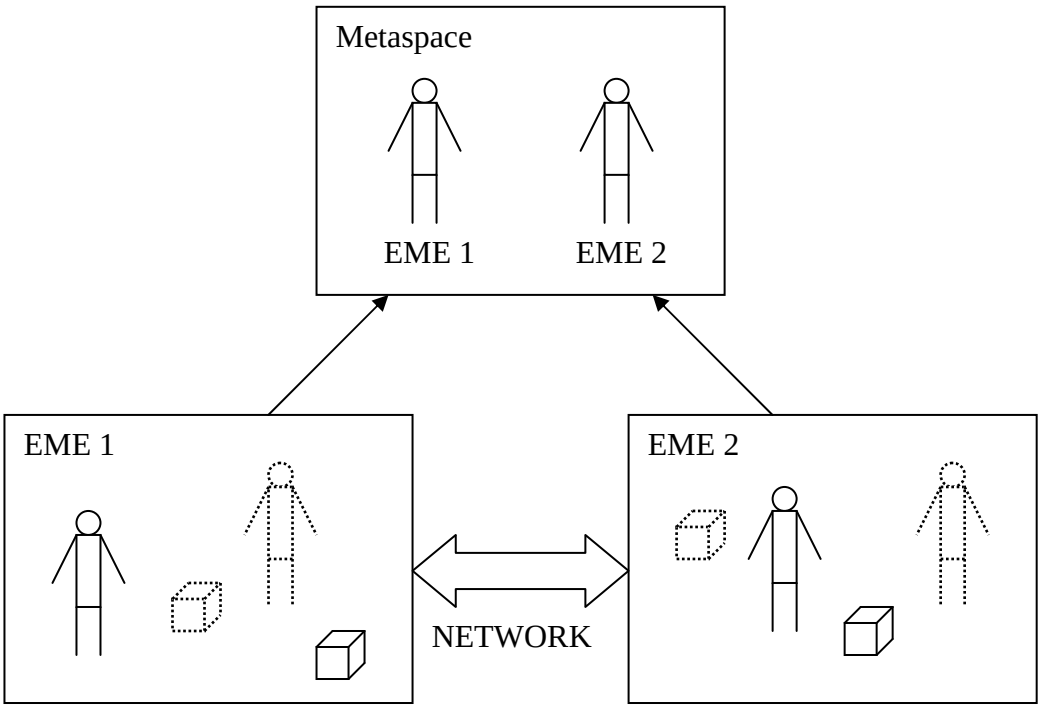


Figure 2.4: two active EMEs connected through the network can be represented as two subjects in a metaspace, one layer above the two EMEs.

[6] In this discussion I consider only active EMEs (i.e., from reactive environments to highly dynamic active environments). Completely passive environments cannot be considered like subjects since they do not perceive nor act. At most, they could be considered as objects: something about this possibility will be said later in this section.

Let's now consider two active EMEs connected through a network (a situation like the one described in Figure 2.1). As previously observed, each EME can be thought to be a subject communicating each other through the network connection. It is thus possible to define a kind of *metaspace*, one layer above the two EMEs, in which the two EMEs can be represented as communicating subjects (see Figure 2.4 in the previous page). In a similar ways, when more active EMEs are connected together like in the network in Figure 2.2, they can be modelled as a collection of subjects interacting in a metaspace one level above the network of EMEs (see Figure 2.5).

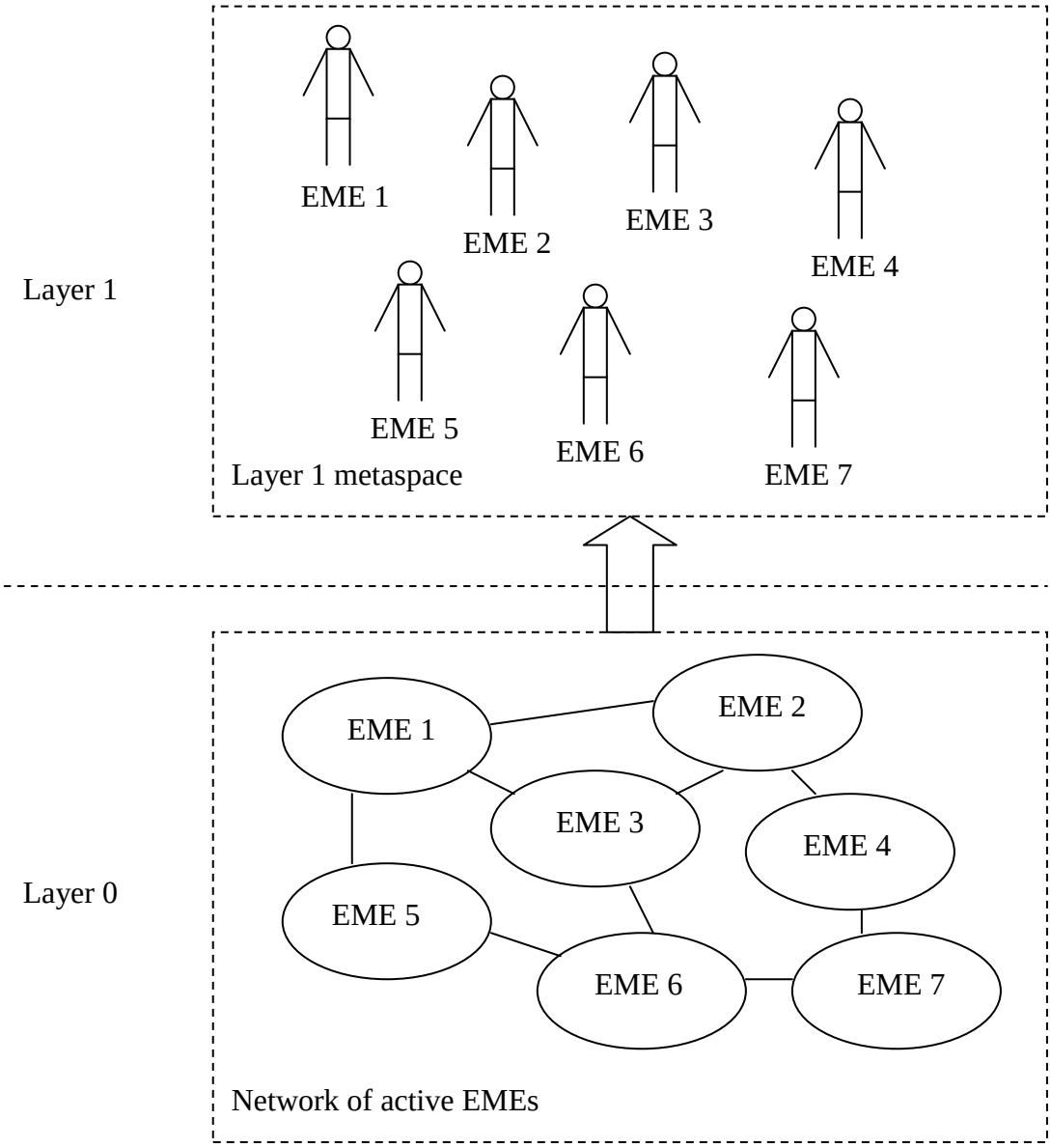


Figure 2.5: a network of EMEs can be represented as a group of subjects in a metaspace, one layer above the EMEs in the network.

As previously discussed, an EME can be directly or indirectly influenced by another EME in the network: similarly each subject/EME can have more or less knowledge about the other subjects/EMEs in the metaspace and their interaction can be more or less strong and tight.
According to this metaphor, the development of a narrative structure along the network of EMEs and the achievement of the performance's narrative and aesthetic goals can be though as the outcome of the interaction (either collaborative or competitive or both) of the subjects/EMEs in the metaspace representing the network. The EMEs can intervene and directly influence what is happening inside them with the aim of enriching the experience of the spectators/participants by controlling the complexity of the interaction, thus helping spectators/participants in understanding the performance contents and enhancing fruition.
If from the one hand each EME can be though having its own storyboard and its own "artistic goals" and real, virtual, and mixed subjects are "actors" collaborating or competing for achieving the "artistic goals" of the EME, on the other hand the metaspace at layer 1 will have its own storyboard and its own "artistic goals", but in this case each EME is an actor in the layer 1 storyboard and EMEs interact by collaborating or competing (or both) for the achievement of the "artistic goals" at layer 1.
Suppose now that two networks of EMEs generate two metaspaces in which the EMEs in the two networks are actors collaborating and/or competing in the context of the storyboard of each metaspace. The metaspace can observe what the subjects/EMEs are doing inside it and can intervene and influence their choices: in other words, the metaspace can be considered as an active environment, and therefore as a subject "perceiving" what the subjects/EMEs are doing inside it and acting accordingly[7].
The two metaspaces can then be grouped as subjects in another metaspace a layer above. The two metaspaces will be "actors" in the storyboard of the new upper level metaspace and will contribute by collaborating and/or competing to the goals of the new metaspace.
This paradigm constitutes the basic structure of Multilayered Integrated Expressive Environments (MIEE). It can be replicated recursively by creating more levels of abstraction, in which each active space or metaspace is considered as a subject in a metaspace one layer above. Each active space and metaspace has his own storyboard and subjects and as a subject itself is part of the storyboard of the metaspace one layer above it (see Figure 2.6 in the following page).
Each group of active spaces belonging to the same metaspace at the upper layer can be considered as part of the same network, i.e., they can be represented in a connected graph. For example, one of the possible "translations" in term of graphs of the MIEE in Figure 2.6 is represented in Figure 2.7. Notice that the edges inside each graph are not univocally determined by the tree structure. Consider for example the three EMEs in the bottom left corner of Figure 2.6: from the tree structure it is only possible to argue that they are connected, but it is impossible to know how they are connected (e.g., if they are fully connected or not). The representation of MIEEs in term of graphs and trees can help

[7] Notice that at this point the metaspace will be usually considered as a virtual environment and a virtual subject, since it usually will not have an objective existence. Its "perceptions" and "actions" with respect to subjects/EMEs will not be physical (like for example generation of audio/visual content in EMEs). Rather, metaspaces will act as software agents interacting with other software agents (the subjects/EMEs). Anyway, sometimes it is possible to find a physical counterpart of metaspaces as it will be described in an example later in this Chapter.

in the design and implementation phase, since traditional and well-know algorithms for traversing graphs and trees can be applied to MIEEs.

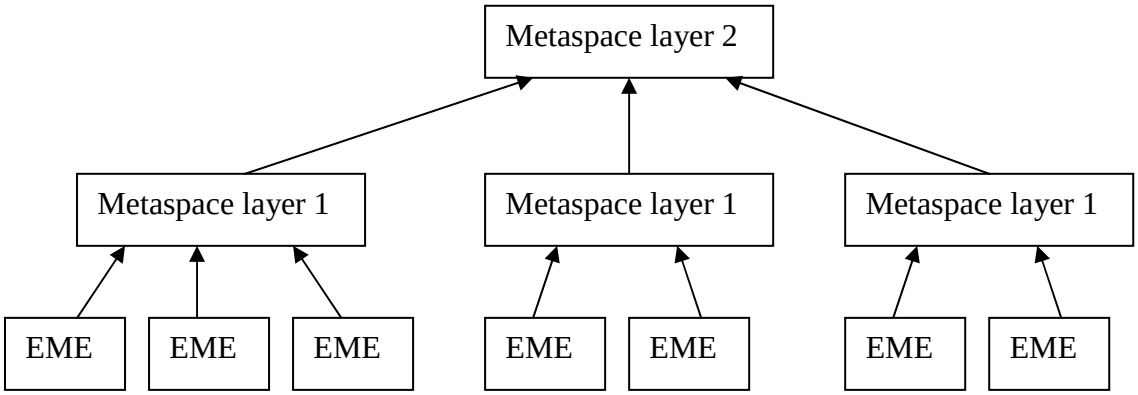


Figure 2.6: structure of a Multilayered Integrated Expressive Environment (MIEE)

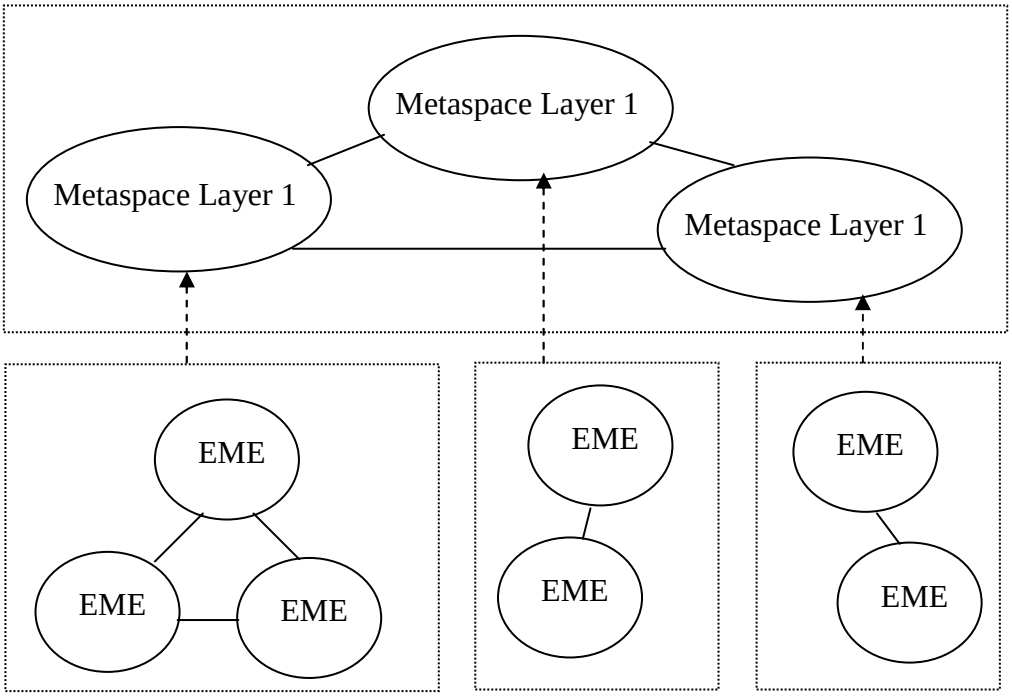


Figure 2.7: representation on term of graphs of the MIEE in Figure 2.6.

How spaces are grouped and connected depends on many aspects, and decisions about this are made during the design of the performance. For example, EMEs belonging to the same geographical region can be connected in the same network. A concrete example will be given at the end of this chapter. Moreover, as previously noticed connections at

each level should be considered as dynamic links, i.e., they could be created and destroyed, and their properties modified depending on the evolution of the performance.
MIEEs are thus *multilayered* since they represent a performance with respect to narrative structures situated at several layers of abstraction.
They are *integrated* since a number of particular aspects of the interactive performance, such as analysis of spectators/participants' behaviour, real-time generation of multimedia output, individuation and application of suitable mappings between analysed behaviour and generated output, management of the whole performance at multiple layers are all grouped and considered under the same conceptual framework.
They are *expressive* since most of the interaction and communication processes taking place inside them (both at the level of "physical" EMEs and at the level of "virtual" metaspaces) are aimed at conveying expressive, emotional content. A discussion about what is considered to be "expressive content" and about the mechanisms through which such expressive content is conveyed in MIEEs will be presented in the next Chapter, dealing with *expressive gestures*.
A final note has to be highlighted about completely passive environments. In discussing EMEs the hypothesis was made that only active environments have to be considered. Such hypothesis is needed because passive environments cannot be considered as subjects since they do not "perceive" nor "act", and therefore the metaphor of environments that like subjects interact in a higher-level metaspace cannot be applied to completely passive environments. This can be a limitation since there are many environments that are completely passive: for example, an important subclass of completely passive environments is the one of traditional theatre spaces where any modification of the environment is decided before the performance and tested during rehearsals. Sometimes, however, it is possible to import completely passive environments in the model by considering them as a special kind of *objects*. In fact, if it is possible to externally control some aspects of the environment (e.g., lights), a subject could use these mechanisms to intervene on the environment. The environment does not "perceive", nor has an internal state, but subjects can use it as an object, by intervening on it through the mechanisms the passive environment provides them.

2.5. Multilayered Integrated Expressive Environments: an example

After discussing the structure of MIEEs in term of their components and the paradigm they implement, a concrete example of possible use of MIEEs is presented to conclude the description.
Up to now, MIEEs have been discussed with respect to a scenario in which they are used to build distributed artistic performances where the narration is structured on multiple layers. The example considers another application scenario: a museum exhibit in which visitors pass through several rooms and installations again following a kind of narrative structure, the narrative structure of the exhibit, and where a goal is enhancing fruition.
Let's start by considering an installation in a room of the museum. Several degrees of complexity are possible, ranging from simply display movies and reproducing audio excerpts to interactive situations where visitors are observed, clones can be generated, audio and visual content produced in real-time depending on visitors' behaviour. The

installation can therefore be regarded as an EME, in which visitors (real subjects) are actively involved in discovering what the exhibit wants to communicate them. Real, virtual, and mixed objects and other real (e.g., robots), virtual (e.g., video and audio clones), and mixed subjects can be involved in the installation. Museum installations that singularly considered can be regarded as EMEs have been developed in several occasions: see for example the installations at "Città dei Bambini" (literally Childrens' City, a permanent science-museum exhibit for children in Genova, Italy) described in (Camurri and Coglio, 1998) and (Camurri and Ferrentino, 1999) and more recently the installations at "Città della Scienza", a permanent science-exhibit in Napoli, Italy.

A room in the museum can contain a certain number of installations connected together trough a local area network. If each installation is considered as an active EME, the room as a whole can therefore be considered as a metaspace in which subjects representing the installations contained in the room collaborate in the context of a higher-level communication objective (or a higher-level narrative structure), namely what visitors are supposed to learn by visiting that room.

Two aspects are worth to be noticed at this point. Firstly, the installations contained in the room should be active EMEs, i.e., they should be able to observe what visitors are doing, to keep and update an internal state, and to act accordingly by dynamically modifying parts of the installation. This means that a certain level of complexity is required in the installation and that the designer has to be careful in finding a good trade off between complexity and understandability when designing and implementing the installation. Simpler and sometimes passive installations can be included as objects, if they provide control mechanisms as discussed at the end of the previous section.

Secondly, this is an example in which the metaspace has a physical correspondence in the museum room. The museum room can be abstracted as an active space inhabited by subjects (the installations) interacting and collaborating toward a common goal: enhancing the fruition of the exhibit.

Let's consider now a further layer of abstraction: for example rooms in the museum can be grouped with respect to thematic areas (i.e., rooms whose installations concern similar issues can be grouped in the same thematic area). A thematic area can thus be considered as another metaspace, collocated at layer 2, inhabited by the rooms that, as subjects, collaborate in the development and in the management of the visit through a narrative path across the thematic area. The museum as a whole can be regarded as a metaspace at layer 3 where all the thematic areas, considered as subjects, interact and collaborate in managing flows of visitors inside the museum. The whole structure is shown in Figure 2.8 in the following page. More levels can be added if needed: for example, if the museum is spread over several buildings, each building can constitute another metaspace at an intermediate layer in between the thematic areas and the whole museum.

Notice that in concrete applications how EMEs and metaspaces have to be grouped in higher layer metaspaces may be quite easy to decide given the application scenario. For example, in the case of the museum here discussed grouping is performed on the basis of location (e.g., all the installations in the same room are grouped in a metaspace) and on the basis of the theme of the exhibit (e.g., all the rooms belonging to the same thematic area are grouped in a metaspace). Similar criteria can be used also in the case of artistic performances, where grouping can be depend for example on the geographical location (e.g., EMEs situated in the same region or country can be grouped together) or on the content (e.g., grouping of EMEs similar in term of storyboard or role of participants).

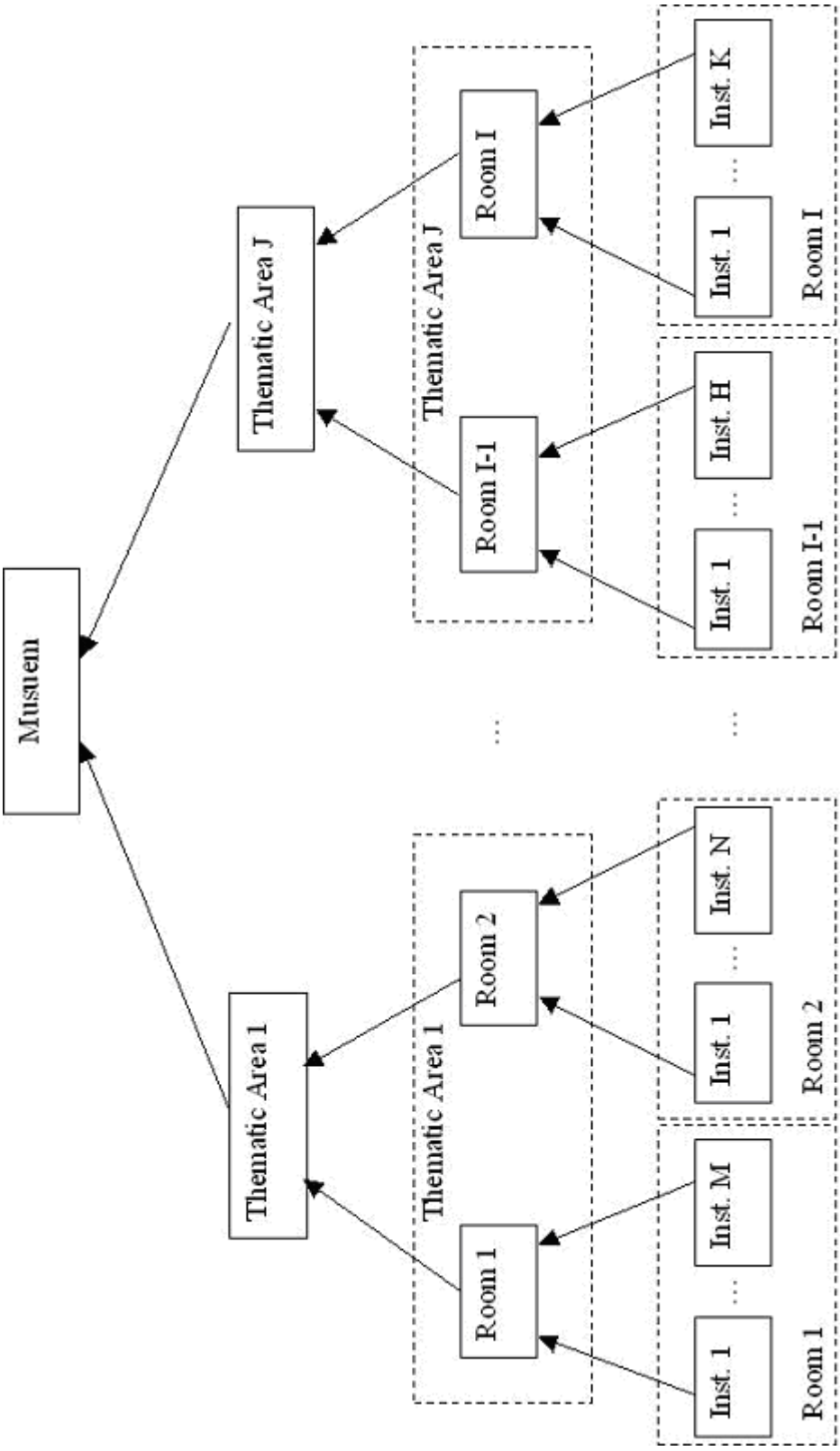


Figure 2.8: a museum modelled as a MIEE.

3. Communicating through expressive gestures

Charter 2 dealt with the structure of Multilayered Integrated Expressive Environments (MIEEs). In this Chapter attention moves on the communicative processes taking place into a MIEE. To this aim, a question already proposed in Chapter 2 is worth to be asked again: how do real, mixed, and virtual subjects communicate? The main interest in this dissertation is in investigating communication through non-verbal channels with a particular focus on music and full-body movement as first-class conveyors of expressive and emotional content. A special kind of MIEEs is therefore envisaged, where information is mainly conveyed through expressive gestures in music and movement[1].
The concept of gesture and, in particular, of expressive gesture therefore plays a key role in understanding the communication mechanisms in non-verbal MIEEs, since it provides a common conceptual framework in which it is possible to analyse the communication process under a multimodal perspective.
This Chapter starts by defining and discussing the concept of expressive gesture and collocating it in the framework of the existing literature about gesture. To make the discussion more concrete, two experiments carried out at the DIST – InfoMus Lab on particular aspects of expressive gesture will be shortly introduced. One of them will be the main subject of the second part of this dissertation.
Attention then moves on expressive gesture as main vehicle of information in non-verbal MIEEs. Three main tasks can be individuated in the communication process:

(i) Analysis of the incoming expressive gestures in order to decode their expressive content;
(ii) Mapping of the decoded expressive content, i.e., making decisions about which expressive content (if any) should be conveyed as a response to the incoming inputs and which expressive gestures are mostly suited to this task;
(iii) Synthesis of the expressive gestures deliberated in (ii).

A possible multilayered and modular architecture for virtual and mixed subjects communicating through expressive gestures is thus described, including modules for analysis, synthesis, and mapping of expressive gestures.
In particular, this Chapter discusses in details a possible structure for the analysis and synthesis components. A multilayered approach allowing multimodal analysis and synthesis of expressive gestures is presented. Moreover, some relevant features of the described architecture are discussed with particular reference to the availability of mechanisms for dynamic customisation of the architecture components in order to adapt in real-time the behaviour of a mixed or virtual subject (e.g., for adapting the behaviour of an observer or of an avatar to the real human subject it is associated with).
The following Chapter will deal with the mapping aspects.

[1] Notice that if from the one hand I now focus on non-verbal communication mechanisms, on the other hand, this does not prevent to have MIEEs in which the verbal aspects play an important role. MIEEs constitute a general paradigm for structuring integrated expressive environments where all kinds of communication are allowed. Here the focus is on non-verbal communication since I believe that it plays the most important role in the performing arts scenario.

3.1. Expressive gesture

The concept of *expressive gesture* is a key issue in this research (see for example Camurri, De Poli, Leman, and Volpe, 2001). Most of the work discussed in this dissertation refers to it. Nevertheless, both the concept of expressive gesture and its role in the communication process are still quite far to be fully understood.
This Section deals with expressive gestures under two complementary perspectives:

(i) From a conceptual point of view: after a short review of some existing definitions of gesture, a definition of expressive gesture is introduced and discussed in its different aspects and with respect to the role of expressive gesture in the communication process between real, mixed, and virtual subjects in MIEEs.

(ii) From an empirical point of view: some ongoing experiments aiming at better understanding non-verbal mechanisms of expressive and emotional communication based on expressive gesture are described. They will be further discussed later in this dissertation. In particular, one of them will constitute the reference work with respect to which the research on analysis of human full-body movement will be described in the second part of this thesis.

3.1.1. Gesture in human-human and human-machine communication

Many definitions of gesture exist in the literature. Generally, they are not in conflict with each other, since each of them focuses on different specific aspects of gesture.
For example, Kendon (1980) says that "for an action to be treated as a gesture it must have features which make it stand out as such". Two issues are relevant in this definition: (i) gesture has features characterizing it and (ii) gesture has "to stand out as such", i.e., it has to emerge, to be evident, to have a kind of unity in its components.
A main stream in literature concerns natural gesture, as a support to verbal communication. For Cassel and colleagues (1990) "a natural gesture means the types of gestures spontaneously generated by a person telling a story, speaking in public, or holding a conversation". McNeill (1992) speaks about "movements of the arms and hands which are closely synchronized with the flow of speech". He also develops a well-known taxonomy dividing the natural gestures that can be generated during a discourse in four different categories: iconic, metaphoric, deictic, and beats. The four categories can be shortly described as follows:

(i) *Iconic* gestures are the air pictures representing some aspect of the object being discussed. For example, they can refer to the shape or the spatial extent of an object.

(ii) *Metaphoric* gestures represent abstract concepts or abstract features of an object. These gestures are especially diverse, most likely because metaphors widely vary from one language or culture to another.

(iii) *Deictic* gestures are pointing motions, i.e., they identify the location of people, places and things.

(iv) *Beats* are little waves of the hand that underscore the value of speech, give accent to words, and help in speaker turn-taking.

It has to be noticed that the categories in the above taxonomy are not intended to be mutually exclusive, i.e., it is possible to have gestures belonging to more than one category at the same time. For example, metaphoric gestures are also iconic gestures. Therefore each category should rather be considered as a continuum (i.e., a gesture can have more or less iconicity, metaphoricity etc.).
Further taxonomies and classifications of gesture can be found in the literature. A summary and a comparison can be found for example in (Zhao, 2001).
Most of the qualitative gesture models (based on psychological, linguistic and cognitive studies) refer to gesture occurring with speech and supporting verbal communication (Kendon, McNeil, Rimé and Schiaratura, Krauss and Hadar: a short survey can be found in the above mentioned Zhao, 2001). McNeill further claims that gestures occur only during speech.
While not neglecting the importance of speech in human-human communication and the role of gestures in supporting speech, the attention is here rather focused on gestures occurring in *non-verbal communication* where the whole communication process is based on the informative content gestures carry. To this aim a broader definition of gesture is needed with respect to the ones mentioned above, a definition taking into account gestures that are not directly associated with speech and conversation.
In this broader perspective, Kurtenbach and Hulteen (1990) define gesture as "a movement of the body that contains information". The fact that gestures have to contain information is important for distinguishing gestures from other movements. For example, Kurtenbach and Hulteen do not consider the act of pressing a button (or a key) as a gesture, since motion does not have any meaning or information associated with it.
The definition by Hummels, Smets, and Overbeeke (1998) goes in the same direction: "a gesture is a movement of one's body that conveys meaning to oneself or to a partner in communication". (Wachsmuth, 1999) says: "for the purpose of this paper it is sufficient to understand "gesture" as body movements which convey information that is in some way meaningful to a recipient". A survey and a discussion of existing definition of gesture can be found in (Cadoz and Wanderley, 2000).
As I will discuss in the following section, the fact that gesture is intended to convey information is a key aspect for defining expressive gesture: in fact, expressive gesture will be distinguished from other kinds of gesture depending on the kind of information it convey, i.e., expressive content.

3.1.2. Gesture in artistic contexts: expressive gesture and expressive content

In artistic contexts and in particular in the field of performing arts, gestures are often not intended to denote things or to support speech as in the traditional framework of natural gesture, but the information they contain and convey is related to the affective/emotional domain. In some specific domains, gestures and their content are strictly codified and stylised as for example in ballet, but in most cases the emotional information does not depend on a defined code, but rather it is associated to dynamically time varying features.
In most of the definition cited above gestures explicitly or implicitly are intended to carry and convey some kind of information. Such a property can be fruitfully used to distinguish and define expressive gestures. In such a perspective, gestures can be

therefore considered "expressive" depending on the kind of information they convey: expressive gestures carry what Cowie et al. (2001) call "implicit messages", and what Hashimoto (1997) calls KANSEI (see also Chapter 1).
That is, they are responsible of the communication of a kind of information (addressed as *expressive content*) that is different and independent, even if often superimposed, to a possible denotative meaning, and that concerns aspects related to feelings, moods, affect, emotional intentions[2].
Let's consider an example: the same action (i.e., the same body movement) can be performed in several ways, by stressing different qualities of movement: it is possible to recognize a person from the way he/she walks, but it is also possible to get information about the emotional state of a person by looking at his/her gait, e.g., if he/she is angry, sad, happy. In the case of gait analysis, we can therefore distinguish among several objectives and layers of analysis: a first one aiming at describing the physical features of the movement, for example in order to classify it (quite a lot of research work can be found in the computer vision literature about gait analysis, see for example Liu et al., 2002); a second one aiming at extracting the expressive content gait coveys, e.g., in terms of information about the emotional state the walker communicate through his/her way of walking. From this point of view, walking can be considered as an expressive gesture: even if no denotative meaning is associated with it, it however communicates information about the emotional state of the walker, i.e., it conveys a specific expressive content. In fact, in this perspective the walking action fully satisfies the conditions stated in the definition of gesture by Kurtenbach and Hulteen (1990): walking is "a movement of the body that contains information".
Moreover, the example of gait outlines another relevant issue, that is more layers of processing are needed to extract and analyse the information contained in and conveyed by a gesture. This consideration will lead in the following to the definition of a multi-layered conceptual framework for analysis and synthesis of expressive gesture.
In the case of the walking action, the expressive gesture (usually) does not have any denotative meaning. This is the most common situation when considering an artistic scenario. It is also possible, however, to consider expressive gestures having a precise semantics, not only in the affective/emotional domain, but also because they are intended to denote things in the outer world. For example, even an iconic, a metaphoric, or a deictic gesture can convey an expressive content through the way in which it is performed. In that case, the expressive content is conveyed in parallel or superimposed to the symbolic meaning. In some cases, the expressive content could also partially or totally modify the intended meaning of a gesture. In fact, the mechanism can be considered in some extent similar to inflections in speech, where a particular inflection (often used to communicate an expressive content) can give a particular meaning to the discourse, sometimes partially or totally different from the original meaning of the words that have been pronounced.

[2] Indeed, it should be noted that in the common meaning of the word "gesture" as reported by dictionaries (some dictionary definitions are for example reviewed in the cited paper by Cadoz and Wanderley), gesture is often defined with reference to expression of feeling and emotion. While from the one hand the scientific definition of gesture as carrier of information gives a broader meaning to the word, on the other hand the definition of expressive gesture (although derived from the definition of gesture) come back closer to its common meaning.

With respect to the main stream in literature of natural gestures and co-verbal gestures, it can be said that if on the one hand, expressive gestures partially include natural gestures, that is, natural gestures can also be expressive gestures, on the other hand a more general concept of expressive gesture is here faced, including not only natural gestures but also musical, human movement, visual (e.g., computer animated) gestures.
Moreover, the concept of expressive gesture here discussed is also somewhat broader than the general concept of gesture as defined by Kurtenbach and Hulteen, since it also considers cases in which, with the aid of technology, communication of expressive content takes place even without an explicit movement of the body, or, at least, the movement of the body is only indirectly involved in the communication process. For example, this can be the case in MIEEs, where expressive content can be conveyed through a continuum of possible ways ranging from realistic to abstract images, sounds and effects: cinematography, cartoons, computer animated characters and avatars, expressive control of lights in a theatre context (maybe in relation with an actor's physical gestures), expressive musical performances, expressive use of sound. Consider, for example, a MIEE in which a theatre performance is taking place: the director, choreographer, composer can ask actors, dancers, musicians, to communicate content through a number of expressive gestures (e.g., dance and/or music phrases, postures, sentences). At the same time, technology allows the director to extend the language he can employ. For example, he can map motion or music features onto particular configurations of lights, in movements of virtual characters, in automatically generated computer music and live electronics. In this way, he can create "extended" expressive gestures that, while still having the purpose of communicating expressive content, are only partially related to explicit body movements: in a way, such "extended expressive gestures" are the result of a juxtaposition of several dance, music, visual gestures, but they are not just the sum of them, since they also include the artistic point of view of the director who created them, and are perceived as multimodal stimuli by human subjects (e.g., spectators).
MIEEs are thus a natural test-bed in which non-verbal communication by means of extended expressive gestures can be studied in all its aspects of analysis, synthesis, and mapping.
The research on expressive gesture here discussed is finalized to two main objectives:

(i) Understanding the mechanisms underlying communication of expressive content through extended expressive gestures (e.g., which features are important, how they can be measured, how features are related to expressive content)
(ii) Developing novel interactive multimedia scenarios (e.g., MIEEs) in which automatic systems enable novel interaction paradigms and allow a deeper engagement of the user, by explicitly observing and processing the expressive gestures the user performs.

In this perspective, MIEEs constitute both environments in which experimenting paradigms of expressive gesture communication and a conceptual platform for designing and developing novel multimedia interactive systems.
Besides giving a definition of expressive gesture, it is needed to empirically study it in order to understand how expressive content is conveyed: a couple of experiments investigating specific aspects of expressive gesture are now introduced.

3.1.3. Experiments on expressive gesture

A first step in the investigation of expressive gesture consists in setting up experiments aiming at individuating the main mechanisms supporting the communication process. Since (i) artistic performances strongly use in their languages such non-verbal communication mechanisms to convey expressive content, and (ii) there is a particular interest in developing expressive gesture applications for artistic scenarios, this research focused on performing arts, namely on dance and music performances, as a test-bed where computational models of expressive gesture and algorithms for expressive gesture processing can be developed, studied, and tested.
As an attempt to shed light in the communication process of expressive content through expressive gestures in artistic scenarios, the attention has been focused to the following two particular aspects:
(i) Expressive gesture as a way to convey a particular emotion to the audience;
(ii) Expressive gesture as a way to emotionally engage the audience.

Each of them has been subject of experiments carried out at the DIST - InfoMus Lab in collaboration with its partners in national and European projects (mainly the EU-IST Project MEGA - www.megaproject.org - in the context of which lot of the work described in this dissertation is collocated).
The ability of expressive gestures to convey emotions has been studied in an experiment carried out in collaboration with the Department of Psychology of the University of Uppsala (Sweden). The experiment considered an archive of dance performances and had the purpose of (i) individuating which motion cues are mostly involved in conveying the dancer's expressive intentions (in term of basic emotions) to the audience during a dance performance and (ii) testing the developed models and algorithms by comparing their performances with spectators' ratings of the same dance fragments. This experiment will be discussed in details in the second part of this dissertation.
The second aspect was investigated trough an experiment aiming at understanding the mechanisms that are responsible of the audience's engagement in a music performance. Spectators exposed to a performance by a professional pianist have been asked to rate with continuous measures their emotional engagement. A statistical analysis has been performed on the spectators' ratings and on a collection of audio and motion cues automatically extracted from the audio and video recordings of the piano performance. Some preliminary results of this second (still ongoing) experiment will be discussed in the conclusions and are reported in (Camurri, Mazzarino, Timmers, Volpe, 2003, and Timmers, Camurri, Volpe, 2003).
If from a scientific point of view these experiments are a starting point for understanding expressive communication, from a technical point of view they constitute the scientific basis on which the design of interactive multimedia systems can be ground.
For example, in the case of MIEEs the experiments give hints on the way in which the input and output components of a virtual or mixed subject should be built. Consider for example a virtual subject having a role of observer: it is responsible of obtaining information about what the user it observes is doing. As already mentioned such information is located and processed along several layers. If the aim is to analyse user's

expressive gestures and to decode the expressive content associated with them, the experiments can help in individuating which features the observer should be able to extract from the user's behaviour and how it should process them[3].

3.2. Virtual and mixed subjects communicating through expressive gestures

Since in a non-verbal MIEE communication mainly takes place through expressive gestures, virtual and mixed subjects have to be endowed with techniques for expressive gesture analysis, mapping, and synthesis. An important objective is the definition of a general architecture that can be considered as a common basis on which virtual and mixed subjects with different tasks and skills can be built. Such a general architecture (i) has to be modular so that different kinds of virtual and mixed subjects can be obtained by replacing components, (ii) must allow multimodal processing of expressive gesture (i.e., virtual and mixed subjects must be able to deal with gestures affecting several channels of perception), (iii) must allow customisation and adaptation of virtual and mixed subjects to real subjects and the environment (e.g., it has to be possible to adapt an observer or an avatar to the human it is associated with).

3.2.1. The "Emotional Agent" architecture

The architecture presented in this Section finds its basis in the "Emotional Agent" architecture developed over the years at the DIST - InfoMus Lab and described in (Camurri and Coglio, 1998). The architecture as originally conceived by its authors is sketched in Figure 3.1.
Five active components can be individuated in it: input, output, rational, emotional, and reactive component. The white and thick arrows represent flows of information in the architecture (e.g., the white arrows connecting the input, rational and reactive components to the emotional component represent emotional stimuli the emotional component receives from the other ones). From a software engineering point of view, such arrows are implemented as buffers on which one component acts as producer and the other as consumer. The black and thin arrows represent parameters affecting the way in which components work (e.g., the two arrows connecting the rational and emotional components are the channels through which rational processing can influence the emotional state of the agent and vice versa). They are implemented as data containers upon which one component has read and write access and the other read only access.
The two dashed arrows represent the information flow from and to the outer world.
The architecture Emotion Agent can be collocated in the literature as an attempt to give both a common structure and some software engineering guidelines to the design and

[3] Of course such experiments are just a starting point: they only consider limited and specific aspects of expressive gesture. It will be therefore very difficult to build an observer able to properly do its tasks in the infinity of situations that can happen in the real world. In my opinion, they should therefore be considered as a first step of a very very long way (someone could say luckily…)

implementation of agents for Multiomodal Environments (MEs). The attention is thus focused on design issues rather than on modelling of biologic mechanisms. In this sense, it differs from other works (see for example Sloman, 1998) where the aim is to model and understand human or animal behaviour.

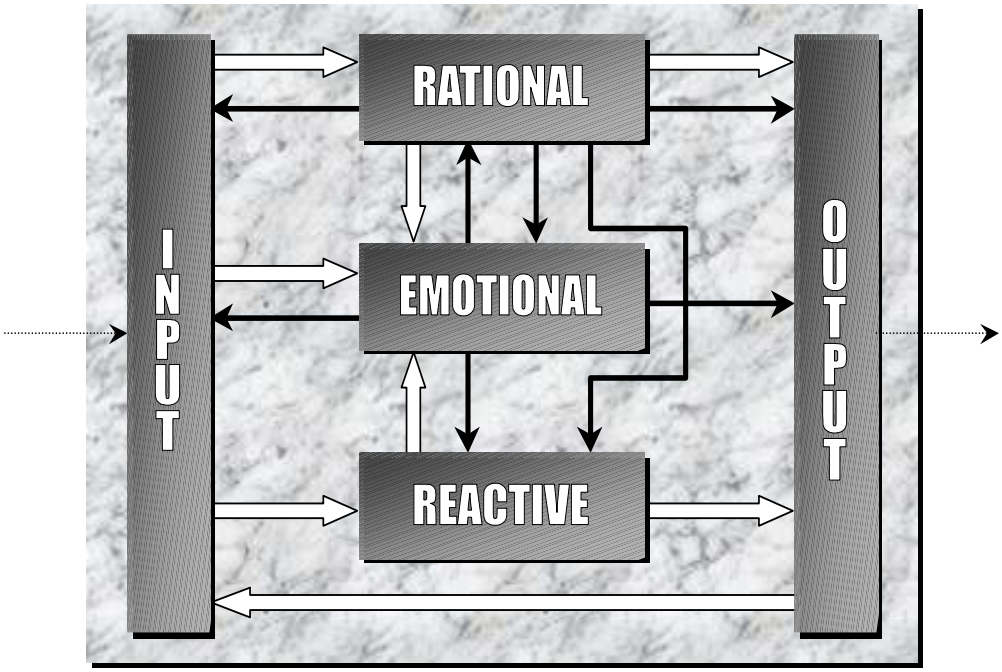


Figure 3.1: the architecture"Emotional Agent" (Camurri and Coglio, 1998)

The architecture Emotional Agent has been used in concrete application scenarios (for example in the permanent exhibit for children "La Città dei Bambini" in Genova as described in Camurri and Coglio, 1998) and its rational and emotional components have been further worked out.
For example, in (Camurri and Ferrentino, 1999) a model of artificial emotions is described that can be employed within the emotional component[4].
In (Camurri and Volpe, 1999, and Volpe, 1999) a possible structure for the rational component is envisaged: a traditional AI production system for automatic reasoning has been endowed with the capability to deal with expressive information. In that implementation of the rational component expressive information could affect several aspects of the rational processing:

(i) The evolution over time of the agent's knowledge about the outer word built on the basis of the inputs it receives.
(ii) The decision making process concerning the selection of the most suitable actions to be accomplished to reach the agent's goals.
(iii) The decision making process concerning the selection of the agent's goals.

[4] This is just an example; more traditional models like the Ortony, Clore, and Collins (OCC) model (Ortony, Clore, and Collins, 1988) may also be employed

The rational component was also able to influence the emotional state of the agent. In particular, the outcome of the agent's actions and success or failure in fulfilling its goals represented positive and negative stimuli for the emotional component.

3.2.2. Internal structure of a virtual and mixed subject

The "Emotional Agent" architecture arises some issues that are worth to be shortly discussed. In the Affective Computing approach (see Chapter 1) three main aspects are considered: (i) machines *recognizing* emotions (ii) machines *expressing* emotions, and (iii) machines *having* emotions. As already explained in Chapter 1 the objectives of the work presented in this dissertation are related to aspects (i) and (ii) since the interest is on extending the artistic languages by acting on the communicated expressive content through technology. To this aim, technology has to provide (i) the possibility to classify and encode in digital format the communicated expressive content in order to process it, and (ii) the ability to produce suitable output to induce emotional reactions in spectators. That is, it is not needed that machines have emotions, humans have them and technology can help the artist in conveying to his/her audience the expressive content he/she wants to convey: the point is how technology can help in communicating emotions and not how machines can feel emotions[5].
However, the adoption of this perspective requires rethinking the Emotional Agent architecture. In fact, the scenario is now the following: a virtual or mixed subject (that could be implemented using the Emotional Agent architecture) observes the expressive gestures through which other subjects try to communicate with it; it processes such gestures, and in turn generates expressive gestures to convey expressive content to the other subjects. Such expressive content can produce several different responses in the receivers, ranging from shifts of attention, to increased engagement, to eliciting of specific emotions. The main tasks can therefore be identified as follows:

(i) Analysis of the incoming expressive gestures in order to decode the expressive content they convey;
(ii) Mapping of the decoded expressive content onto suitable outputs, that is making decisions about if it is needed to answer to the incoming inputs, what expressive content should be expressed, and how it should be expressed;
(iii) Synthesis of expressive gestures to convey the expressive content decided in (ii)

In this way the virtual or mixed subject does not have an emotional state, rather it has an expressive content to communicate. That is, the virtual subject does not have an emotional state representing the emotional stimuli it received and a given personality, but it decides which expressive content is most suitable in the current conditions. If from the one hand, such expressive content may depend on incoming emotional information (e.g., the virtual subject could recognize an emotion in the expressive gestures of other

[5] In other words, the aim is not to create machines able to replace humans also with respect to emotional aspects, rather the focus is on how machines can help humans in express themselves: in this perspective this research is "human-driven", i.e., humans are the source of requirements for designing machines capable to support them rather than to replace them.

subjects) and on the personality traits the virtual or mixed subject wants to show, on the other hand the virtual or mixed subject explicitly makes decisions about the expressive content and how to communicate it, i.e., it is more concerned with rational processing rather than with "emotional processing". In other words, the mapping task in (ii) can be assigned to the rational component of the Emotional Agent architecture without the need of explicitly including an emotional component. An explicit emotional component would be needed for virtual subjects having ("feeling") emotions: in this case, the emotional component would contain an emotional state corresponding to the felt or perceived emotion, while the rational component would be responsible of the "consciously" deliberated emotions. The actual output would consist in some kind of juxtaposition of the two aspects. But since what has been said above this possibility tends to be excluded. Moreover, if the emotional component were kept, the only way the rational component would have to decide the expression of a given expressive content would consist in influencing the emotional component. But in this way the rational component would not have any guarantee to succeed in the task, since success would depend on the internal mechanisms of the emotional component. Of course, this is exactly what happens in humans, i.e., it is often difficult to simulate an emotion (cognitively deliberated) when another emotional state is actually present, but it has to be remembered that here the aim is not to model as precisely as possible human behaviour, rather the focus is on designing and implementing subjects able to analyse and convey suitable expressive content[6].

These observations lead to a revision of the way in which virtual and mixed subjects are structured. In particular, their architecture reflects the tasks they are responsible of (i.e., expressive gesture analysis, mapping, and synthesis). A first step consists therefore in roughly identifying the input component of the Emotional Agent architecture with the analysis process, the output component with the synthesis process and the three intermediate components with mapping. Figure 3.2 in the following page shows this correspondence with some details for the mapping component. The Expressive Gesture Analysis and Synthesis components will be discussed in the next Section, the mapping component in the next Chapter.

The white and tick arrows again represent flows of information, while the black and thin arrows represent influences that a component exerts on another one.

It should be noticed however that the architecture in Figure 3.2 should be considered more as a conceptual framework rather than as a software engineering design (as the Emotional Agent architecture partially was), although some guidelines for its software implementation could be given. For example, in a possible implementation an approach similar to the one employed in the Emotional Agent architecture can be considered: thus white arrows can be again buffers on which one component acts as producer and the other one as consumer, while black arrows can be implemented as data containers upon which one component has read and write access and the other read only access.

The second step consists in going inside each component and in analysing how information flows and is processed. It should be noticed that if from the one hand the first step could be considered as a simplification (maybe excessive but needed from a certain point of view) of the original Emotional Agent architecture, on the other hand the

[6] Of course, this does not mean that I am not at all interested in modeling human emotional mechanisms; this also is an interesting problem, but it is another research issue that is not the subject of this dissertation.

second step leads to a more detailed (but still general enough) analysis of the problem. The Emotional Agent architecture does not make any commitment about the internal structure of its components, even if some possible scenarios are envisaged. Here instead a possible internal structure (especially for the input and output components) is discussed that, while being general enough to build a wide variety of different and customisable virtual and mixed subjects, is at the same time detailed enough to allow to organize analysis and synthesis of expressive gesture in a unified conceptual framework under a multimodal perspective.

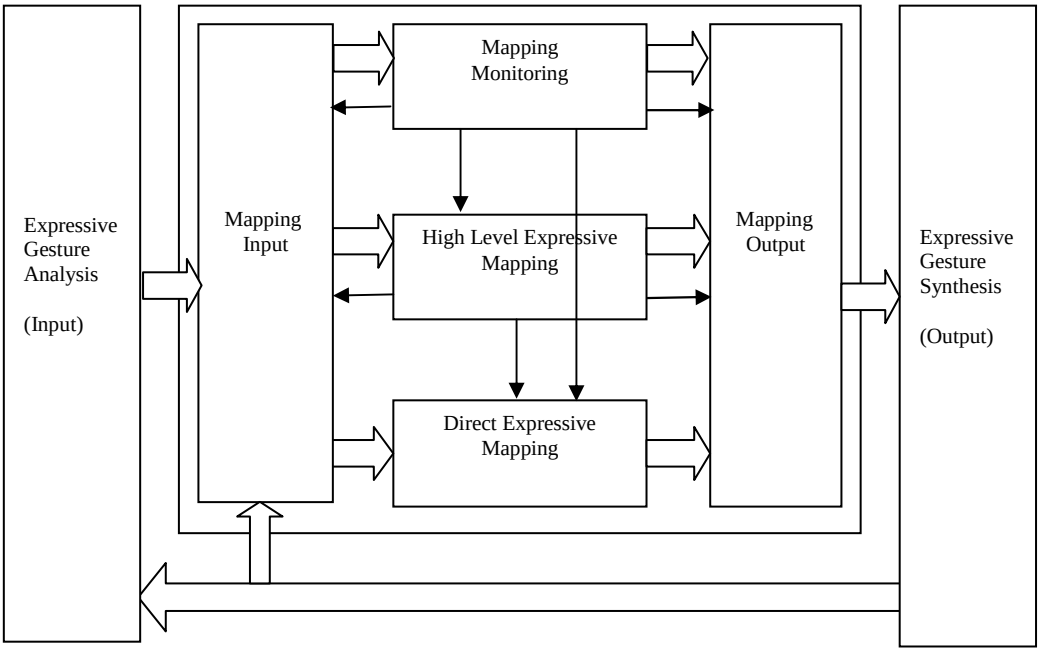


Figure 3.2: internal structure of a virtual or mixed subject

3.3. Analysis and synthesis of expressive gesture in virtual and mixed subjects

While going on with the task of describing in more details the information flow and processing of each component in a virtual or mixed subject, this Section considers the input and output components and illustrates a unified conceptual framework underlying both aspects. The next Chapter will focus instead on the mapping component and will describe a collection of possible mapping strategies a virtual or mixed subject could apply depending also on its expressive autonomy.
The unified conceptual framework here discussed has been developed in collaboration with partners in the EU-IST project MEGA. In particular, it has been conceived and discussed with Prof. Marc Leman at IPEM – Ghent University (Belgium) and with researchers at IPEM (Camurri, De Poli, Leman, 2001; Camurri, De Poli, Leman, Volpe, 2001).

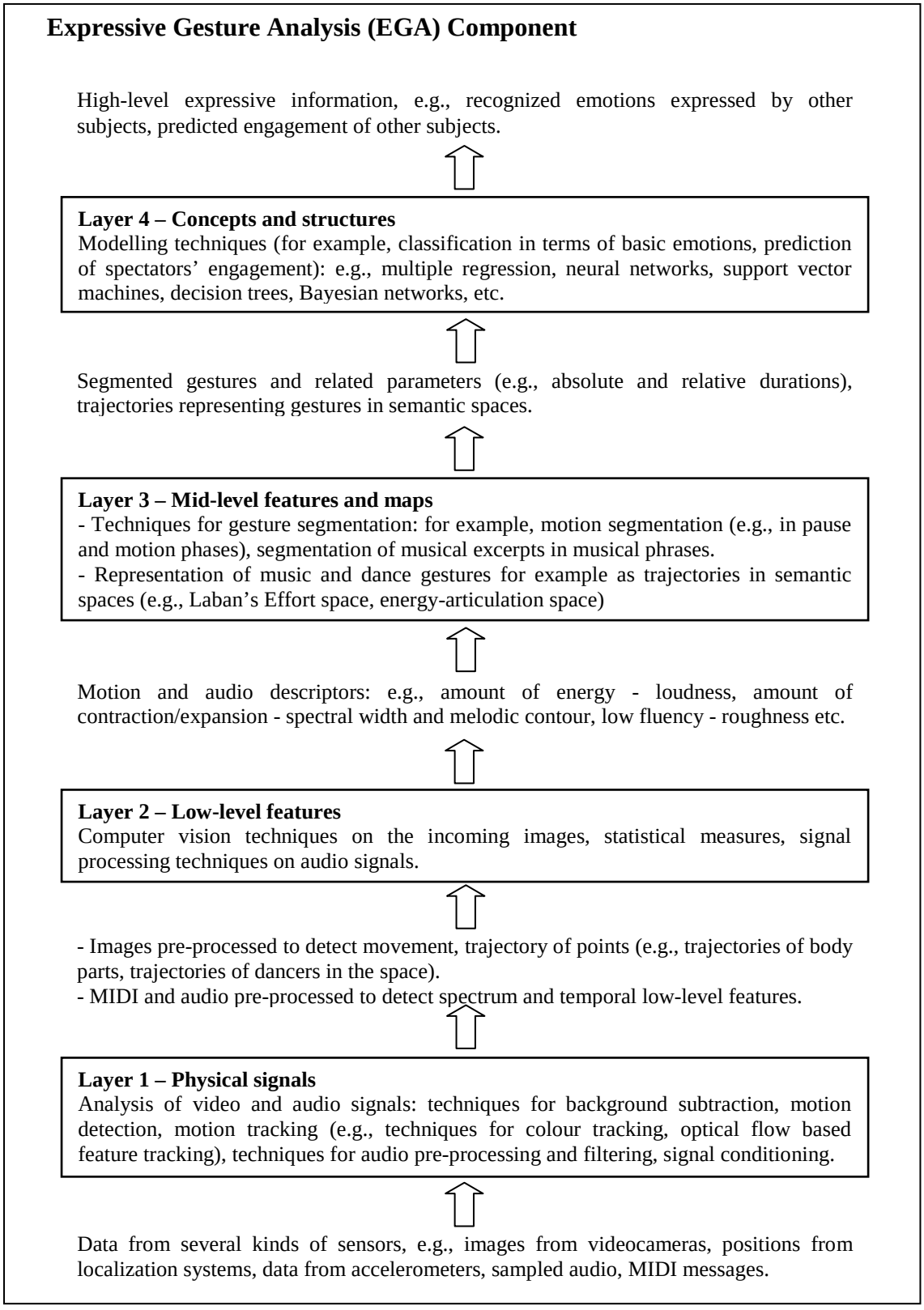


Figure 3.3: the multilayered framework for analysis of expressive gesture

Since, as discussed above, analysis and synthesis of expressive gesture is as a process involving several layers of abstraction, a multi-layered architecture is envisaged in which analysis is carried out by progressively extracting higher-level information from lower-level signals and vice versa for synthesis.
Such a multilayered approach enables to split up the problem of expressive gesture analysis, synthesis and mapping into different sub-problems. Given the nature of expressive gesture as a gestalt-like entity, a straightforward approach is to split up the gestalt-like entity in terms of features and parameters, events and gestures, gestural spaces, and concepts. This allows a bottom-up and a top-down definition of the notion of expressive gesture that can be employed the one for analysis and the other for synthesis.
Moreover, the multilayer approach also integrates, by means of "cross-modal mix" modules, features emerging from different physical input channels (e.g., audio, visual, sensors such as accelerometers or haptic devices). These integrations are conceived to be possible at different levels too.
Let's describe the multilayered architecture by considering the analysis side. According to this framework, analysis is accomplished by four subsequent layers of processing dealing with different kinds of inputs/outputs, ranging from low-level physical signals to high-level expressive information.
The Expressive Gesture Analysis (Input) component of a virtual or mixed subject is thus composed by four sub-components hierarchically ordered on the basis of the level of abstraction of the kind of information they process (see Figure 3.3 in the previous page).
In the figure, the multilayered structure is represented in its four layers, each one with its inputs and its outputs. Inside the boxes representing each layer a short list is included of possible techniques that may be applied at that level.
As stated above, such a multilayered structure should be considered under a multimodal perspective, i.e., it aims at integrating analysis of audio, video, sensor signals.
Layer 1 (Physical Signals) receives as input information captured by the sensors of a computer system. Physical signals may have different formats strongly dependent on the kind of sensors that are used. For example, they may consist of sampled signals from tactile, infrared sensors, signals from haptic devices, frames in video, sampled audio signals, MIDI messages. In this context the word "sensors" is often related to both the physical sensors employed and to the algorithm used to extract a given set of low-level data. It is therefore possible to speak of "virtual sensors" or "emulated sensors". For example, in the case of analysis of movement through videocameras, a CCD camera can be an example of physical sensor, while the optical flow, the motion templates, or the positions of certain points in the frame sequence are examples of data extracted from "virtual sensors" implemented by the cited algorithms. Layer 1 applies pre-processing, filtering, signal conditioning, and audio and video analysis techniques to the incoming rough data to obtain cleaner data and further signals derived from the rough input. For example, in the case of video analysis of human movement two types of output are generated: pre-processed images and trajectories of body parts.
Layer 2 (Low-level features) gets as input the pre-processed signals coming from Layer 1 and applies algorithms to extract a collection of low-level descriptors. The employed techniques range from computer vision algorithms, to signal processing, to statistical techniques (that can be applied on the extracted data). The extracted low-level descriptors are features that psychologists, musicologists, researchers on music perception,

researchers on human movement, and artists deemed important for conveying expressive content. In the case of analysis of expressive gesture in human movement, examples are the amount of contraction/expansion, of stability, of rotational movements. Important cues are those related to the Effort dimensions described in Rudolf Laban's Theory of Effort (Laban, 1947, 1963): these features will be extensively described in the second part of this dissertation. In the case of music such feature are related to tempo, loudness, pitch, articulation, spectral shape, periodicity, dynamics, roughness, tonal tension and so on: a similar conceptual framework and a taxonomy of audio features worked out in the context of audio mining can be found in (Leman et al., 2003, and Leman et al., 2001). Notice that analogies can be found among features in movement and in music, e.g., amount of motion – loudness, contraction/expansion – melodic contour or spectral width, bounded, hesitant movement – roughness.

Layer 3 (Mid-level features and maps): in (Camurri, De Poli, Leman, Volpe, 2001) this layer is described in these terms: "In this layer, the purpose is to represent expression in gestures by modelling the low-level features in such a way that they give an account of expressiveness in terms of events, shapes, patterns or as trajectories in spaces or maps." The layer receives data from Layers 1 and 2 and has two main tasks: segmenting expressive gestures and representing them in a suitable way. Such a representation would be the same (or at least similar) for gestures in different channels, e.g., for expressive gestures in music and dance. Data from several different physical and virtual sensors are therefore likely to be integrated in order to perform such a step. Each gesture is characterized by the measures of the different cues extracted in the previous step (e.g., speed, impulsiveness, directness, etc. for movement, loudness, roughness, tempo, etc. for music). Segmentation is a relevant problem at this level: the definition of expressive gesture does not help in finding precise boundaries. For example, in the second part of this thesis, a motion phase in dance will be considered as an expressive gesture (and segmentation will be done on the basis of the detected amount of motion). In facts, this is quite an arbitrary hypothesis: sub-phases of a motion phase (e.g., the phase of motion preparation) could also be considered as expressive gestures as well as sequences of motion and pause phases. Several possibilities are open for the common representation Layer 3 generates as its output. For example, an expressive gesture can be represented as a point or a trajectory in a semantic space[7]. Clustering algorithms could then be applied in order to group similar gestures and to distinguish different ones. Another possible output is a symbolic description of the observed gestures along with measurements of several quantities describing them.

Layer 4 (Concepts and structures) collects inputs mainly from Layers 2 and 3 and is responsible to extract high-level expressive content form expressive gestures. It can be organized as a conceptual network mapping the extracted features and gestures into (verbal) conceptual structures. For example, in the two experiments previously sketched in this Chapter the focus was on the four basic emotions (anger, fear, grief, and joy) in

[7] Whether the representation has to be a point or a trajectory depends on how the low-level features are processed in Layers 2 and 3. For example, if a vector containing the averages of the low-level features is calculated along the time duration of a gesture or a gesture is considered as a single event, it could be represented as a point in a multidimensional space. If instead more values for each feature are available (e.g., local values, or averages along gesture sub-phases) or if a gesture is considered as a sequence of events (as it is likely to be) a trajectory is a more appropriate representation.

the first one, and on spectator engagement (from a certain point of view something near to arousal in the valence-arousal model) in the second one. Other outputs are possible: for example, a structure could be envisaged describing the Laban's conceptual framework of gesture Effort, i.e., Laban's types of Effort such as "pushing", "gliding", etc. (see Laban, 1947, 1963, and later in Chapter 5). Several different machine learning techniques can be used for building such structure, ranging from statistical techniques like multiple regression and generalized linear techniques, to fuzzy logics or probabilistic reasoning systems such as Bayesian networks, to various kinds of neural networks (e.g., classical back-propagation networks, Kohonen networks), support vector machines, decision trees.

The conceptual architecture sketched above is conceived for analysis, i.e., the Expressive Gesture Analysis component of a virtual or mixed subject can be implemented following these guidelines. Anyway, a similar structure can be employed also for synthesis. Let's consider for example Layer 4: it may consist of a network in which expressive content is classified in term of the four basic emotions anger, fear, grief, and joy depending on current measures of low and mid-level cues. If instead of considering the framework from a bottom-up perspective a top-down approach is taken, an emotion the virtual subject intends to convey can be translated by a similar network structure in values of low and mid level cues to be applied to generated audio and/or visual signals. In this way the Expressive Gesture Synthesis component of a virtual or mixed subject can be obtained by using a similar multilayered architecture.

3.3.1. Customising analysis and synthesis

A virtual or mixed subject can evolve over time. It can change the spatial and time perspective under which it observes its environment (e.g., it can observe the whole scene or only particular aspects of it). It often needs to be adapted to a real subject it is associated with. Consider for example a virtual subject in an EME playing the role of avatar of a real subject inhabiting another EME. The avatar could be customised in order to reproduce the attitudes of the real subject associated with it. If for example the real subject pays a particular attention to light changes, the avatar could be customised in order to have a similar attitude. It is therefore needed to provide mechanisms for adapting the behaviour of the analysis and synthesis Layers, while preserving at the same time the modularity of the conceptual framework. This can be obtained by including some intermediate modules in between the Layers of the analysis and synthesis framework (see Figure 3.4 in the following page).

Suppose for example to have a module at Layer 3 able to extract the "scenic presence" of a dancer. The "scenic presence" would be a mid-level feature that could be used by Layer 4 for classifying the dancer's current expressive intention.

Modules in Layer 1 and 2 provide Layer 3 with the information needed to perform this task. On the basis of this information Layer 3 calculates an index of scenic presence. Suppose now that such an analysis is done by an observer associated to a human real subject (e.g., a spectator) paying particular attention to light. In this case, lighting on stage can strongly affect (in a non-linear way) the scenic presence index. For example, if the dancer were standing in a lighted area in front of the stage, his/her scenic presence

would sensibly grow. There is therefore the need to emphasize in a non-linear way this parameter. This can be done by means of an intermediate (between Layer 3 and Layer 4) mathematical module which takes as inputs the calculated index of scenic presence, the outputs from the physical layer (stage coordinates, lighting position and intensity), and the outputs of the low-level features layer (e.g., amount of detected motion to understand whether the dancer is or is not moving), and generates as output a modified (enhanced) index of scenic presence.

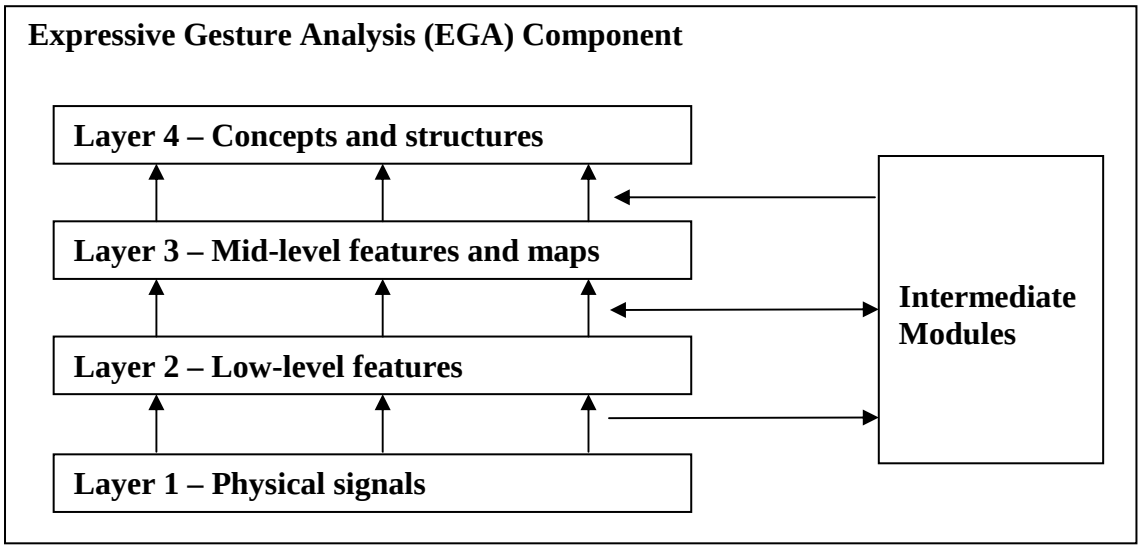


Figure 3.4: role of the intermediate modules in the expressive gesture analysis process

It should be noted that the mid/high-level feature "index of scenic presence" already implicitly depends on the actual values of the low-level features from Layers 1 and 2: the mechanism implemented by the intermediate module allows to adapt and tune the index according to the desired focus of attention and the attitudes of a modelled human subject.
Intermediate modules add flexibility to the analysis and synthesis architecture, since they allow the definition of "archetypical" models of features (e.g., algorithms to calculate them in "standard" conditions), while keeping separate (i) their evolution over time given specific contexts, and (ii) different biasing due to "personality", attitudes and focus of interest. Moreover, intermediate modules allow dynamic evolution of a virtual subject: while the layers extract features by using "archetypical" algorithms, the use of one (dynamic) or more intermediate modules can induce dynamic biases on the calculated features depending on the evolution of the performance.
Several possibilities are available for implementing intermediate modules, such as for example (i) non-linear algebraic functions, (ii) look-up tables, (iii) threshold switches, (iv) rule-based systems, (v) decision-making and data fusion modules.

4. Mapping of expressive gestures

As mentioned in Chapter 3, three main tasks can be individuated in expressive gesture processing: (i) analysis of expressive gesture, i.e., decoding the expressive content that expressive gestures contain and convey, (ii) mapping of expressive gestures, i.e., deciding which expressive content (if any) to convey given the incoming inputs and which expressive gesture to use to convey it, and (iii) synthesis of expressive gestures, i.e., generating suitable expressive gestures depending on the decisions made by the mapping function. A virtual or mixed subject in a non-verbal MIEE should be able to accomplish all these three tasks.
Chapter 3 presented a possible internal architecture for a virtual or mixed subject and discussed analysis and synthesis of expressive gesture by describing a conceptual framework that both (i) allows to consider and organize the analysis and synthesis processes under a multimodal perspective, and (ii) provides some guidelines for implementing the Expressive Gesture Analysis and Synthesis components of a virtual or mixed subject. Mechanisms for adapting virtual or mixed subjects with respect to the attitudes of a real subject were also described.
This Chapter concludes the conceptual discussion on expressive gesture by dealing with the mapping problem. The components related to mapping of the architecture for virtual or mixed subject presented in the previous Chapter will be analysed in details. Possible techniques for making decisions about which expressive content to convey and how to convey it will be shortly described. In particular, a software module developed for such a task will be presented. The Chapter will finish with a discussion on expressive autonomy: in fact, mapping is strictly dependent on expressive autonomy since the actual degree of expressive autonomy a virtual or mixed subject has strongly affects its capabilities to make decision about expressive content, i.e., mapping.

4.1. A multilayered model for mapping

Mapping of expressive gestures involves two main aspects:

(i) Making decisions about if, when, and how to answer to incoming inputs. This problem is strictly related with the paradigm of interaction that is employed. For example, in (Rowe, 2001, 1993) interactive systems for music are distinguished with respect to interaction paradigms in two subclasses: instrument paradigm systems “that treat the machine contribution as an extension or augmentation of the human performance”, and player paradigm systems considering the machine as an interlocutor. More in general, a distinction can be made between systems (and situations) in which a continuous mapping of inputs onto outputs is needed and systems (and situations) in which a sort of dialog takes place. These two conditions can be again considered as extreme boundaries of a continuum of possible intermediate situations and virtual and mixed subjects can be though to be continuously and dynamically evolving over time along this continuum.

(ii) Making decisions about which channel to use for the response. That is, once the virtual or mixed subject decided to try to convey a given expressive content, the most suitable ways (i.e., expressive gestures) have to be chosen in order to do it. This obviously depends on the subject's capabilities (of course, a subject endowed with only audio outputs can only generate sound and music outputs). If many possible output channels are available, the subject has to decide which one or which ones are most suitable given the actual context.

Mapping can take place on several layers: for example it is possible to associate a decoded emotional intention (e.g., one of the four basic emotions anger, fear, grief, and joy) to the generation of expressive gestures conveying the same or another emotional intention (e.g., it might be possible to create a kind of empathic agent showing its understanding of one's emotional state by displaying the same emotional state or, conversely, a subject answering to an emotional intention by displaying the opposite one[1]): in this case mapping would take place at Layer 4 in the conceptual framework depicted in Section 3.3. It is also possible to directly associate values of cues extracted by modules at Layer 2 with values of similar cues involved in synthesis: for example a movement performed with high energy can be associated to a musical excerpt played loudly. In this case mapping takes place at Layer 2. While moving bottom-up along the Layers, it is likely that interaction mechanisms move along the continuum from continuous mapping to dialogical mapping.
The general architecture for virtual and mixed subjects presented in Chapter 3 considers three main layers (components) for mapping. They are shown in Figure 4.1.

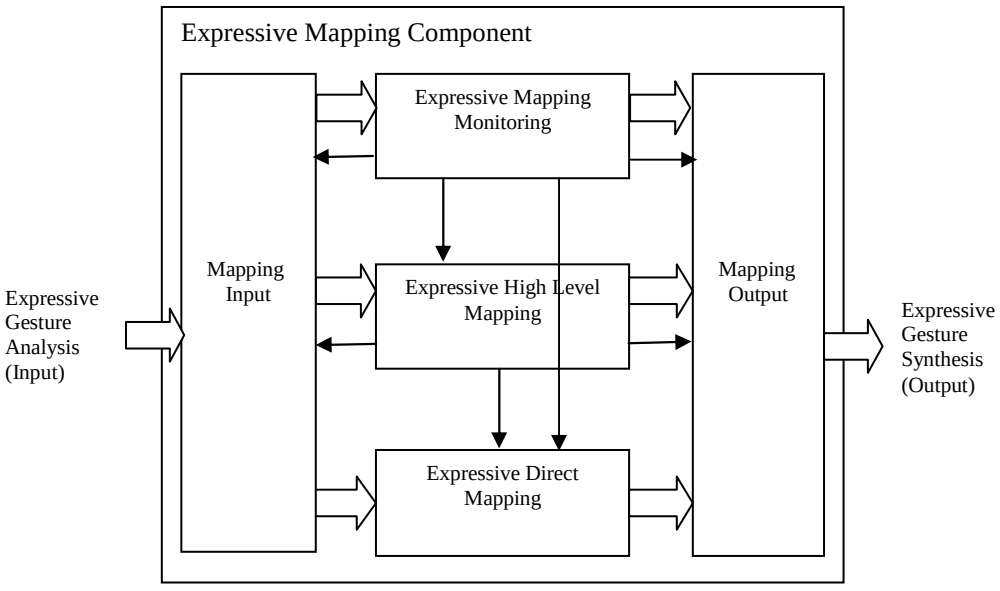


Figure 4.1: structure of the mapping component of a virtual or mixed subject

[1] Indeed, the last one would be quite a sadistic subject since it would be happy when you are sad!

All of them can receive inputs from the four layers of the Expressive Gesture Analysis component and can influence the four layers of the Expressive Gesture Synthesis component. The three layers correspond to three different kind of mapping strategies ordered with respect to increasing complexity. They include:

(i) *Expressive direct mapping*, such as for example in the case of algebraic functions associating, without any dynamics, features of the incoming expressive gestures to features of generated expressive gestures (e.g., mapping of movement cues onto parameters of algorithms for sound synthesis or post-processing and visual media).

(ii) *Expressive high-level indirect mapping*, including reasoning and decision-making processes. For example, consider a software module able to make decisions based on the incoming decoded expressive content: it could select an algebraic function as the ones in (i) within a collection of possible algebraic functions, thus allowing direct mapping to be adapted to the current context, i.e., implementing an adaptive and dynamic direct mapping.

(iii) *Expressive mapping monitoring*, i.e., algorithms trying to measure the effectiveness of the lower mapping layers with respect to the overall goals of the subject and of the performance. They can modify and adapt the processing of the lower mapping layers (e.g., by modifying decision parameters or changing possible collection of algebraic functions) as a result of their evaluations.

As usual, the white and thick arrows represent flows of data between the components, while the black and thin arrows represent influences that a component exerts on another one. Implementation guidelines as the ones discussed in Chapter 3 are still valid.
The three typologies of expressive mapping will now be discussed in more details. Notice that the term "expressive" is explicitly mentioned while speaking about mapping: this should help in avoiding misunderstandings with respect to other possible uses of the terms "mapping" and "mapping strategies" that can be found in the literature. In particular, with "expressive mapping" and "expressive mapping strategies" here I do not intend the association of features of analysed gestures to emotional categories[2], nor I mean the association of the physical movement of a performer with the generated sound of an (hyper) instrument[3].

4.1.1. Expressive direct mapping

With *expressive direct mapping* I intend an association without any dynamics of expressive features of analysed expressive gestures with parameters of synthesised expressive gestures. For example, the actual position of a dancer on the stage can be mapped onto the reproduction of a given sound. Expressive direct mapping is often associated with the lower levels of the conceptual framework for analysis and synthesis discussed in the previous Chapter: for example parameters calculated in Layer 2 (e.g.,

[2] This would be a "vertical" mapping between layers of the analysis framework. Here I am instead concerned with a "horizontal" mapping between layers in analysis and layers in synthesis.
[3] This kind of mapping could be indeed included in the concept of "horizontal" mapping I am dealing with: it is just a possible particular aspect in a broader scenario.

amount of motion – loudness) can be used to control particular features in the real-time generation of audio and visual content. Moreover, direct expressive mapping is also more likely to be employed in continuous mappings rather than in dialogic ones.
Expressive direct mapping allows obtaining simple reactive behaviours in virtual or mixed subjects, and therefore it can be associated with the reactive component of the Emotional Agent architecture (see Chapter 3 and Camurri and Coglio, 1998).
Several possible implementations are available for expressive direct mapping such as for example the following:

(i) Collections of pre-defined condition-action rules, i.e., set of rules associating given configurations of parameters coming from the analysis side with given configurations of synthesis parameters. For example, sounds or visual outputs can be associated with regions in the space, the use of a given post-processing with a given expressive cue, the automatic recognition of a given expressive gesture with the automatic generation of another expressive gesture.

(ii) Collections of algebraic functions, calculating values of synthesis parameters depending on values of analysed expressive cues. Let's call $\underline{u}$ a vector of expressive cues: $\underline{u}$ will contain numerical values calculated by the analysis component (e.g., the occupation rate of a region in the environment, the calculated fluency of a movement, the roughness of a musical excerpt). Let's also call $\underline{s}$ a vector of synthesis parameters (for example the parameters of a physical model for sound generation, the parameters controlling the movement of a computer generated character). If $\underline{s} \in S$ and $\underline{u} \in U$, an expressive direct mapping strategy can be though as a function $\underline{m}$: $S \rightarrow U$, $\underline{s} = \underline{m}\ (\underline{u})$ algebraically connecting analysis parameters with synthesis parameters. It should be noticed that while the complexity of the algebraic function can be freely increased according to any possible need, it anyway remains a static function, i.e., the mapping it induces does not change anymore once the function is defined and put at work.

4.1.2. Expressive high-level indirect mapping

Expressive high-level indirect mapping strategies can be associated with explicit use of reasoning techniques, and can therefore be related to the rational component of the Emotional Agent architecture. They are characterized by:

(i) A state evolving over time (that is, they are dynamic processes): such a state can be updated for example by applying some kind of reasoning technique to the available information.

(ii) Decisional processes, i.e., the system could make decisions based on the incoming information from analysis and the acquired knowledge. Such decisions can concern the kind of expressive content to produce and how to convey it, and can be related for example to the narrative structure of a performance.

Production systems and decision-making algorithms can be employed to implement this kind of mapping strategies. Let's consider again the vector $\underline{u}$ of expressive parameters returned by the analysis algorithms and the vector $\underline{s}$ of synthesis parameters, $\underline{s} \in S$ e $\underline{u} \in$

U. Let's consider also a collection of K possible expressive direct mapping algebraic functions: $\underline{m}_k$: S → U, $\underline{s} = \underline{m}_k(\underline{u})$ with $k = 1...K$.
Each function $\underline{m}_k$ directly maps a given configuration of analysis parameters onto a given configuration of synthesis parameters, as previously described while talking about expressive direct mapping.
The K direct mapping functions can be considered as possible alternatives among which a higher level module can choose depending on the available and incoming information. A decision-making algorithm can thus be employed to select the direct mapping function $\underline{m}_{k^*}$ that results to be the most suitable in the given situation (for example, in a given moment of a performance): it is therefore possible to have a collection of expressive direct mapping strategies among which a choice is made by a higher level mapping strategy (the decision-making algorithm).
A particular but relevant case is represented by linear direct mapping functions. These function can be written as $\underline{s} = \underline{m}_k(\underline{u}) = M_k \underline{u}$, where M_k is an $m \times n$ matrix (being $\underline{s}$: $m \times 1$ and $\underline{u}$: $n \times 1$). In this case the decision-making algorithm has to choose among K matrices $M_1 ... M_K$, representing the K linear direct mappings.
Mechanisms can be included providing smooth transitions between direct mappings, i.e., when the decision-making algorithm decides to change the underlying direct mapping, the smoothness and the time duration along which the change has to take place can be decided as well (in a sense, smoothness and time duration could be mapping parameters as well).
This paradigm can be further iterated, leading to hierarchies of mapping functions: suppose for example that H sets of direct mapping functions are available. Each set contains respectively $K_1 ... K_H$ functions. A first decision can then be made about which of the H sets should be considered. A second decision will concern which of the K_{h^*} functions in the selected set has to be employed with the incoming analysis data.
Notice that a similar approach can be applied even if the direct mapping is implemented through condition-action rules and sets of condition-action rules. A decision-making algorithm can be employed to decide which of K rules that can be applied in a given situation (i.e., whose conditions are matched) should be employed. In the literature of classical production systems this problem is usually addressed as the "conflict resolution" problem (see for example Russell and Norvig, 1995) and it is usually solved by employing simple algorithms (e.g., selection of the rule having the highest priority, selection of the most specific rule). If many sets of rules are available at the same time a two-step procedure as the one described for multiple set of algebraic functions can be considered.

4.1.3. Expressive mapping monitoring

A further layer of processing can be envisaged influencing both direct and indirect mapping. Such layer concerns the evaluation of the effectiveness of the currently employed mapping strategies whether they are direct or indirect. Effectiveness can be considered under several aspects: for artistic performances it can be related to the audience's engagement; in a museum scenario it could be associated to visitors' fruition

of the museum exhibit. Such a measure could be the result of a direct evaluation by spectators, in case it is not possible to calculate it automatically[4].
Once a measure of effectiveness is available, it can be used to make decisions aiming at improving the overall performances of the virtual or mixed subject by modifying and adapting its behaviour in order to maximize effectiveness.
Decisions made at the expressive mapping monitoring layer can influence both expressive direct and indirect mappings. Expressive direct mapping is affected through mechanisms similar to the ones described above, i.e., collections of functions or rules can be replaced by more suitable sets. But expressive mapping monitoring can operate also on indirect mapping. Suppose for example that many decision-making algorithms are available in the indirect mapping component. Many possibilities for interpreting and decoding information coming from analysis could also be available. Measures of effectiveness could thus be used to select among the available decision-making algorithms which one is most suitable given the measured effectiveness and the current situation (e.g., the part of the performance which is actually taking place). It could also be possible to dynamically adapt the way in which mapping processes the incoming inputs or sends information to the synthesis component.

4.1.4. The expressive mapping input and output subcomponents

Before talking about a possible implementation of indirect mapping, let's conclude the description of the structure of the expressive mapping component by shortly describing the mapping input and output subcomponents.
The main role of the input component is to encode the information coming from analysis in a way that can be processed by the mapping components. For example, if direct mapping is implemented as a set of condition-action rules, the input component has to encode the information coming for analysis according to the syntax of the condition part of the rules. Another task the mapping input component is responsible for is dispatching information coming from the four layers of the analysis framework to the appropriate subcomponents of the mapping component. For example, it is more likely that the output of an emotion classifier at Layer 4 will be sent to the indirect mapping sub-component, rather than to the direct mapping one.
Conversely, the main task of the output sub-component is to translate the output of the mapping components as required by the algorithms implemented in the synthesis Layers. For example the action part of a condition-action rule could needs to be translated in a vector of control parameters for the synthesis algorithms. The output subcomponent also has to dispatch information to the four layers of the synthesis framework. For example, it can send to the physical layer the name of a MIDI score that has to be played, and to the conceptual layer the emotional intention according to which expressive deviations on performance parameters have to be calculated.

[4] In fact the expressive mapping monitoring layer should be though as a conceptual layer. At the moment no commitment is done about the possibility to partially or fully implement it. Anyway, it should be noticed that experiments like the ones sketched in Chapter 3 and further discussed in the following are investigating the possibility to measure spectator's engagement.

4.2. The Affective Decision Maker (ADM)

Direct mapping can be implemented by means of standard mathematical functions or classical production systems. Here attention focuses on a pilot implementation of an expressive indirect mapping component employing decision-making algorithms to make decisions among a collection of possible expressive gestures to be generated as output by a virtual or mixed subject.
The Affective Decision Maker (ADM) is a deterministic multiattribute decision maker that can be influenced in its processing by the expressive information the analysis layers extract from expressive gestures[5]. The ADM operates on a set of alternatives among which a choice has to be done. Each alternative is characterized by the values of a set of attributes. Each attribute represents a criterion with respect to which the decision is made. The table made by the values of the attributes for each alternative is called decision table and is the main internal data structure of the ADM.
Consider, for example, that you have to buy a new PC: the overall cost, the CPU clock frequency, the amount of RAM etc. are relevant aspects (attributes) that you have to consider in order to make your choice. The final choice will be made among several PCs (alternatives), each one characterized by a particular given value for each attribute.
In the case of the ADM both the attributes and the decision-making mechanisms are (or can be) related to expressive content.
The ADM selects among alternatives taking into account two kinds of information: (i) information about the environment in which it operates (Environmental Information) and (ii) information about the expressive content coming from the analysis layers (Expressive Information).
Consider, for example, a museum application. A virtual subject observes visitors' behaviour and tries to catch their attention for improving fruition and possibly making more interesting their visit. In this context, the positions of visitors inside a room or the number of visitors can be considered as Environmental Information; the detected posture of a visitor, some properties of his/her movement communicating an expressive content (e.g., interest) can be considered as Expressive Information. In some other contexts the distinction is not so clear. In the case of a dance performance in a theatre, for example, the position of a dancer on the stage could be considered Environmental Information, but the same position carries also an expressive content (e.g., in relationship with the scenery) so that it could be considered Expressive Information as well. In fact, what should be considered Expressive and what should be considered Environmental Information strongly depends on the specific application: the decision is usually up to the designer of the application.
The ADM returns as output the alternative it selected on the basis of the incoming Expressive and Environmental Information, of the current decision making algorithm and of the algorithms that have been used to update its internal data structures.
Figure 4.2 shows the internal structure of the ADM. It is divided up into two main components: an information-processing module and a decision-making module. The

[5] The Affective Decision Maker described in this section is inspired to previous studies investigating the rational component of the Emotional Agent architecture. In particular, a component similar to the ADM has been there used for selecting among several possible goals of the agent (Camurri and Volpe, 1999).

information-processing module manages the internal data structures: a decision table and an array of decision parameters. The decision maker contains the decision-making algorithms and is responsible to make the decision.

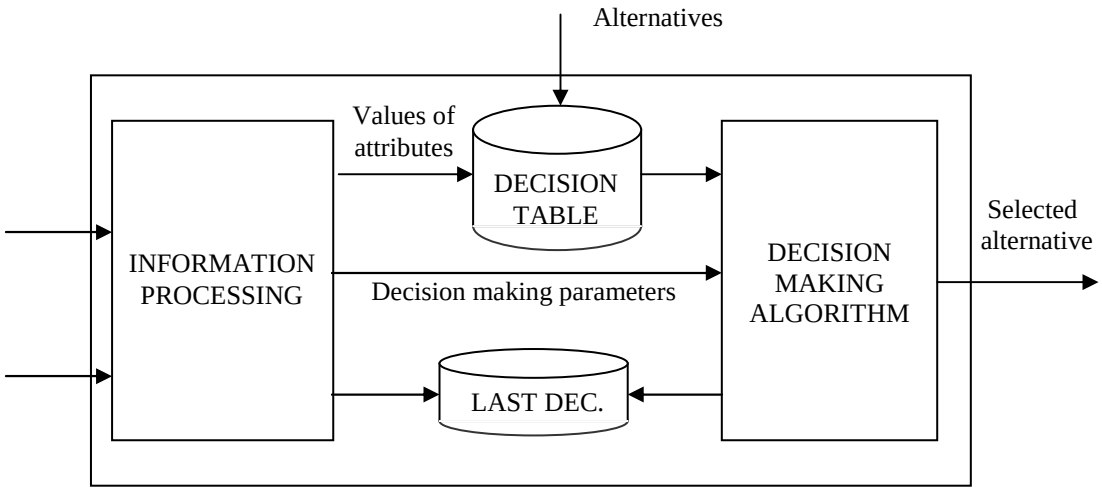


Figure 4.2: the internal structure of the ADM

During the initialisation phase, an initial decision table is loaded in the internal decision table data structure. Such data structure is a matrix containing a row for each alternative and a column for each attribute. The values of the attributes in the decision table are normalized to 1. An array of decision parameters is also maintained: it contains a weight for each attribute and a parameter α used when the Hurwicz and Hurwicz Modified decision-making algorithms are selected (see below the description of the five decision-making algorithms the ADM implements at the moment). The weights are in the range [0, 1] and their sum must be 1. The parameter α is also in the range [0, 1]. In the initialisation phase α is set to 0.5 and the weights are set to 1/(Number of Attributes).
At run-time, the decision table and the decision parameters are updated according to the input information. Then, (if needed) the currently selected decision-making algorithm chooses an alternative according to the actual decision table and decision parameters. The last decision can be taken into account to avoid choosing it again.
Both the update and decision-making algorithms can be dynamically changed at run-time, for example depending on the expressive mapping monitoring component.

4.2.1. ADM's decision-making algorithm

Before looking in more details at the mechanisms the ADM uses to deal with expressive content, the five decision-making algorithms the ADM currently implements are here shortly reviewed. Four of them are classical multiattribute decision-making algorithms; the fifth is a modification of the classical Hurwicz algorithm[6].

[6] See (Camurri and Volpe, 1999) and (Volpe, 1999) for a more detailed description.

Let consider x_{ij} to be the value of the j[th] attribute for the i[th] alternative. Let be *m* the number of alternatives and *n* the number of attributes. The decision table will be a matrix having the following structure:

x_{11}	x_{12}	…	x_{1j}	…	x_{1n}
x_{21}	x_{22}	…	x_{2j}	…	x_{2n}
…	…	…	…	…	…
x_{i1}	x_{i2}	…	x_{ij}	…	x_{in}
…	…	…	…	…	…
x_{m1}	x_{m2}	…	x_{mj}	…	x_{mn}

Let be $\underline{w}$ a vector of weights. $\underline{w}$ contains a weight for each attribute, therefore its dimension is *n*. The weights must sum to 1 (that is, $\sum_j w_j = 1$)

Note that the algorithms can select not just one alternative: it can happen that some alternatives result equivalent for a given algorithm. The algorithms, therefore, return a set of alternatives. A choice has to be made among equivalent alternatives in this set. In the current implementation, just the first alternative in the set (according to the order in which the alternatives are stored in the decision table) is returned as output by the ADM. The five decision-making algorithms are the following:

(i) *MAXIMAX*. The set A^* of the indexes of the selected equivalent alternatives is:
$A^* = \{ i : i = \arg\max_i (\max_j x_{ij}) \}$.

The algorithm calculates the maximum value for each row (that is, each alternative has a "score" equal to the value of the attribute having the maximum value with respect to the other attributes). Then the alternative with the maximum score is selected. If two or more alternatives have the same maximum score they are all part of the set of selected equivalent alternatives.

(ii) *MAXIMIN*. The set A^* of the indexes of the selected equivalent alternatives is:
$A^* = \{ i : i = \arg\max_i (\min_j x_{ij}) \}$

The algorithm calculates the minimum value for each row (that is, each alternative has a "score" equal to the value of the attribute having the minimum value with respect to the other attributes). Then the alternative with the maximum score is selected. If two or more alternatives have the same maximum score they are all part of the set of selected equivalent alternatives.

(iii) *HURWICZ*. The set A^* of the indexes of the selected equivalent alternatives is:
$A^* = \{ i : i = \arg\max_i [\alpha \min_j x_{ij} + (1-\alpha) \max_j x_{ij}] \}$

The algorithm calculates the maximum and minimum values for each row. A "score" consisting in a trade-off between the value of the attribute having the maximum value with respect to the other attributes and the value of the attribute having the minimum value represents each alternative. The trade-off is obtained through the parameter α. Then the alternative with the maximum score is selected.

If two or more alternatives have the same maximum score they are all part of the set of selected equivalent alternatives.

(iv) *HURWICZ MODIFIED*. The set A^* of the indexes of the selected equivalent alternatives is:

$A^* = \{ i : i = \arg\max_i [\alpha \max_j z_{ij} + (1-\alpha) \min_j z_{ij}] \}$, where $z_{ij} = w_j x_{ij}$.

The algorithm is very similar to the previous classical Hurwicz algorithm. The only difference is that the weights of the attributes are also taken into account. So, in this case the trade-off is made not just on the values of the attributes, but also on the weighted values of the attributes (and the weights can depend on the expressive content).

(v) *SAW (Simple Additive Weighting).* The set A^* of the indexes of the selected equivalent alternatives is:

$$A^* = \left\{ i : i = \arg\max_i \sum_j w_j x_{ij} \right\}$$

In this case, the "score" given to each alternative is a weighted sum of the values of the attributes, where the weights are the weights associated to each attribute. Then the alternative with the maximum score is selected. If two or more alternatives have the same maximum score they are all part of the set of selected equivalent alternatives.

It should be noticed that depending on the values contained inside the decision table and the weights assigned to each attribute, different decision-making algorithms can obtain different choices. Notice also that:

- Peaks in the values of an attribute can lead the MAXIMAX algorithm to choose the alternative with a high value for an attribute, even if the values for the other attributes are quite low.
- MAXIMIN often selects an alternative having values around the mean values.
- MAXIMAX, MAXIMIN and classical Hurwicz algorithms do not use any information about the weight assigned to the attributes. When one of these algorithms is selected the weights and the way in which the weights are updated do not influence the decision-making process.
- The MAXIMAX strategy could be seen as an "optimistic" strategy: it takes the best among the best values, while the MAXIMIN strategy could be considered as a "pessimistic" or "prudent" strategy since it takes the best among the worst values. In this perspective, if the Hurwicz or Modified Hurwicz algorithms are used, the parameter α can be seen as a measure of the trade-off between the optimistic and pessimistic point of view.

4.2.2. Affective Decision Maker: some issues

After describing the structure of the ADM and the algorithms it employs, some further relevant issues need to be discussed and clarified: which are the mechanisms through which expressive content influence decision making? How can the ADM be adapted to

dynamically changing environments (this is an important aspect for a virtual or mixed subject that should be able to adapt its behaviour)? How should the ADM decide that it is time to make another decision since the previous one is not suitable anymore? How can a suitable set of attributes be selected? This section tries to answer to some of these questions.

1. Why affective? As it was described up to this point, the ADM could seem to be a common deterministic, multiattribute decision maker: in fact, it can be used in such a way, if needed. Anyway, the ADM was designed principally to support decision making in contexts where expressive/emotional information has a particular importance and directly affects the decision making process. In the ADM architecture, expressive content can influence the decision making process in three different stages:

(i) When the decision table is updated: the algorithm updating the values in the decision table takes into account the Expressive Information input. In this way, the values of the attributes can directly reflect the expressive/emotional content. Note that the attributes themselves can be defined as related to expressiveness: e.g., the expressive importance of an action can be an attribute in the decision process leading to the selection of an action (while other attributes can be related to the actual utility of an action).

(ii) When the decision parameters are updated: the algorithm updating the decision parameters also takes into account the Expressive Information input. In this way, attributes can be differently weighted according to expressive content. For example, if an attribute represents a measure of how much a certain action is supposed to catch the attention of a user, the weight of this attribute should be increased when the information in input shows a low degree of attention from the user.

(iii) Directly in the decision-making phase: for example, the Hurwicz and Hurwicz Modified decision-making algorithms contain the parameter α that can be updated accordingly to the Expressive Information input. In particular, as already discussed, the parameter α can be seen as a measure of the trade-off between “optimistic” and “pessimistic” decision-making strategies.

2. Dynamic update of algorithms and decisional contexts. The update and decision-making algorithms can be dynamically changed at run-time. This feature provides a further mechanism for expressive content to influence the decision making process and, at the same time, allows the ADM to adapt its behaviour to changing decision making contexts. In fact, changing the update algorithm means changing the way in which the expressive (and environmental) information influences the updating of the decision table and decision parameters (according to the first mechanism described above). Changing the current decision making algorithm means (i) changing the way in which the internal data structure are considered in order to make a decision (ii) changing the way in which expressive information influences the decisional process (e.g., if the current Hurwicz Modified Algorithm is replaced by the MAXIMAX Algorithm two of the three “affective” mechanisms described above, attribute weighting and α, are not working anymore). Basically, the need of changing the update and decision-making algorithms emerges from the need of adapting the behaviour of the ADM to dynamically changing

decisional contexts. For example, in a (quite long) dance performance several update algorithms could be used in the different phases of the performance in order to modify and adapt the mapping between Expressive Information and Decision Table/Decision Parameters so that the mapping can be better suited to the content (and in particular expressive content) of each phase of the performance. The decision about when to change the current algorithms and which ones should be used is up to other mapping modules (e.g., a mapping monitoring component) or to a human supervisor. The information used by the mapping monitoring component or by the human supervisor in order to make such a decision also has an important role: for example, if a measure of effectiveness could be defined measuring how much the previous decisions were good, the algorithms could be replaced when such a measure goes under a given threshold.
The same expressive information could be used in more subtle ways too: changing the update algorithms on the basis of some expressive information means that expressive information is used to state how expressiveness itself should influence the decision making process. For example, suppose that an analysis module were able to detect happiness: depending on this information the current algorithms could be replaced by some new more "optimistic" ones, that is, not only it is possible to change some values in the tables, or some weights, or the α parameter depending on the recognized degree of happiness, as described in the previous section, but it is also possible to decide to completely change even the algorithms that are employed to calculate such values.

3. Temporal scope of decisions. In the previous sections, when describing the behaviour of the ADM at run-time, it was said that the ADM makes a new decision "if needed". How is it decided if a new decision is needed? In the current implementation the ADM just exports a command and makes a new decision each time such a command is given. Sending this command is up to another component, e.g., a mapping monitoring component (that could be implemented by another ADM) or a human supervisor.

4. Selecting a suitable set of attributes. Multiattribute decision makers are commonly used in order to make financial decisions (e.g., in economy). In such contexts, it is usually well known how to characterize an alternative, i.e., the sets of attributes are quite well defined for any given problem and methods are available to obtain the values for each attribute. For example, consider you have to buy a new car and you can select among several models. Each model will represent an alternative and each model (alternative) will be characterized by the values of a set of features (attributes). It's quite well known what are the relevant features for a car (e.g., average fuel consumption, maximum speed, price etc.) and it's quite easy to obtain such data. Thus, once filled the decision table, the decision maker is able to make its choice. But, when we move on contexts where expressiveness has an important role, it could be difficult to find a set of attributes characterizing an alternative and then measure their values. Suppose, for example, you have to select an audio or video fragment among a collection of available fragments. You should be able (i) to characterize each fragment with some numerical values related to expressiveness (ii) to assign such values in order to fill the decision table. In practice, you have to answer a question similar to this one: "Given the inputs I have (for example, measured features of the movement of a dancer, a recognized basic emotion in a music performance...), why should I select this video fragment instead of

another one?" Often this is not an easy task. Some good attributes could measure how much an alternative is suitable with respect to some measured features (for example, how much playing a given audio or video fragment is suitable with respect to the rate of occupation by a dancer of a certain region on the stage): in this case, the (affective) decision maker would select the alternative (e.g., the audio or video fragment) that seems to be the most suitable given the inputs.

4.3. Expressive autonomy[7]

As already sketched in Chapter 2, expressive autonomy plays a role of paramount importance in designing interactive systems for artistic applications. It is related to the role of technology in the performance and to the relationship between automatic decisions and decisions made by the director/creator of the artistic performance.
In the framework of MIEEs in which virtual and mixed subjects interact with real subjects the problem of the expressive autonomy can be introduced by proposing a question: to which extent can a virtual or mixed subject make autonomous decisions? That is, does the virtual or mixed subject have to follow the instructions given by the director, the choreographer, the composer, (in general the creator of a performance or of an installation) or is it allowed some degree of freedom in its behaviour?
This question was firstly asked in (Camurri, Coletta, Ricchetti, Volpe, 2000) where expressive autonomy was defined by taking as example a dialog between a dancer and a robot. The issue was raised by the design and implementation of a robot-dancer interaction (in the context of the performance "L'Ala dei Sensi", held in Fearrara, Italy, in November 1999) in which the robot was an interpreter (that is, conveying some expressive content) of a predefined "score of movements".
In fact, many hours in rehearsals were spent to obtain the desired behaviour and once obtained, the robot was not allowed anymore to deviate too much from the expected behaviour. A similar situation can be found in most music and theatre performances. The performer is often asked to convey the expressive content that the director, the composer, the choreographer intends (or intended) to communicate.
In general, a virtual or mixed subject in a MIEE can have different degrees of expressive autonomy. According to the definition given in the cited paper, the expressive autonomy of a virtual or mixed subject in a MIIEE is defined as the amount of degrees of freedom that a director, a choreographer, a composer (or in general the designer of a MIEE or the author of an application involving communication of expressive content) leaves to the subject in order to make decisions about the most suitable expressive content to convey in a given moment and about the way to convey it (i.e., which expressive gestures have to be generated to convey it).
It should be noticed that expressive autonomy is therefore somewhat different with respect to autonomy as intended in Artificial Intelligence and Robotics (see for example Russell and Norvig, 1995): in fact, expressive autonomy does not concern the amount of

[7] The concept of expressive autonomy has been introduced and discussed in (Camurri, Coletta, Ricchetti, and Volpe, 2000) from which this Section partially derives.

built-in knowledge the subject (agent) contains nor its capabilities to make decisions on its own on the basis of the feedback coming from its physical sensors.
It is immediately evident from the definition that expressive autonomy is strictly related to mapping, being mapping the process devoted to choosing the expressive content that has to be conveyed and the expressive gestures suited to convey it. In fact, a virtual or mixed subject having no expressive autonomy would not need any mapping component: mapping strategies would be selected in advance by the author of the performance and the only task of the subject would be to execute what the author already decided.
Let's consider again the example of the robot interacting with a dancer: if during a performance the robotic subject is asked to perform in an expressive way a sequence of movements that the director predisposed and that has been repeatedly tuned during a number of rehearsals, then the robot is only minimally expressively autonomous or it is not expressively autonomous at all. In this case, it is just needed to pre-program the robot with the sequence of movements that have to be performed and with the way in which they have to be performed (and that was decided and tested by the director): no components for expressive direct or indirect mapping are needed, except for possible recovering from unexpected situations.
Of course, this is not always the case: if, as it happens for performers, some degrees of freedom are still present and the subject (e.g., the robot) is flexible, versatile and rational enough to intervene when necessary to add nuances to its behaviour coherently with the performance, then it could be said that the subject is expressively semiautonomous. That is, it plays the role the author or the director assigned to it, but it can still make decisions for example about the way of conveying expressive content. For example, an expressively semiautonomous robotic subject could choose which expressive gestures (e.g., which style of movement) are most suitable to generate in order to appear happy, in a part of a performance during which the director wants the robot to appear happy. A semiautonomous subject has therefore to be provided with mapping components, although they would be allowed to directly control only certain aspects of the subject's behaviour, while other aspects would be pre-programmed as in the previous case.
It is also possible (and this is the most interesting situation) to design MIEEs in which automatic subjects have a high degree of expressive autonomy: it is for example the case of the installation at the permanent science exhibit for children "La Città dei Bambini", developed in Genova, Italy, in 1997 (see Camurri and Coglio 1998; Camurri and Ferrentino, 1999) where a robot played the role of guide for visitors or was a visitor itself. More recently (opening in November 2001), at the permanent science exhibit "La Città della Scienza" in Napoli, Italy, a virtual character was developed inhabiting five different computer stations. A human supervisor (e.g., a mime) controls one of the stations and directs movements and expressions of the virtual character by means of sensor systems (e.g., a data glove). The supervisor also gives his/her voice to the virtual character. The remaining stations (four or all the five stations if the supervisor is not present) are automatically managed, that is the virtual character is endowed with suitable mapping strategies enabling him to autonomously interact with visitors. For making decisions about its behaviour, the virtual character uses both inputs coming from the museum environment and captured by microphones and videocameras and information about possible narrative structures in its dialog with visitors.

In such a case, although the subject could have to follow a narrative thread, however it can choose what expressive content to convey in order to increase the interest of its audience: the author of the application builds a narrative structure, and the subject is assigned with the task to instantiate/interpret it in a suitable way given its current audience and context. The actual degree of expressive autonomy, however, can depend on the structure and dynamics of the narration and can vary over the time during the visit. For example, some schema could be provided (e.g., derived from sociological studies) within which and on the basis of which narration takes place. The subject can therefore be allowed of a high degree of expressive autonomy within such schemas[8].
Complete expressive autonomy implies that at a given moment the subject is completely free to choose the expressive content it wants to convey as well as the way to convey it. Complete expressive autonomy therefore implies the existence of a full mapping component implementing all the mechanisms previously described.
With respect to expressive autonomy subjects can be placed along a continuum, having at one of its extreme points the completely controlled subject, on the other the completely expressive autonomous subject, and in between all the degrees of semi-autonomy. The continuum is represented in Figure 4.3.

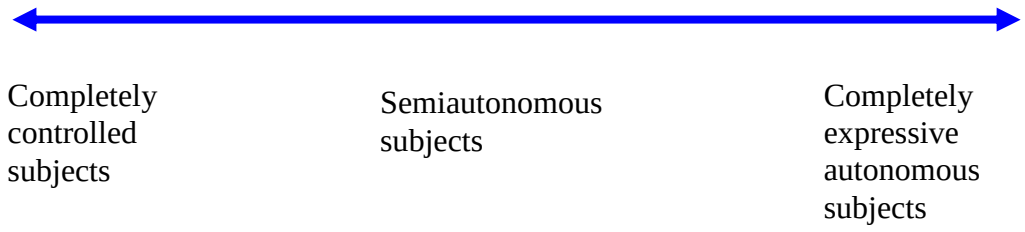


Figure 4.3: the expressive autonomy continuum

Notice that the expressive autonomy continuum is very similar to the continuum sketched for Active EMEs in Chapter 2. In fact, in the framework of MIEEs, EMEs are regarded as subjects at a higher-level metaspace: therefore they have a given degree of expressive autonomy. Thus a subject representing a completely passive EME in the higher-level metaspace will not have any expressive autonomy, while a subject representing a highly dynamic active environment will have a high degree of expressive autonomy.
Notice also that the continuum should be intended in a dynamic way, i.e., a subject can dynamically change its current degree of expressive autonomy. For example a robot can be completely controlled in a certain part of a performance and then "come to life" in another part by acquiring a high degree of expressive autonomy. Of course, this implies that the robot should be endowed with the needed expressive mapping components and that it should be possible to dynamically enable and disable mapping components during the performance.

[8] Notice that schemas can help in matching the needs of the author with the needs and the capabilities of technology: in fact, if from the one hand they constrain the expressive autonomy of the created subjects, on the other hand they also constrain the expressive autonomy of the author who should organize the performance or the installation around them, once selected.

The required degree of expressive autonomy (i.e., the position of a subject in the expressive autonomy continuum) is crucial also from the point of view of implementation. In fact, a high degree of expressive autonomy requires the subject to have more sophisticated capabilities in order to make its expressive choices. Thus, while the design and the implementation of a subject with a limited degree of expressive autonomy may result quite simple (e.g., only expressive direct mapping might be needed), a high expressively autonomous subject might need to be equipped with different kinds of expressive mapping components (both direct and indirect mappings and mapping monitoring).

As a final remark, I have to notice that the framework depicted so far is mostly a conceptual framework. A big challenge is now to implement it fully or partially, i.e., to design and implement tools being high-level and flexible enough to allow authors to build artistic performances (or other applications, like museum applications) based on MIEEs inhabited by expressive subjects with different degrees of expressive autonomy, without they have to worry about the technological issues and the underlying complexity.

The EyesWeb open software platform developed over the years at the DIST – InfoMus Lab (see Appendix A for a more detailed description) is a first step in this direction, but more high-level capabilities (e.g., the definition of a high-level language describing MIEEs and their content) would be needed and should be developed.

PART 2

ANLYSIS OF EXPRESSIVE GESTURE IN HUMAN FULL-BODY MOVEMENT

5. Expressive gesture in human full-body movement

After introducing MIEEs, their components, and expressive gesture as a main non-verbal communication channel, the focus now moves on a specific aspect: analysis of expressive gesture in human full-body movement. That is, human full-body movement is investigated as an example of conveyor of expressive content in interaction within a MIEE. A collection of techniques for analysis of expressive gesture in human full-body movement at different layers is discussed in the framework of an experiment carried out at the DIST InfoMus Lab in collaboration with the Department of Psychology of the University of Uppsala (Sweden) in the context of the EU-IST MEGA project.
From a scientific point of view the experiment tries to give some first answers to questions like the following ones: which are the features in expressive gesture that are mainly responsible of conveying expressive content? How can they be measured? How is the temporal dynamics of such features related to the communication of different expressive contents? Is it possible to build a classifier able to automatically classify expressive gestures on the basis of the expressive content they convey? Are the outputs of the automatic classifier consistent with spectators' perception of expressive gestures?
From a technical point of view the experiment is intended to shed light on possible design and implementation of a virtual or mixed subject capable to observe expressive gestures in full-body movement of people interacting with it and to decode the conveyed expressive content associated with them. This objective implies the implementation of the conceptual framework illustrated in Chapter 3 (adapted to this specific application) by developing and applying the techniques operating at each level on data coming from sensors and further processed by the subject.
In this perspective this Chapter is devoted to present the theoretical framework with respect to which motion analysis is carried out, with particular reference to the sources driving the investigation: mainly, theories by psychologists (e.g., Wallbott, Argyle, Boone and Cunningham) and choreographers and researchers on human movement (e.g., Laban). Possible different perspectives and approaches to human full-body motion analysis are discussed. The experiment in its initial hypotheses and its methodology is also described.
The following two Chapters will go deeper in the discussion and will deal with two different topics: the employed techniques for extracting expressive cues from motion at different layers in Chapter 6, the classification problem and an attempt to deal with it in Chapter 7.

5.1. Background and sources

The analysis of expressive gesture in human-full body movement described in this dissertation is inspired to several sources ranging from approaches grown in the traditional fields of science (e.g., psychology) and engineering (e.g., biomechanics) to approaches derived from theories from art and humanities (e.g., choreography, music

composition). In a sense, this work can be considered as an attempt to bridge the gap between these two fields toward the common goal of understanding expressive gestures and exploiting their communicative power under a scientific perspective (i.e., a deeper understanding of non-verbal communication channels), an engineering perspective (i.e., building enhanced and effective interactive systems for several different application domains), and an artistic perspective (i.e., exploiting the means technology provides in order to enrich language and to pioneer novel art forms).
Main sources on which the approach here adopted finds its foundations come from:

(i) Research and theories on KANSEI and emotion arousal/appraisal (e.g., the Hashimoto's theory on KANSEI Information Processing sketched in Chapter 1);
(ii) Biomechanics, techniques for motion capture and computer vision;
(iii) Research and theories from art and humanities on communication of expressiveness in dance (e.g., Rudolf Laban's Theory of Effort, Laban 1947) and music (e.g., Pierre Schaeffer's Sound Morphology, Shaeffer 1977);
(iv) Research and theories from psychology on non-verbal communication of expressiveness (Wallbot 1980, Argyle 1980, Boone and Cunningham, 1998);

Since KANSEI Information Processing has already been described in Chapter 1 and biomechanics and techniques for motion capture and computer vision mainly deal with technical aspects of motion detection and processing described in Chapter 6, here the focus will be on the two last sources: art and humanities and psychology.

5.1.1. Theories from art and humanities

A classical approach frequently employed in machine learning consists in creating an explicit description of a studied phenomenon in term of a collection of parameters. The values of such parameters forming a vector of parameters for every available sample in the training set are then used to perform analysis (i.e., recognition, classification, regression). For example, starting from a human movement signal, this approach builds a description in terms of expressive cues (such as fluentness, directness, energy, etc.), shapes, and phrasing. While from the one hand this approach is commonly used in machine learning, on the other hand it finds some basis also in theories from art and humanities. As an analogy, in music this would be equivalent to recreate a "score" starting from a sound signal, or, better, to build a representation of the signal in terms of a vocabulary similar to Pierre Schaeffer's Morphology (Schaeffer, 1977).
As already sketched in Chapter 2, Schaeffer's Morphology is an attempt to describe and study "concrete music" where music objects extend the traditional musical instruments with sounds coming from the real life, produced by concrete objects. In this direction, Schaeffer's Morphology is an approach supporting musicological analysis of such music. Morphological qualities based on perceptual features enable segmentation of continuous streams of a (concrete) sound signal: segmentation and identification of music objects are based on perceptual cues such as "grain", "texture", "allure" etc. Analogies can be investigated with analysis in human movement where similar problems can be envisaged (e.g., segmentation of a continuous stream of movement data, identification of motion primitives, extraction of a collection of perceptual cues). From such a comparative

analysis it may be possible to individuate a collection of features having a similar role in both music and movement domains[1].
Some research works showing analogies between music (e.g., level envelopes of tones) and movement (e.g., force patterns in walking) can be found in the literature (Sundberg, Friberg, Frydén, 1994; Friberg and Sundberg, 1999; Friberg, Sundberg, Frydén, 2000).
Rudolf Laban's Theory of Effort (Laban and Lawrence 1947, Laban 1963) provides a similar "cues language" in the domain of human movement (e.g., in dance). The Theory of Effort is one of the main inputs for the analysis carried out in this work and it is therefore worth to be described in some more details.
In the Theory of Effort, Laban points out the dynamic nature of movement and the relationship among movement, space and time. Laban's approach is an attempt to describe, in a formalized way, the main features of human movement without focusing on a particular kind of movement or dance expression. In fact, it should be noticed that while being a choreographer, Laban did not focused only on dance, but rather he envisaged in his theory the whole complexity of human movement including dance expression but also extended to everyday movements like the ones performed by workers in their usual activities.
The basic concept of Laban's theory is *Effort* considered as a property of movement.
From an engineering point of view it can be considered as a vector of parameters identifying the qualities of a movement performance. It has to be noticed that Theory of Effort describes the *quality of movement*. That is, it is not concerned with, for example, degrees of rotation of a certain joint or the moment that has to be applied, rather it considers movement as a communication media and tries to extract parameters related to its expressive power.
The effort vector can be regarded as having four components generating a four-dimensional "effort space" whose axes are Space, Time, Weight, and Flow[2]. During a movement performance such effort vector describing the motion qualities moves in the effort space. Laban investigates the possible paths followed by the vector and the expressive intentions that may be associated with them[3].
Each effort component is measured on a bipolar scale, the extreme values of which represent opposite qualities along each axis.
Space refers to the actual direction of a motion stroke and to the path followed by subsequent strokes (a sequence of directions). If the movement follows these directions smoothly the space component in the effort space is considered to be "flexible", while if it follows them along a straight trajectory it will be marked as "direct".
Time is also considered with respect to two different aspects: an action can be "sudden" or "sustained", which allows the binary description of the time component of the effort

[1] Of course, it is often not straightforward to find a direct connection between features in music and features in movement. Anyway, concepts that are similar in the two fields can be worked out in order to individuate a collection of features having similar roles in the two domains.
[2] In his original theory Laban considered a three-dimensional space defined by the Space, Time, and Weight components. The fourth component, Flow, was intended as a kind of modifier with respect to the three basic components.
[3] Notice that expressive content is more likely to be encoded in the trajectory of the effort vector in the effort space, rather than in its absolute position. That is, expressive gestures are characterized by the effort dynamics (i.e., variations) along time, rather than by the values of the effort components at a given time.

space. Moreover, in a sequence of movements, each of them has a given duration in time: the ratio of the durations of subsequent movements gives the time-rhythm, as in a music score and performance.
Weight is a measure of how much strength and weight is exerted in a movement. For example, in pushing away a heavy object it is necessary to use a strong weight, while in handling a delicate and light object, the weight component has to be light.
Flow is a measure of how bound or free a movement, or a sequence of movements, appears. Laban describes it in these terms: "In an action capable of being stopped and held without difficulty at any moment during the movement, the flow is *bound*. In an action in which it is difficult to stop the movement suddenly, the flow is *free* or *fluent*" (Laban, 1963, p. 56).
The two extremes of each bipolar scale along each axis can be interpreted as "indulging" with respect to a given dimension (e.g., light and flexible movements indulge in weight and space) or "fighting" against it (e.g., quick and bound movements fight against time and flow). A graphical notation is also provided for describing movements with different effort qualities. Figure 5.1 summarizes in a table the eight extreme qualities along each effort axis and shows the graphical notation Laban developed to describe effort.

Axes	Indulging Effort	Fighting Effort
Space	*Flexible*	*Direct*
Time	*Sustained*	*Quick*
Weight	*Light*	*Strong*
Flow	*Free*	*Bound*

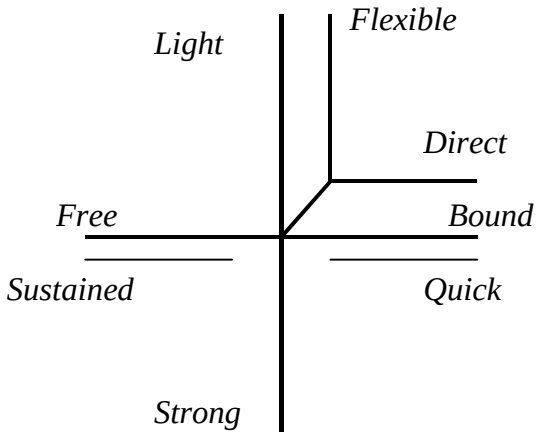


Figure 5.1: efforts table and graphical notation

Laban's basic theory considers mainly the first three components of effort (Space, Time, and Weight) to develop a description of human movement. By considering the three-dimensional space built on these three axes and the opposite qualities for each effort component, it is possible to identify eight combinations of the Space, Time, and Weight components (addressed as basic efforts), corresponding to states that the movement can assume in its development. These eight combinations can be considered as the vertexes of a cube in the effort space whose axes are Space, Time, and Weight. Such a cube is represented in Figure 5.2. The eight basic efforts and their qualities are summarized in Table 5.1.

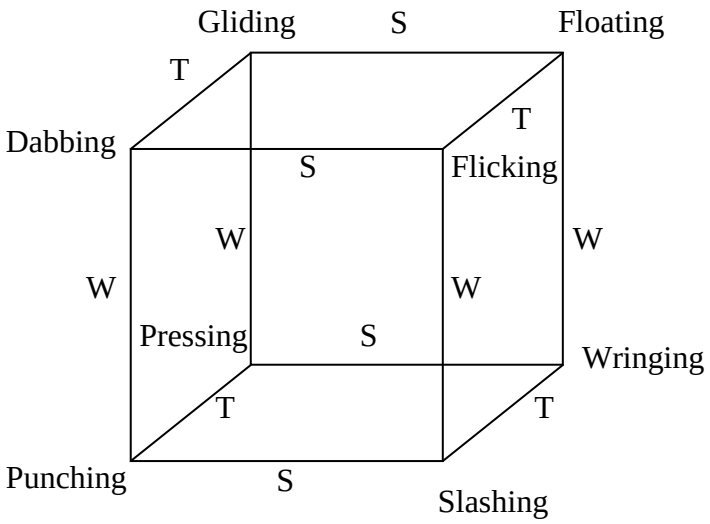


Figure 5.2: the effort cube (Laban, 1947). A basic effort is associated to each vertex. The letter on each edge indicates the effort component that changes when moving from one vertex to an adjacent one (S = Space, T = Time, W = Weight).

Basic Effort	Space	Time	Weight
Pressing	*Direct*	*Sustained*	*Strong*
Flicking	*Flexible*	*Sudden*	*Light*
Punching	*Direct*	*Sudden*	*Strong*
Floating	*Flexible*	*Sustained*	*Light*
Wringing	*Flexible*	*Sustained*	*Strong*
Dabbing	*Direct*	*Sudden*	*Light*
Slashing	*Flexible*	*Sudden*	*Strong*
Gliding	*Direct*	*Sustained*	*Light*

Table 5.1: the eight basic effort and their qualities as described in (Laban, 1963)

In a recent study (Zhao, 2001), four neural networks have been trained to recognize the two extreme qualities for each effort component. The training set consisted of a collection of arm movements whose features have been extracted by using both motion capture systems and videocamera based systems.

5.1.2. Research in psychology

Research on human movement as a mean for human-human communication has been widely developed in psychology. Several researchers (see for example Argyle, 1980) stressed the importance of full-body gestures (body postures and movements) in non-verbal human-human communication.
A main research field concerns visual perception of human movement and its qualities. For example, in his well-known investigation on point light displays Johansson (1973) showed that human observers are able to still get a vivid impression of human movement even if only points of light placed on body joints are displayed. Beside the importance of the result (i.e., lot of information about the moving person can be subtracted away yet maintaining a vivid impression of the performed movement), this also allows to produce and manipulate motion stimuli (as trajectories of body joints) in a relatively easy way.
Following the path opened by Johansson, other researchers devoted their work to visual perception of movement focusing on specific qualities (e.g., genre and identity of the moving person) or on specific actions (e.g., walking). Developmental studies have also been carried out. A short review of the research work on visual perception of human movement can be found for example in (Pollick, 2003).
A field of particular interest for the work described in this dissertation concerns the relationships between movement and expressiveness with respect to two main aspects: (i) which are the movement features that are mainly responsible to convey expressive content (what I call "expressive cues") and (ii) how these expressive cues relate to a particular expressive content, or, in other words, how an expressive message is encoded in the dynamics of these cues. In the tradition of the work by Johansson, it has been shown that it is possible for human observers to perceive emotions in dance from point light displays (Walk and Homan, 1984; Dittrich et al., 1996). Pollick (2001) analysed recognition of emotion in everyday movements (e.g., drinking, knocking) and found significant correlations between motion kinematics (in particular speed) and the activation axis in the two-dimensional space having as axes activation and valence as described by Russell (1980) with respect to his circumplex structure of affect. Wallbott (1980) in his paper dealing with measurement of human expression after reviewing a collection of works concerning movement features related with expressiveness and techniques to extract them (either manually or automatically), classified these features by considering six different aspects: spatial aspects, temporal aspects, spatio-temporal aspects, aspects related to "force" of a movement, "gestalt" aspects, categorical approaches. Boone and Cunningham (1998) starting from previous studies by De Meijer (1989, 1991) identified six expressive cues involved in the recognition of the four basic emotions anger, fear, grief, and happiness, and further tested the ability of children in recognizing emotions in expressive body movement through these cues. Such six cues are "frequency of upward arm movement, the duration of time arms were kept close to the body, the amount of muscle tension, the duration of time an individual leaned forward, the number of directional changes in face and torso, and the number of tempo changes an individual made in a given action sequence" (Boone and Cunningham, 1998). It has to be noticed that in their paper Boone and Cunningham distinguish between propositional and nonpropositional aspects of movement (Buck, 1984, cited in Boone

and Cunningham, 1998). This distinction can be related to what already said about gestures that can be expressive both if they have a specific denotative meaning and if they do not have it. In fact, propositional movements are intended as established signs transmitting a given meaning (e.g., a raised hand to indicate stop). Specific movements corresponding to emotion stereotypes can also be considered as propositional (e.g., a clenched fist to show anger or raised arms to demonstrate joy). Non-propositional movements are instead embodied in the direct and natural emotional expression of body movement based on fundamental elements such as tempo and force that can be combined in a wide range of movement possibilities. Therefore, non-propositional movements do not rely on specific movements, but build on the quality of movements i.e., how movements are carried through, for instance whether it is with lightness or heaviness (Camurri, Lagerlöf, Volpe, 2003). Non-propositional movements are thus expressive gestures that do not have a denotative meaning, i.e., that kind of expressive gesture that more often is encountered in performing arts.

5.2. Perspectives of analysis

Human full-body movement can be analysed under different views and perspectives. Several aspects contributing in encoding expressive content in expressive gestures have to be taken into account by a virtual or mixed subject observing the motion of a user/participant interacting with it. Moreover, the way in which each of these aspects is dealt with strongly influences the design and implementation of virtual and mixed subjects. Some of these aspects (partially following the classification of movement cues described in Wallbott, 1980) are shortly discussed in this Section.

5.2.1. Space views

A first aspect concerns the space under analysis, i.e., which extent is considered and which level of detail is assumed in analysis with respect to the spatial component. In his book "Modern Educational Dance" Laban (1963) introduces two relevant concepts: the *Kinesphere*, referred also as *Personal Space*, and the *General Space*, the whole space surrounding the Kinesphere In particular Laban says: "Whenever the body moves or stands, it is surrounded by space. Around the body is the sphere of movement, or Kinesphere, the circumference of which can be reached by normally extended limbs without changing one's stance, that is, the place of support. The imaginary inner wall of this sphere can be touched by hands and feet, and all points of it can be reached. Outside this immediate sphere lies the wider or "general" space which man can enter only by moving away from their original stance. He has to step outside the borders of his immediate sphere and create a new one from the new stance, or, in other words, he transfers what might be called his "personal" sphere to another place in the general space. Thus, in actual fact, he never goes outside his personal sphere of movement, but carries it around with him like a shell" (Laban, 1963, p. 85).

A first distinction can thus be done between analysis in the Personal Space and analysis in the General Space. This distinction does not only determine the spatial extent on which analysis has to be carried out (e.g., in case of dance performances, the space occupied by the body of a dancer for the Personal Space, and the whole stage for the General Space), but it also affects the kind of techniques employed for analysis. In fact, even if analogies can be found among features in the Personal Space and in the General Space, different techniques can be needed to extract them. The next Chapter will illustrate algorithms for analysis in the Personal Space, while Chapter 8 will discuss a reference model for analysis in the General Space.
Further subdivisions can be done depending on the envisaged level of detail in both the Personal and the General Spaces. For example, it is possible to consider the motion of only one person (e.g., a dancer) within the General Space or the motion of a group of persons in order to analyse the behaviour of the group as a whole. In the Personal Space it is possible to consider global features, such as for example the global amount of detected motion or the contraction/expansion of the whole body (examples will be discussed in Chapter 6) or local features like those describing the motion of a given joint or of a given part of the body (e.g., head, hands, feet).
In a perspective from wide to narrow these different spatial points of view can be summarized as follows:

(i) Global properties in the General Space, i.e., behaviour of a group considered as a whole in the General Space;
(ii) Local properties in the General Space, i.e., behaviour of single individuals, separately analysed, in the General Space;
(iii) Global properties in the Personal Space, i.e., behaviour of the body considered as a whole in the Personal Space;
(iv) Local properties in the Personal Space, i.e., behaviour of given parts of the body, separately analysed, in the Personal Space

It should be noticed that this subdivision should not be considered as a rigid and static one, but rather as a continuum of possibilities through which the focus of attention of a virtual or mixed observer moves, depending on the current needs. Many analyses at each of the four levels of detail can be carried out in parallel and their results integrated toward a global interpretation of the detected movement.

5.2.2. Time views

Time also plays a very important role in analysis, mainly with respect to the time interval on which analyses are carried out. The time interval can vary from a few milliseconds (e.g., one frame from a videocamera) to several minutes (a whole performance) and it depends on the evolution of the performance and its narrative structure (e.g., in a dance performance) as well as on considerations about how movement is perceived by humans with respect to time. The adopted mapping strategies (whether continuous or dialogical, see Chapter 4) can also strongly influence the duration of time that is considered for analysis, continuous mapping often requiring quite short time windows on which fast

computations have to be performed, while dialogical mapping needs to segment expressive gestures and analyse them in their whole duration.
The problem has been dealt with also (and maybe mainly) in the framework of analysis of music. For example, in a taxonomy of descriptors of musical audio worked out by Marc Leman and colleagues in the context of audio mining (Leman et al., 2001, 2003) a distinction is made among non-contextual "low level descriptors obtained from a frame-based analysis of the acoustical wave" (e.g., onsets, offsets, roughness), mid-level descriptors "derived from musical context dependencies within time-scales of about 3 seconds" (e.g., beat, short rhythmic patterns, short interval sequences, tonal tension) and allowing through segmentation an event-based representation of musical objects, and high-level descriptors that "typically involve learning and categorization beyond the representation of the *now*", referring to time intervals longer than 3 seconds, and related to the cognitive and emotional/affective domains. High-level features are related to long-term memory processes, while low and mid-level features are mainly dealt with by the short-term memory.
A similar approach can be envisaged also for the time aspect of motion descriptors. That is, it is possible to distinguish between descriptors calculated on different time scales:

(i) Low-level descriptors, calculated on a time interval of a few milliseconds (e.g., one or a few frames coming from a videocamera). For example the current amount of contraction/expansion can be calculated on just one frame (see the description of the Contraction Index in Chapter 6), i.e., on 40 ms with the common sample rate of 25 fps.

(ii) Mid-level descriptors, calculated on a movement stroke (in the following also referred as "motion phase"), on time durations of a few seconds. Examples of such descriptors are the overall direction of the movement in the stroke (e.g., upward or downward) or its directness (i.e., how much the movement followed direct paths). At this level it is possible to obtain a first segmentation of movement in strokes that can be employed for developing an event-based representation of movement. In fact, strokes or motion phases can be characterized by a beginning, an end, and a collection of features including both mid-level features calculated on the stroke and statistical summaries (e.g., average, standard deviation), performed on the stroke, of low-level features (e.g., average body contraction/expansion during the stroke).

(iii) High-level descriptors related to the conveyed expressive content (but also to cognitive aspects) and referring to sequences of movement strokes or motion (and pause) phases. Time intervals in the case of dance performances range from a motion phrase (some seconds), to a microdance (a short dance fragment with a duration up to a few minutes, see in the following), to a whole dance performance (several minutes)[4].

With respect to the general framework illustrated in Chapter 3, low-level descriptors can be collocated at layers 1 and 2, mid-level descriptor at layer 3, and high-level descriptors at layer 4.

[4] It should be noted that it seems that even one or few motion strokes can already convey expressive content. See for example the work by Pollick (2001) in which human observers were able to distinguish emotions in very short everyday movements (e.g., drinking or knocking).

It should be noted that the time aspect is on a large extent complementary with respect to the space aspect described above, i.e., it is possible to have low-level, mid-level, and high-level descriptors for the movement of a limb in the Personal Space (consider for example the amount of research in analysis of arm movements), for the movement of the full-body in the Personal Space (the main subject of the experiment discussed in the following), and for the movement of individuals and groups in the General Space as for example in studies about the global behaviour of visitors of a museum exhibit.

5.2.3. Spatio-temporal views: "polyphony"

Another important aspect concerns how movements are interleaved and orchestrated in space and time. Parallelism and orchestration can be viewed both with respect to space, i.e., orchestration of gestures of different body parts of the same or different dancers (e.g., a coordination of different features and gestures in arms and legs such as elasticity and high-energy in legs and low-energy and inertial direct movement in arms can reflect happiness), and with respect to time, i.e., subsequent movement strokes connected with fluency versus "broken" and hesitating motion, aspects similar to articulation in music.
"Polyphony" is also related to what Laban (1963) calls "effort rhythms", i.e., sequences of basic efforts preceded by a preparation phase and followed by a termination phase. Such sequences stress in different ways the four effort components of Space, Time, Weight, and Flow depending on which components change and which remain constant during the sequence. By stressing one component (e.g., directness in space) with respect to another different expressive contents can be conveyed.

5.2.4. Motion with respect to its target

In everyday life motion is often a goal-directed action, i.e., movement is intended to reach a given target in the space. Something similar can happen also in dance, the artistic expression of movement: one or more dancers can tend to reach a target on the stage. As a consequence, spectators perceive a sort of arousal, "the need for" a target rising from the observed movements. A target can be made explicit through the design of the scenery and through specific dialogue mechanisms between humans on stage: for example, a dancer following another one or escaping from another one.
The reach of a target and the physical effort (fatigue) it costs is an important element in analysis, even if often difficult to measure.
The way in which a target position is approached can also be relevant for analysis. A collection of features can be extracted describing how the target is approached in space and time (e.g., in a direct and sudden way or with flexible and sustained movements).
Beside the importance of this kind of analysis in non-verbal communication through expressive gestures, the way in which a target is reached plays also a relevant role in other application domains, such as for example in therapy and rehabilitation where measures like the directness of the trajectory followed to reach the target can be used as evidence for diagnosis of particular pathologies and for therapy monitoring.

5.2.5. Postures

Not only motion strokes are important for decoding the expressive content conveyed through movement, but also pauses (or pause phases) in between strokes can have a relevant role. During a pause it is likely that a particular posture is assumed. Argyle (1980) discusses the importance of postural attitudes in non-verbal communication: postures are used to express interpersonal attitudes, emotions, and personality traits. Moreover, a given gesture assumes a different expressive "strength" according to the postural conditions in which it is made (e.g., the posture assumed before and after the gesture or the posture of the whole body when only a part of it, for example a limb, is performing the gesture). In interpersonal communication, postural attitudes define the basic style of communication: for example relaxing, indifferent, curious.
Analysis of postures during pauses in the movement can therefore be needed to fully understand the expressive content associated with expressive gestures in human full-body movement.

5.3. Approaches to analysis of expressive gesture in movement

After introducing the main aspects that have to be taken into account when analysing expressive gesture in human movement, here two approaches are presented that can be followed in performing the analysis. The first one is a bottom-up approach that in the framework of the conceptual architecture described in Chapter 3 starts from processing of physical signals for extracting movement features and tries to decode the expressive content by using the motion descriptors that are obtained at the subsequent layers. The second one, following the tradition starting from research on point light displays by Johansson (1973), proceeds by progressively subtracting information from a rich stimulus in order to find the cues that are mainly involved in expressive content communication.
Until now the discussion concerned expressive gesture in human movement from a general point of view even if lot of times movement in dance has been considered as a useful example. Even the work by Laban addressed not only dance, but also more general aspects like movements of workers in their everyday activities and, in fact, the research on expressive gesture described in this dissertation mainly concerns the development of interactive multimedia systems enabling novel interaction paradigms and allowing a deeper engagement of the user in a number of different application scenarios.
However, a particular focus is put on performing arts and on artistic performances because of the strongly use in their languages of non-verbal communication mechanisms to convey expressive content. Dance and music performances therefore constitute an ideal test-bed where computational models of expressive gesture and algorithms for expressive gesture processing can be developed, studied, and tested. The two approaches will be therefore described with reference to dance performance and they will be applied to a collection of dance performances.

5.3.1. Bottom-up approach: microdances

The bottom-up approach tries to individuate and model expressive cues by studying a reference archive of microdances recorded for this purpose. With "microdance" it is meant a short video fragment containing enough information to be able to decode and classify expressive content.
Microdances can be useful to individuate that features that are mainly responsible of conveying expressive content. In particular, analysis of microdances can provide experimental evidence with respect to the cues that choreographs and psychologists already identified: this is mainly obtained by an analysis of differences and invariants in the same choreography performed with different expressive intentions. For example, a comparison can be done between a choreography performed in a "neutral" way, i.e., didactically and without any expressive intention, and the same choreography performed with expressive intentions corresponding to the basic emotions fear, grief, anger, and joy. At a first stage, such a comparison can be done by hand or through annotations asked to choreographers and dance experts.
Once a set of possible expressive cues is individuated algorithms can be developed for automatically extract them from the available microdances. Techniques can then be applied for expressive content classification.
Microdances can also be used for testing the developed models and algorithms. Human observers evaluate each microdance. The outputs obtained by the developed algorithms and models are then compared with spectators' rating of the same dance fragment, thus allowing evaluation of the performances of the algorithms.
Notice that the same approach can be applied at different layer with respect to the conceptual framework described in Chapter 3: at layer 2 for a perceptual validation of the extracted low/mid level cues (e.g., spectators can be asked to evaluate the amount of perceived motion and the results compared with the outputs of the algorithm computing the amount of detected motion), at layer 3 for a perceptual validation of gesture segmentation (segmentation of dance gestures by spectators and by a segmentation algorithm can be compared), at layer 4 for expressive content classification (for example, classification of dance fragments with respect to the four basic emotions anger, fear, grief, and joy, performed by spectators and by classification techniques can be compared).

5.3.2. The subtractive analysis approach

The subtractive approach starts from the work by Johansson on point light displays. It aims at identifying those features that are mainly responsible of expressive content communication by progressively reduce information from the initial stimulus. With respect to the works on point light displays, this approach does not only intend to show that recognition is possible with reduced information but it also attempts to evaluate the contribute and the weight of the contribute that different kinds of information bring to expressive content decoding and understanding.

An aspect on which the subtractive approach can be employed is related to emotional arousal, i.e., the effective engagement of spectators exposed to artistic stimuli.
In this case, for example, the inputs to the subtractive analysis are genuinely artistic live performances and their corresponding video recordings. A reference archive of artistic performances, chosen after a strict interaction with artists and performers, has to be built. Image processing techniques can then utilized to gradually subtract information from the video recordings. For example, parts of the dancer's body could be progressively hidden, deforming filters could be applied (e.g., blur), the frame rate could be slowed down, etc.
Each time information is reduced spectators are asked to rate the intensity of their engagement in a scale ranging from negative to positive values (a negative value meaning that the video fragment rises some feeling in the spectator but such a feeling is a negative one). The transitions between positive and negative rates and a rate of zero (i.e., no expressiveness has been found by the spectator in the analysed video sequence) would help to identify what are the movement features carrying expressive information.
Of course, a deep interaction is needed between the image processing phase (i.e., the decisions on what information has to be subtracted) and the rating phase.
In the following of this dissertation and in particular in the experiment that will be now described and that will provide the framework in which the developed algorithms will be discussed, the bottom-up approach will be mainly followed. Nevertheless, experiments using the subtractive approach are also currently carried out at the DIST - InfoMus Lab, even if they are not subject of this discussion.

5.4. An experiment on analysis of expressive gesture in dance performance[5]

As an example of analysis of expressive gesture in dance performance, an experiment is now discussed, carried out in collaboration with the Department of Psychology of the University of Uppsala (Sweden) in the framework of the EU-IST MEGA project.
The aim of the experiment is twofold: (i) individuating which motion cues are mostly involved in conveying the dancer's expressive intentions to the audience during a dance performance and (ii) testing the developed models and algorithms by comparing their performances with spectators' ratings of the same dance fragments.
In particular, in the case of this experiment expressive gesture has been analysed with respect to its ability to convey emotions to the audience. The study is in fact focused on the communication through dance gesture and recognition by spectators of the four basic emotions: anger, fear, grief, and joy.
After outlining some research hypotheses a collection of motion descriptors (expressive cues) has been identified and algorithms developed to extract them. The algorithms have been applied on twenty microdances constituting the reference archive for the experiment. At the same time, spectators have been asked to indicate the expressive

[5] The description of the experiment is partially taken from the following papers:
Camurri A., Lagerlöf I., Volpe G., "Emotions and cue extraction from dance movements", International Journal of Human Computer Studies, in press, 2003.
Camurri A., Mazzarino B., Timmers R., Volpe G., "Multimodal analysis of expressive gesture in music and dance performances", V International Gesture Workshop, Genova, 2003.

content (i.e., the basic emotions) they were able to identify in the dances. Finally, ratings from spectators have been compared with the results of the employed classification techniques. This Section describes the experiment, its hypotheses and methodology. The following two Chapters will deal respectively with algorithms for extraction of expressive cues and with the obtained results.

5.4.1. Research hypotheses

The research hypotheses are grounded on the role in dance expressive gesture of the Laban's dimensions of Space, Time, Weight, and Flow. In particular, as a result of a joint discussion with the psychologists in Uppsala the following aspects emerged.

- The Space dimension is considered in its aspects related to Laban's Personal Space by measuring to what extent limbs are contracted or expanded in relation to the body centre, how much movements are direct or flexible (i.e., tend to follow straight trajectories or smooth ones), which direction results to be prevalent in motion (for example, Boone and Cunningham, 1998, showed that joy is characterized by a higher amount of upward movements);
- The Time dimension is considered in terms of overall duration of the whole performance, of duration of motion strokes (i.e., pause and motion phases), and of tempo changes (that also contribute to the underlying structure of rhythm or flow in the movement).
- The Weight dimension is considered with respect to the amount of tension and dynamics in movement: since the technical difficulties arising when measuring aspects like movement tension, weight has been mainly associated with the vertical component of acceleration.
- The Flow dimension is considered in terms of analysis of shapes of speed and energy curves, frequency/rhythm of motion and pause phases, amount of acceleration and deceleration during motion phases.

In the hypotheses discussed with the psychologists these expressive cues are associated in different combinations to each emotion category. For example, in (Lagerlöf and Djerf, 2002), also reported in (Camurri, Lagerlöf, and Volpe, 2003), the table in the following page can be found.

5.4.2. Description of the experiment

An experienced choreographer was asked to design a choreography such that it excluded any propositional gesture or posture and it avoided stereotyped emotions.
In Uppsala, five dancers performed this same dance with the four different emotional expressions: anger, fear, grief and joy. Each dancer performed all the four emotions. The dance performances were video-recorded by two digital videocameras (DV recording format) standing fixed in the same frontal view of the dance (a spectator view). One camera obtained recordings to be used as stimuli for spectators' ratings. The second video camera was placed in the same position but with specific recording conditions and

hardware settings to simplify and optimise automated recognition of movement cues (e.g., manual shutter). Dancers' clothes were similar (dark), contrasting with the white background, in an empty performance space without any scenery. Digitised fading eliminated facial information and the dancers appeared as dark and distant figures against a white background.
The psychologists in Uppsala then proceeded in collecting spectators' ratings: the dances were judged with respect to the perceived emotion by 32 observers, divided in two groups. In one group ratings were collected by "forced choice" (choose one emotion category and rate its intensity) for each performance; the other group was instead instructed to use a multiple choice schemata, i.e., to rate the intensity of each emotion for all the four emotions for each performance.
At the same time, at the DIST - InfoMus Lab motion cues have been extracted from the video recordings and models for automatic classification of dance gestures in term of the conveyed basic emotion have been developed.
In the next Chapter, the algorithms used for extracting motion cues will be presented. An extended discussion of the output of the computational and statistical models and a comparison with spectators' ratings will be included in Chapter 7.

Basic Emotion	Expressive Cues
Anger	Short duration of time Frequent tempo changes, short stops between change Movements reaching out from body centre Dynamic and high tension in the movement Tension builds up and then "explodes"
Fear	Frequent tempo changes Long stops between changes Movements kept close to body centre Sustained high tension in movements
Grief	Long duration of time Few tempo changes, "smooth tempo" Continuously low tension in the movements
Joy	Frequent tempo changes Longer stops between changes Movements reaching out from body centre Dynamic tension in movements Changes between high and low tension

Table 5.2: association between expressive cues and conveyed basic emotions according to the hypotheses of the experiment (Camurri, Lagerlöf, Volpe, 2003).

6. Automated extraction of expressive cues[1]

This Chapter illustrates the techniques that have been developed in order to extract expressive cues from human full-body movement. In particular, such techniques have been applied to the twenty dance performances included in the reference archive recorded for the experiment described in Chapter 5. Most of these expressive cues have then been employed to classify fragments (motion strokes) of the dance performances with respect to the four basic emotions anger, fear, grief, and joy (see Chapter 7).
According to the sources and the research hypothesis outlined in Chapter 5, such expressive cues include:

- Global measures in the Personal Space (i.e., cues describing the movement of the full body) such as global amount of detected motion, amount of contraction/expansion, orientation of the body (i.e., an elliptical approximation of the body silhouette has been used and the orientation of the axes has been considered as approximating the orientation of the body), overall motion direction;
- Measures inspired to psychological researches such as Boone and Cunningham's global amount of upward movement or measures involving the dynamic of the contraction/expansion of the body (e.g., the amount of time limbs are kept close to the body);
- Cues inspired by the Rudolf Laban's Theory of Effort such as directness (i.e., how much the trajectory of a movement is direct or flexible), impulsiveness, fluency, or coming from more recent studies based on Laban's Theory (e.g., Zaho 2001).
- Cues inspired by analogies with audio analysis, e.g., Inter Onset Intervals, frequency analysis;
- Kinematical measures such as velocity, acceleration, average and peak velocity and acceleration.

The expressive cues and the algorithms developed to extract them will be now presented with reference to the layered conceptual framework discussed in Chapter 3 and instantiated on the particular task of analysis of expressive gesture in human full-body movement (see Figure 6.1).
The techniques here described together with modelling techniques such the one discussed in the next Chapter (decision trees) or other data mining and machine learning techniques (e.g., neural networks, support vector machines, multiple regression, fuzzy sets) are the basic bricks for building the Expressive Gesture Analysis (EGA) component of a virtual or mixed subject (observer) in a MIEE.
All the algorithms have been implemented as a collection of software modules for the EyesWeb open architecture (see www.eyesweb.org, Appendix A, and Camurri, Coletta, Peri, Ricchetti, Ricci, Trocca, Volpe, 2000). In particular, they constitute the core of the EyesWeb Expressive Gesture Processing Library (see Appendix B, and Camurri, Mazzarino, Volpe, 2003).

[1] The algorithms illustrated in this Chapter have been partially discussed in several papers, see for example (Camurri, Trocca, Volpe, 2002) and (Camurri, Lagerlöf, Volpe, 2003).

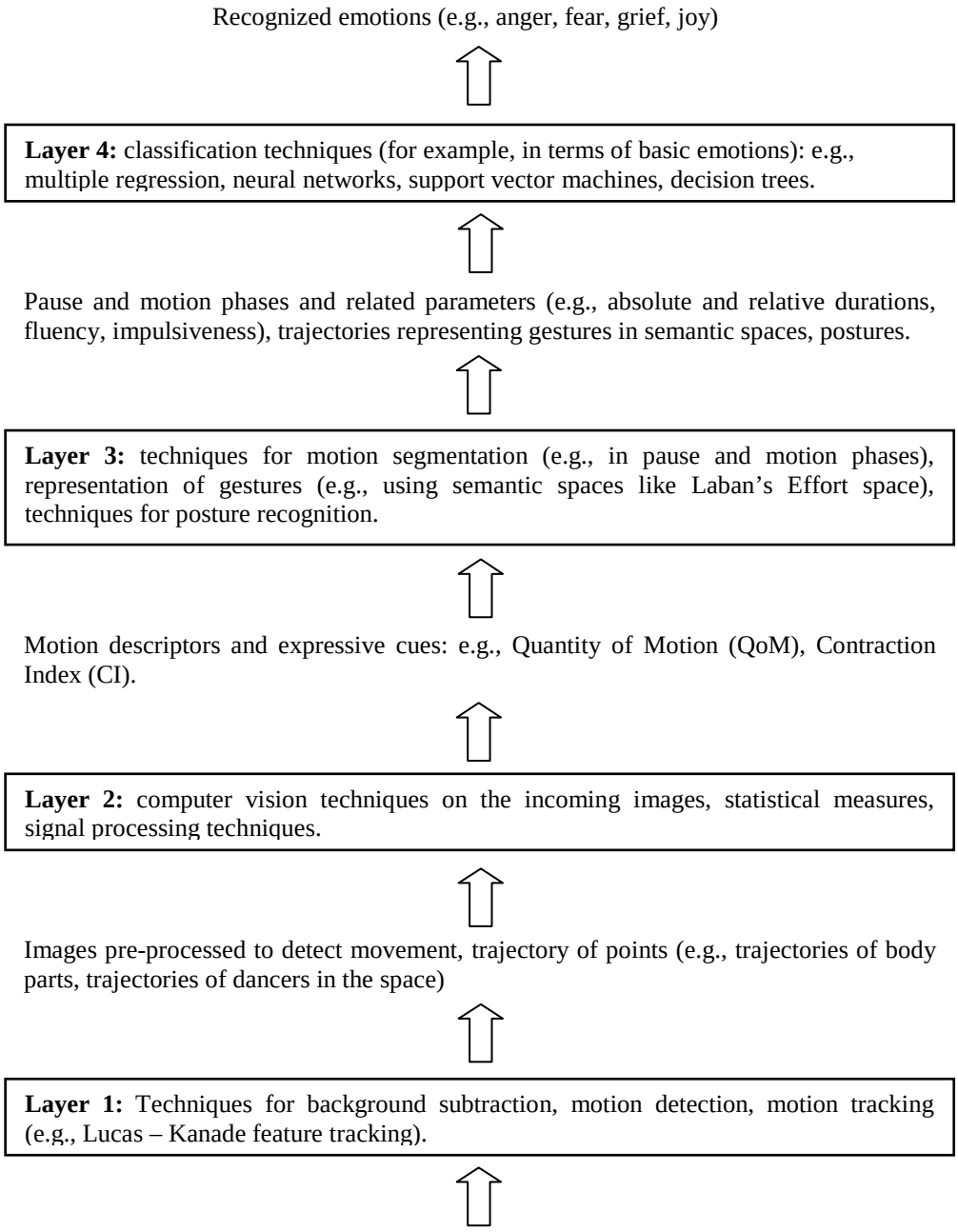


Figure 6.1: the conceptual framework described in Chapter 3, instantiated for analysis of expressive gesture in human full-body movement and, in particular, in dance performance (Camurri, Lagerlöf, Volpe, 2003).

6.1. Layer 1: processing of physical signals

Layer 1 is responsible of processing the information coming from sensors in order to detect and obtain information about the motion that is actually occurring[2]. It receives as input images from one or more videocameras and, possibly, information from other sensors (e.g., accelerometers).
A common set-up that is often employed due to the relative easiness in preparing it and in the management of the incoming information consists in acquiring images with just one videocamera (monocular vision) at 25 fps not interleaved.
Set-ups with two or more videocameras can also be considered. They range from systems employing two videocameras (e.g., for stereoscopic vision) to systems using many of them. For example, well-known motion capture systems use a quite huge number of videocameras (e.g., 12 or 14) disposed along a circle positioned around the location where movements are going to be performed. In this case, techniques (sometimes quite computationally expensive and not real-time) are needed for integrating information coming from each videocamera.
Sometimes it is also possible to use on-body markers and sensors (e.g., accelerometers), even if it should be noticed that the particular application field (artistic performances) often does not allow the use of on-body sensors since it would be too much constraining and disturbing for dancers.
For the sake of easiness, here algorithms are illustrated with reference to the simplest set-up, i.e., just one fixed videocamera acquiring not interleaved frames at the frame rate of 25 fps. Furthermore, since the analyses here discussed refer to the Personal Space, the movement of just one dancer is considered. However, it has to be noticed that most of the described techniques can be extended for possible use in more complex set-ups (at least 2 videocameras) and with more than one dancer.
Layer 1 generates two kinds of output: processed images (e.g., the silhouette of the dancer, see Figure 6.2) and trajectories of body parts (both points on the body without any specific reference to anatomical parts, and points representing the movement of specific joints or parts like head, hands, feet).
Layer 1 accomplishes its task by means of consolidated computer vision techniques usually employed for real-time analysis and recognition of human motion and activity: see for example the temporal templates technique for representation and recognition of human movement described in Bobick and J. Davis (2001). It should be noticed, however, that in contrast to Bobick and J. Davis research, here the aim is not at detecting or recognizing a specific kind of motion or activity.
In the following some basic techniques (e.g., background subtraction, feature tracking) usually employed at this step are shortly illustrated without going in too many details that would be outside the scope of this dissertation. Detailed descriptions can be found in computer vision books and in papers from the computer vision community (like for example the cited Bobick and J. Davis, 2001).

[2] In a way Layer 1 can be considered as a layer of virtual sensors, i.e., including both the employed physical sensors (e.g., videocameras) and the algorithms used to extract a given set of low-level data.

6.1.1. Silhouette extraction

A first step that is often (even if not always[3]) needed and sometimes critical is the extraction of the silhouette of the dancer. Background subtraction techniques can be employed allowing separating the silhouette of the dancer from the background. The most general formula for background subtraction is:

Silhouette[*t*] = *Threshold*(*Frame*[*t*] - *Background_Image*)

being *Frame*[*t*] a frame acquired at instant *t*, *Background_Image* an image of the background without any foreground object (i.e., in this case no dancers), and *Threshold* a function that given the difference image extracts from it pixels belonging to a certain range of value.
In the simplest case the background image is a static image, recorded once and never updated, that holds a picture of the background. Figure 6.2 shows a dancer's silhouette extracted with EyesWeb by using this method (the output image is a monochromatic representation of the silhouette, a median filter has been used to eliminate noise).
There are two major drawbacks in this simple implementation that can be extremely critical in artistic applications: light conditions (changes in lights produce a degradation of the performances), or, worse than that, details of the background can change. In such cases performances can degrade fast requiring at least a re-calibration of the threshold value or the acquisition of a new background image.

(a) (b)

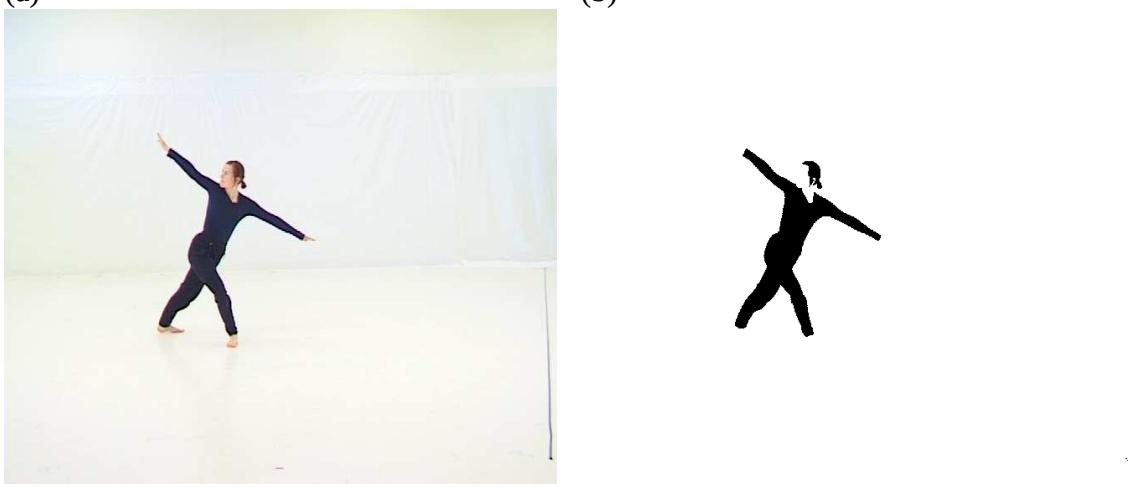


Figure 6.2: silhouette extraction using EyesWeb: (a) the incoming video frame; (b) the dancer's silhouette extracted with EyesWeb.

[3] For example if a colour-tracking algorithm is employed, usually background subtraction is not needed.

Improvements can be obtained by acting on two different aspects: the background image and the threshold function.
As for the background image, a main improvement consists in continuously updating it instead of using a static image. Several strategies can be adopted, considering and mixing the incoming frames in a new background image: for example, the background image can be updated with portions of the current frame that are recognised to belong to the background, or the output of an integrator operating on the grabbed frames can be used to generate a new background image from the incoming frames. Algorithms for adaptive background subtraction can be found in the computer vision literature. More complex approaches make also use of statistical models of the background
The threshold function can be made more complex in order to improve performances. For example, it can consider the background image in order to apply different thresholds to different areas in the image. This lead to a threshold function defined as follows:

Threshold(*Frame*[*t*] - *Background_Image, Background_Image*)

This kind of function uses available knowledge about the background in order to detect the silhouette. For example, the threshold can be more restrictive in case of bright backgrounds and less restrictive in case of dark ones. Many different thresholds can in principle be applied to differently dark and light areas. Virtually, it would possible to have up to 256 different thresholds in b/w images coded with 8 bits per pixel, even if the background subtraction process would soon become unmanageable.
Research on background subtraction is still a very active field in computer vision and the most recent developments can be found in computer vision journals and in the proceedings of the main computer vision conferences.

6.1.2. Silhouette Motion Images (SMIs)

A straightforward use of the dancer's silhouette extracted through the previously described background subtraction techniques is represented by Silhouette Motion Images (Trocca, 2001; Camurri, Trocca, Volpe, 2002). A Silhouette Motion Image (SMI) is an image carrying information about variations of the silhouette shape and position in the last few frames. SMIs are generated by the following formula:

$$Silhouette\_Motion\_Image[t] = \left\{ \sum_{i=1}^{n} Silhouette\ [t-i] \right\} - Silhouette[t]$$

The SMI at frame *t* is generated by adding together the silhouettes extracted in the previous *n* frames and then subtracting the silhouette at frame *t*. The resulting image contains just the variations happened in the previous frames. If *n* is the number of frames on which the SMI is calculated and $n = 1$, then the SMI carries information about the instantaneous variations of the silhouette. Working with a higher *n* allows capturing more information about the shape of motion and results are smoother, because the effect is

similar to average filtering. Figure 6.3a shows an SMI with $n = 4$. In the figure the SMI is the grey area, while the darker contour shows the most recent silhouette.
SMIs are inspired to motion-energy images (MEI) and motion-history images (MHI) (Bradsky and J. Davis, 2002, Bobick and J. Davis, 2001). They differ from MEIs in the fact that the silhouette in the last (more recent) frame is removed from the output image: in such a way only motion is considered while the current posture is skipped. Thus, SMIs can be considered as carrying information about the "amount of motion" occurred in the last *n* frames. Information about time is implicit in SMIs and is not explicitly recorded.
An extension of SMIs can be considered, which also takes into account the internal motion in silhouettes (see Figure 6.3b). In such a way it is possible to distinguish between global movements of the whole body in the General Space and internal movements of body limbs inside the Personal Space.

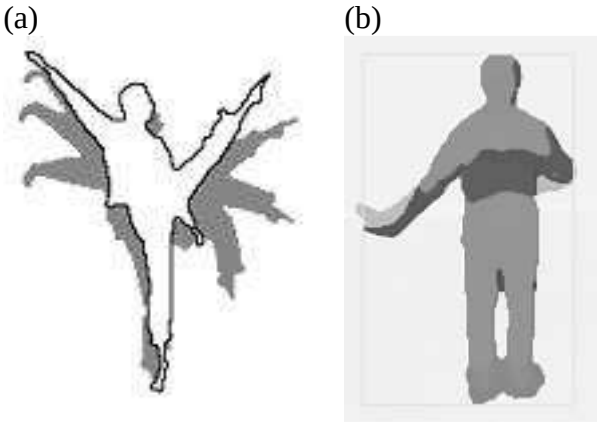


Figure 6.3: examples of Silhouette Motion Images (SMIs): (a) an SMI with time window (*n*) of 4 frames; (b) measure of internal motion using SMIs (the dark grey area inside the silhouette).

6.1.3. Motion tracking

Motion tracking is a very wide field in the computer vision literature. Here only the Lucas-Kanade tracking algorithm (Lucas and Kanade, 1981) is shortly described since it has been employed for the experiment discussed in this dissertation. It provides as output the trajectories of a redundant number of points randomly positioned on the moving body: no in formation about position of joints or body parts is available, rather the obtained trajectories can be processed in order to extract some global measures (e.g., speed calculated as average on all the trajectories). However, it should be noticed that other techniques such as for example skin colour tracking are available in EyesWeb for extracting positions and trajectories of specific body parts (e.g., hands and head).
The Lucas-Kanade feature-tracking algorithm allows tracking the movements of a certain number of points in a sequence of images. It is often used to track the movement of a

mobile camera while recording a scene with fixed objects. Anyway, it can be used with interesting results also to track moving objects filmed by a fixed camera.
The algorithm works as follows:

1. Select points in the image *n* that seems to be easy to track, i.e., well defined edges (Shi and Tomasi, 1994)
2. At image (frame) *n+1*, calculate the local optical flow in the neighbour of the selected points.
3. For each point estimate the new position and the reliability of such prediction. If reliability is under a certain threshold the point is considered lost.
4. Increment *n* and repeat step 2 and 3.

Usually the user specifies how many points have to be tracked and the algorithm attempts to follow them. However, the procedure that finds features classified as "good to track" can find less than the specified amount of points and such points can be marked lost after a few frames. It is therefore necessary to call again the procedure that selects points that can be tracked reliably and, among the proposed points, selects a few of them in order to maintain almost constant the number of tracked features. This step is called substitution.
This algorithm has been implemented in EyesWeb using the Open Computer Vision Library (OpenCV). However, at the moment of the implementation the OpenCV library did not provide a function that can directly perform feature substitution. This has been done by calling again the function that finds the "good to track" features and selecting in the resulting list those that are above a minimum distance from those already tracked.
The main disadvantage of the Lucas-Kanade feature tracking is that the selected points can fall everywhere in the image, i.e., either on the background or on the body of the dancer. In order to guarantee that some of the selected points will fall on the dancer, it is necessary to track a high number of points. Furthermore it is impossible to know where, on the body of the dancer those points are, and those that are attached to the background are a major waste of resources. It is possible to discriminate between points on the background and on the body by observing their velocities (points on the background should be still, even if noise can make this statement false). Another problem is point substitution. If the substitution happens while the dancer is moving, the blur produced by motion may prevent new points to be attached to the body.
These problems have been partially solved by combining motion tracking with a background subtraction module that extracts the dancer's silhouette. The silhouette is used as a mask to extract from the original frame just the image of the dancer. The resulting image is then sent to the tracker. The silhouette can also be used to estimate if the dancer is moving or not, and the point set can be completely or partly re-initialised when the dancer is still or moving slowly.
Figure 6.4 shows the output of the Lucas-Kanade feature tracker included in the EyesWeb Expressive Gesture Processing Library. In particular, Figure 6.4a shows the points that have been selected for tracking, while Figure 6.4b displays the trajectories of the tracked points (in a time window of 1 s).
Information motion detection and tracking provides to the upper levels is therefore encoded in two different forms: positions and trajectories of points on the body (as the

output of the Lucas-Kanade feature tracker), and images directly resulting from the processing of the input frames (e.g., dancer's silhouettes, SMIs).

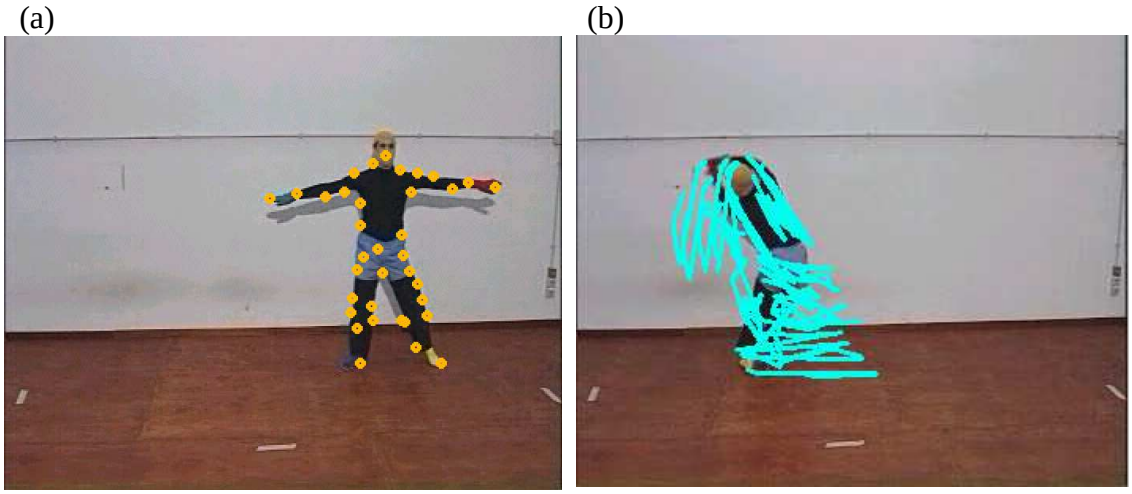


Figure 6.4: Lucas-Kanade feature tracking: (a) the points that have been selected for tracking; (b) the trajectories of the tracked points.

6.2. Layer 2: motion descriptors and expressive cues

Layer 2 is responsible of the extraction of a set of motion descriptors and expressive cues from the data coming from low-level motion detection and tracking. Its inputs are the processed images and the trajectories of points on body coming from Layer 1. Its output is a collection of motion cues describing movement and its qualities. To accomplish its task, Layer 2 employs computer vision, statistical, and signal processing techniques. The expressive cues that have been employed in the discussed experiment are now reviewed and algorithms for their extraction described.

6.2.1. Quantity of Motion (QoM)

Quantity of Motion (QoM) is computed as the area (i.e., number of pixels) of an SMI (e.g., the number of pixels in the grey area in Figure 6.3a). It can be considered as an overall measure of the amount of detected motion, involving velocity and force. QoM can be thought as a first and rough approximation of the physical momentum $q = m \cdot v$, where m is the mass of the moving body and v stands for its velocity. The shape of the QoM graph is close to the shape of the graphs of velocity of a marker put on a limb. QoM has two main problems: (i) the measure depends on the distance from the camera; (ii) difficulties emerge when comparing measures from different dancers. Such problems

can be (at least partially) solved by scaling the SMI area by the area of the most recent silhouette. The following formula is thus obtained:

Quantity_of_Motion = Area(SMI[t, n]) / Area(Silhouette[t])

In this way, the measure becomes relative, i.e., independent from the camera's distance (in a range depending on the resolution of the videocamera), and it is expressed in terms of fractions of the body area that moved. For example, it is possible to say that at instant *t* a movement corresponding to the 2.5% of the total area covered by the silhouette happened.

6.2.2. Contraction Index

The Contraction Index (CI) is a measure, ranging from 0 to 1, of how the dancer's body uses the space surrounding it in terms of contraction/expansion of the body with respect to its centre of gravity. For example, Figure 6.5 shows two conditions characterized by different values of the Contraction Index: a high value (near to 1) in Figure 6.5a where the body fills almost completely the rectangle enclosing it (usually called "bounding rectangle"), a low value (near to 0) in Figure 6.5b where limbs (especially arms) are kept quite far from the centre of gravity.

(a) (b)

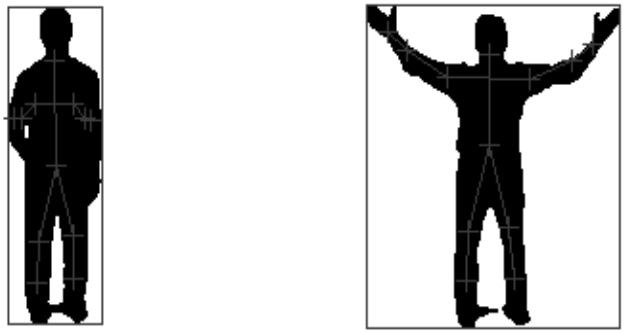


Figure 6.5: two conditions characterized by different values of the Contraction Index: a high value (near to 1) in (a) and a low value (near to 0) in (b).

The algorithm computing the CI combines two different techniques: the individuation of an ellipse approximating the body silhouette and computations based on the bounding rectangle. The former is based on an analogy between image moments and mechanical

moments: in this perspective, the three central moments of second order build the components of the inertial tensor of the rotation of the silhouette around its centre of gravity: this allows to compute the axes (corresponding to the main inertial axes of the silhouette) of an ellipse (see Figure 6.6) that can be considered as an approximation of the silhouette: eccentricity of such an ellipse is related to contraction/expansion; orientation of the axes is related to the orientation of the body (Kilian, 2001). The second technique used to compute CI is related to the bounding rectangle, i.e., the minimum rectangle surrounding the dancer's body (see Figure 6.5). The algorithm compares the area covered by this rectangle with the area currently covered by the silhouette. Intuitively (see Figure 6.5a and b), if the limbs are fully stretched and not lying along the body, this component of the CI will be low (near to 0), while, if the limbs are kept tightly nearby the body, it will be high (near to 1). While the dancer is moving the CI varies continuously. Even if it is used with data from only one camera, its information is still reliable, being almost independent from the distance of the dancer from the camera. A use of this cue consists of sampling its values at the end and the beginning of a movement stroke, in order to classify that movement as a contraction or expansion.

Figure 6.6: computation of an ellipse approximating the dancer's silhouette. The axes of the ellipse allow estimating the orientation of the body; its eccentricity is related to body contraction/expansion

A further use of the Contraction Index consists in analysing its dynamics along time (for example by computing the amount of time CI remained above a given threshold, i.e., the body has been contracted, during a motion stroke). This kind of analysis is related to one of the expressive cues individuated by Boone and Cunningham (1998): the duration of time arms are kept close to the body, that could be generalized in order to take into

account not only the movement of the arms, but also the contraction/expansion of the whole body.

6.2.3. Features extracted from motion trajectories

Layer 1 provides Layer 2 with processed images (e.g., SMIs) and trajectories of points located on the dancer's body (e.g., the output of the Lucas-Kanade feature tracker). QoM and CI are example of expressive cues extracted from processed images. Expressive cues can also be obtained from motion trajectories, like for example cues depending on the shape and the geometry of the analysed trajectory (e.g., direction, length, directness) and kinematical cues (e.g., velocity, acceleration). Here some of the geometrical cues will be introduced while in the following cues related to motion kinematics will be discussed.
Two of the most straightforward cues that can be extracted from trajectories recorded during a motion stroke are motion length and motion direction.
Motion length is computed by adding together the lengths of the segments composing the trajectory (i.e., all the segments joining two subsequent points of the sampled trajectory). It can give indirect indication about the complexity of a movement, its directness, and its dynamics (e.g., if it has been either fast or slow, or, using Laban's terminology, quick or sustained). However, information motion length provides usually needs to be integrated with other cues in order to be able to draw some conclusions about these properties. For example motion length contributes to a more reliable measure of motion directness: the Directness Index (DI).
The Directness Index for a motion trajectory is calculated as the ratio between the length of the straight trajectory connecting the first and the last point of the motion trajectory and the sum of the lengths of each segment constituting the motion trajectory (i.e., the motion length). Therefore, the more the Directness Index is near to one, the more direct is the motion trajectory (i.e., the motion trajectory is "near" to the straight one). Further aspects can be taken into account in order to improve the computation of the Directness Index, e.g., the deviations of the sampled points of the motion trajectory with respect to the straight one can be calculated and their average and standard deviation analysed. The Directness Index can contribute to the analysis in the Laban's dimension of Space, i.e., its values can be used for a first rough estimate of how much a motion stroke is direct or flexible.
Motion direction is calculated by measuring the angle of the vector joining the first and last sampled points of the motion trajectory in the last n frames. The selected value of n determines how much motion direction refers to an instantaneous or short-term direction (low values of n) or to a sort of average direction during a motion stroke (high values of n or direction calculated on a whole motion stroke). Motion direction is related to another expressive cue that Boone and Cunningham (1998) found to be relevant for expressive content decoding: the frequency of upward arm movement. In this case too the original cue by Boone and Cunningham can be generalised to consider not only arm, but also full-body upward movements. For example, the duration of time in a motion stroke in which upward movement has been detected can be computed and considered as a tendency of the motion stroke to be upward directed.

As an example, Figure 6.7 shows a sampled trajectory (red line) and the computed motion direction (the green segment, whose length is proportional to the overall displacement).

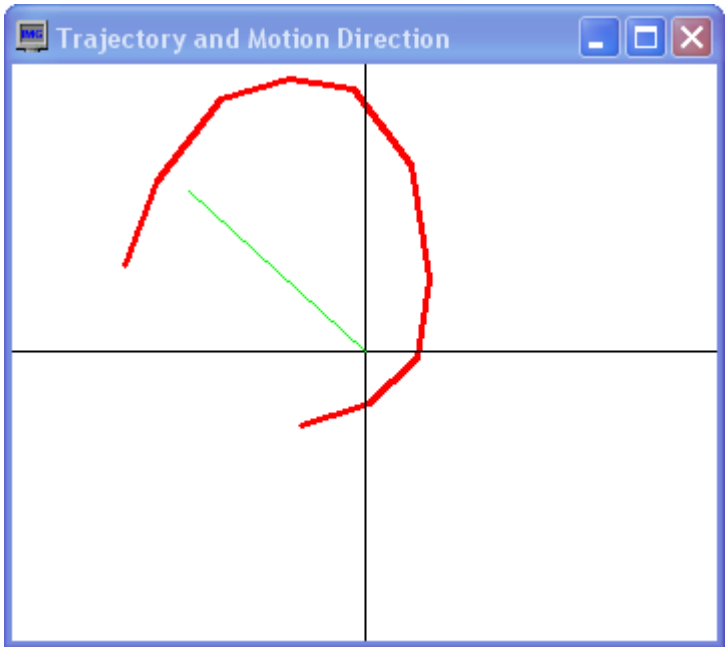


Figure 6.7: a motion trajectory and the computed motion direction

In the case that many motion trajectories are available at the same time (e.g., the Lucas-Kanade feature tracker can provide a redundant number of points distributed over the whole body), averages can be computed on them in order to obtain cues reflecting the global properties (i.e., motion length, direction, directness) of the analysed stroke.

6.2.4. Kinematical measures

Trajectories can also be analysed with respect to kinematical aspects. In particular, velocity and acceleration can be extracted by using well-known techniques for approximated numerical derivative. For example, the software modules included in the EyesWeb Expressive Gesture Processing Library allow calculating velocity and acceleration by using either the asymmetric backward numeric derivative or the symmetric numeric derivative whose formulas (for velocity) are reported here below. Notice that in the formulas x and y stand for the x and y coordinates of the position of a sampled point in the motion trajectory, while i is the index related to time: so $x(i)$ is the x coordinate of the currently sampled point while $x(i\text{-}1)$ is the x coordinate of the last sampled point. Notice also that symmetric numerical derivative introduce a delay of one sample, i.e., 40 ms in case the usual sampling frequency of 25 fps is employed.

Asymmetric backward numeric derivative (for velocity):

$$\begin{cases} v_x(i) = \dfrac{x(i) - x(i-1)}{\Delta t} \\ v_y(i) = \dfrac{y(i) - y(i-1)}{\Delta t} \end{cases}$$

Symmetric numeric derivative (for velocity):

$$\begin{cases} v_x(i-1) = \dfrac{x(i) - x(i-2)}{2\Delta t} \\ v_y(i-1) = \dfrac{y(i) - y(i-2)}{2\Delta t} \end{cases}$$

A low-pass filter can be applied in order to reduce the noise introduced by the numerical derivative operation.
As already explained for motion length, direction, and directness, if many motion trajectories are available at the same time (e.g., from the Lucas-Kanade tracker) further computations can be carried out to extract global motion features. In particular descriptive statistics (e.g., average, standard deviation, maximum) can be calculated:

(i) *Along time*: for example, average and peak values calculated either on running windows or on all the samples in a given time interval (e.g., the average velocity of the hand of the dancer during a given motion stroke)
(ii) *Among trajectories*: for example, average velocity of groups of trajectories available at the same time (e.g., the average instantaneous velocity of all the tracked points located on the arm of a dancer).

6.3. Layer 3: gesture segmentation and representation

Layer 3 is responsible of segmenting motion in order to individuate motion strokes, i.e., motion and non-motion (pause) phases.
It is also in charge of extracting further higher-level expressive cues that are the result of an analysis of the segmented movement and of the obtained sequence of motion and pause phases. Examples of such cues are the temporal duration of motion and pause phases compared with the total duration of the dance performance, impulsiveness, and fluency.
Segmentation and cue extraction is performed from the input (i.e., the expressive cues) coming from Layer 2. In a certain extent, motion and pause phases can be associated with movement gestures[4]. The output of Layer 3 can thus be considered as a first representation of gestures in term of values of expressive cues associated with them.

[4] As already discussed in Chapter 3, associating motion and pause phases with gestures is quite a rough approximation since gestures can be observed shorter than a single motion phase or covering some of them. Anyway, it can be taken as a starting point for a first analysis that hopefully will lead to understand more subtle aspects.

6.3.1. Motion segmentation

A straightforward way to individuate movement strokes and therefore to segment movement in motion and pause phases is to apply a threshold on the detected energy or amount of movement. As a first approximation, the QoM measure has been therefore used to perform such segmentation.
QoM is related to the overall amount of motion and its evolution in time can be seen as a sequence of bell-shaped curves (*motion bells*). In order to segment motion, a list of these motion bells has been extracted and their features (e.g., peak value and duration) computed. For this task an empirical threshold can be defined on the QoM: for example, according to a threshold that has been used in several applications, the dancer is considered to be moving if the area of his/her motion image (i.e., the QoM) is greater than 2.5% of the total area of the silhouette. Figure 6.8 shows motion bells after automated segmentation: a motion bell characterizes each motion phase.

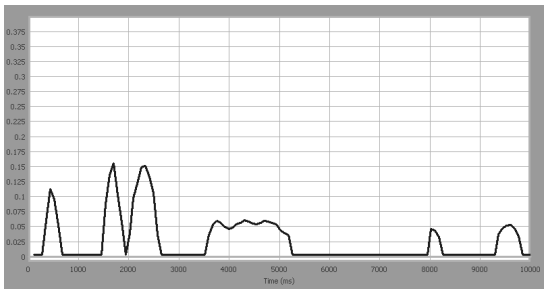


Figure 6.8: motion segmentation

The described segmentation technique based on only Quantity of Motion, even if sufficient in most cases in the experiment under exam, can be improved in several ways. From the one hand further cues can be taken into account to perform the segmentation task. For example, techniques based on analysis of the speed and acceleration profiles can be considered: (Bindiganavale, 2000) uses the zero-crossings of acceleration to detect changes in motion, (Zhao, 2001) uses a method based on zero-crossings of acceleration and curvature to segment the trajectories followed by the arm joints.
On the other hand, more detailed segmentation can be obtained. For example, Laban says that “almost any work-operation or expressive gesture shows the following pattern: preparation – one or several main efforts – termination” (Laban, 1963, p. 75). A further goal of segmentation is therefore to individuate sub-phases within a motion phase.
Finally, it should be remarked that segmentation algorithms based on extracted cues could obtain a segmentation that is different from the segmentation that a human observer could perform. In other words, detected segments could be different from perceived segments. Perceptual experiments would be needed in order to compare the motion phases obtained from algorithms with the motion phases perceived by humans. As a result of such experiments some kind of curve of perceived motion could be obtained replacing or modifying the curve of detected motion computed by the algorithms. This observation raises issues that are beyond the scope of this dissertation: it

is just worth to be noticed that while in more traditional application (e.g., video surveillance) and in application fields in which the focus is on measuring movement what is important is the detected motion, in applications like the ones envisaged in this work (i.e., expressive interaction in artistic contexts) it could be more relevant to consider the perceived motion.

6.3.2. Fluency and impulsiveness

Motion segmentation can be considered as a first step toward the analysis of the rhythmic aspects of dance. Analysis of the sequence of pause and motion phases and their relative time durations can lead to a first evaluation of dance tempo and its evolution in time, i.e., tempo changes, articulation (in analogy with music legato/staccato). Parameters from pause phases can also be extracted to individuate real still standing positions from active pauses involving low-motion (hesitating or oscillation movements).
Furthermore, motion fluency and impulsiveness can be evaluated. They are related to Laban's Flow and Time axes.
Fluency can be estimated starting from an analysis of the temporal sequence of motion bells. A dance fragment performed with frequent stops and restarts (i.e., characterized by a high number of short pause and motion phases) will result less fluent than the same movement performed in a continuous, "harmonic" way (i.e., with a few long motion phases). The hesitating, bounded performance will be characterized by a higher percentage of accelerations and decelerations in the time unit (due to the frequent stops and restarts), a parameter that has been demonstrated of relevant importance in motion flow evaluation (see, for example, Zhao 2001, where a neural network is used to evaluate Laban's Flow dimension).
A first measure of impulsiveness can be obtained from the shape of a motion bell. In fact, since QoM is directly related to the amount of detected movement, a short motion bell having a high pick value will be the result of an impulsive movement (i.e., a movement in which speed rapidly moves from a value near or equal to zero, to a peak and back to zero). On the other hand, a sustained, continuous movement will show a motion bell characterized by a relatively long time period in which the QoM values have little fluctuations around the average value (i.e., speed is more or less constant during the movement).

6.3.3. Gesture representation

Several kinds of representation can be envisaged. One possibility consists in producing a symbolic description of the analysed sequence of movements. This representation can be useful because it can be understood by a human researcher in a relatively easy way and also used by an automatic system. In fact, motion and pause phases would be represented as motion objects (or gestures) in analogy with music objects: they would be characterized by a beginning, an end, a time duration, and a collection of values of motion cues either the values of motion cues continuously collected during the whole

phase, or single values either summarizing a continuous cue (e.g., averages) or related to cues that can be directly calculated on a whole phase (e.g., motion length).
For example, depending on the Contraction Index a motion phase can be seen as a contraction phase (if the value of CI at the end of the phase is higher than the one at the beginning) or as an expansion phase. It is therefore possible to obtain a description like the following one:

Contraction(Start_Frame, Stop_Frame, Initial_Value_CI, Final_Value_CI, other cues…)
Expansion(Start_Frame, Stop_Frame, Initial_Value_CI, Final_Value_CI, other cues…)

Another possibility is to build a representation in terms of points or trajectories in multidimensional semantic spaces, i.e., spaces whose axes are expressive cues having a relevant influence with respect to the conveyed expressive content. As discussed in Chapter 3, whether the representation has to be a point or a trajectory depends on how the low-level features are processed in Layers 2 and 3. For example, if a vector containing the averages of the Layer 2 expressive cues is calculated along the time duration of a motion phase (gesture) or a motion phase is considered as a single event, the gesture/motion phase could be represented as a point in the multidimensional space. If instead more values for each cues are available (e.g., local values, or averages along sub-phases) or if a gesture is considered as a sequence of events (as it is likely to be) a trajectory is a more appropriate representation.
Figures 6.9a and b show an example of such kind of representation in a 2D space.

(a) (b)

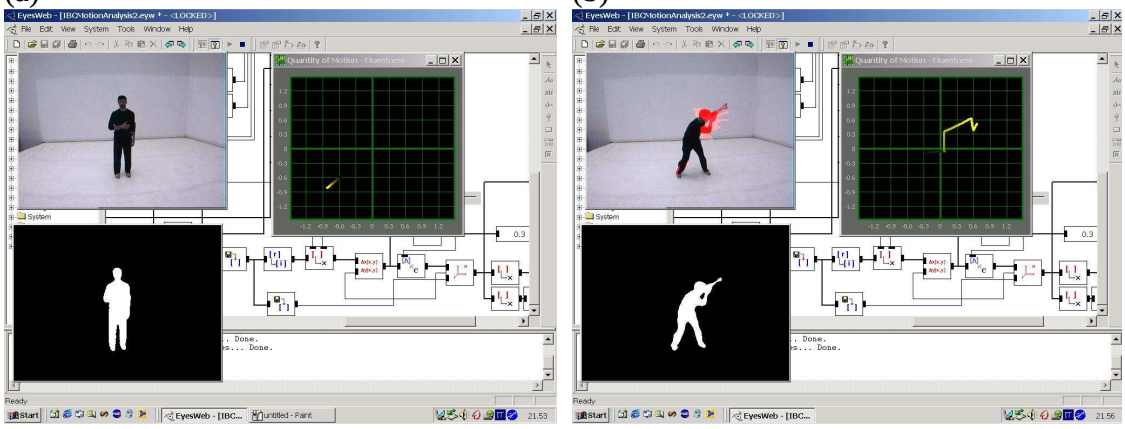


Figure 6.9: representation of gestures as trajectories in a 2D space.

The two dimensions are Quantity of Motion and fluency. In Figure 6.9a the dancer is not moving: the current position in the space (window in the right) is moving toward the bottom left parts of the 2D space (yellow stripe), a position characterized by low QoM and low fluency (i.e., the amount of pause phases is dominating the amount of motion phases). In figure 6.9b, a high-energy gesture is displayed. The red shadow around the dancer (the SMI) in the upper-left window of Figure 6.9b is proportional to the QoM and

the position in the space (the yellow stripe in the right window) is moving toward the top-right region in that window, characterized by high QoM and high fluency.
It should be noticed that this is just an example. In fact, a semantic space would need experimental results in order to be established: its dimensions have to be proved to be meaningful and possibly uncorrelated. However, the two expressive cues here indicated (QoM and fluency) are good candidate dimensions for such an expressive space being the former related to energetic aspects of movement and the latter to temporal, rhythmical aspects, even if at the moment no evidence has still been provided to support this hypothesis.
Once gestures are represented as trajectories in a semantic multidimensional space cues can be extracted from these trajectories. In particular, it is possible to extract the cues previously discussed with respect to motion trajectories also from trajectories of expressive gestures in semantic spaces. The values thus obtained can be used as input to algorithms (e.g., clustering techniques) for grouping similar trajectories, i.e., similar gestures, in order to interpret them. Notice that if motion gestures and music gestures could be represented in the same (or a similar) expressive space, algorithms could be used for grouping and analysing such gestures in a multimodal perspective.

6.3.4. Posture recognition

As already discussed in Chapter 5, not only motion is important in expressive content communication. Pauses also have a role of paramount importance. During a pause the body may assume a particular posture and body postures can be considered as expressive gestures having a relevant role in conveying expressive content to the audience (see for example Argyle, 1980).
Algorithms for posture recognition have thus been implemented in the EyesWeb Expressive Gesture Processing Library, even if not directly employed in the discussed experiment.
One of them, robust enough to be employed in real-time performances is based on Hu moments (Hu, 1962), a set of seven moments, which are translation, scale and rotation invariant, and have been widely used in computer vision for shape discrimination.
The algorithm employs a nearest-neighbour technique. For each considered (normalised) posture Hu moments are calculated and stored in a matrix. During each pause phase, Hu moments are calculated on the incoming (normalised) silhouette. Euclidean distances are computed between the Hu moments of the silhouette in the current frame and the Hu moments of each candidate posture accordingly to the following formula:

$$d_p(\underline{u},\underline{v}) = \left(\sum_{k=1}^{n} |u_k - v_k|^p \right)^{\frac{1}{p}}$$

where $\underline{u}$ and $\underline{v}$ are the Hu moment vectors (having $n = 7$ elements), and p is the degree of the distance. p can be provided as parameter to the algorithm. When $p = 1$, the 1-distance is obtained:

$$d_1(\underline{u},\underline{v}) = \sum_{k=1}^{n} |u_k - v_k|$$

When $p \leq 0$, the algorithm calculates the ∞-distance, i.e.,

$$d_\infty(\underline{u},\underline{v}) = \max_{k=1..n} |u_k - v_k| .$$

The posture corresponding to the minimum distance is candidate for recognition.
Mechanisms are provided to recognize a posture only if its duration in time is long enough to consider it effectively as a posture: each time a posture is recognized, the last N recognized postures (where N is provided as a parameter to the algorithm) are considered and compared against a threshold provided as parameter as well. The threshold represents a percentage of recognition during the last N recognitions: if the most frequently recognized posture among the last N postures has a percentage of recognition above this threshold, such a posture is recognized, otherwise no posture is recognized. For instance, if the threshold is set to 90% and N is set to 10, posture number 3 will be recognized only if the array of the indexes of the last 10 recognized postures contains the index 3 at least nine times. In this way, postures cannot be recognized if their time duration is too short with respect to the value of N, the threshold, and the sampling frequency. For instance, if N = 10, the threshold is 100%, and the frame rate is 25 Hz, a posture can be recognized only if its duration is longer than 10x(1/25) = 0.4 s.
The algorithm returns also a confidence index in the range [0, 1] describing how much it is confident to have correctly recognized a given posture. The confidence index is determined by comparing the value of the calculated minimum distance with the distance immediately larger than the minimum one, similar values of the two distances meaning an ambiguous recognition.
Figure 6.10 shows the five postures that have been used to test the algorithm.

Figure 6.10: five normalised postures used to test the discussed posture recognition algorithm.

7. Classification of dance fragments

The expressive cues described in Chapter 6 have been extracted from the 20 dance fragments recorded for the experiment sketched in Chapter 5. In this Chapter a model is introduced for such data attempting to classify them with respect to the four basic emotions (anger, fear, grief, and joy) corresponding to the dancers' expressive intentions. Dance fragments have been segmented in motion and pause phases as described in Chapter 6. To this aim, an empirical threshold on the QoM has been defined for these dances, corresponding to the 2.5% of the average value of the QoM computed along each whole dance fragment. A vector of 18 expressive cues has then been extracted for each motion phase. According to the research hypotheses described in Chapter 5, such cues include:

- Cues related to the time duration of motion and pause phases: duration of the current motion phase, duration of the last pause phase;
- Cues related to the amount of movement (energy): average, standard deviation, and peak value of the Quantity of Motion along a motion phase;
- Cues related to body contraction/expansion: average and standard deviation of the Contraction Index along a motion phase;
- Cues related to the use of space: Directness Index, length and overall direction of motion trajectories along a motion phase;
- Cues derived from the cited studies by Boone and Cunningham (1998): amount of upward movement, dynamics of the Contraction Index;
- Kinematical cues: average, standard deviation, and peak value of speed along a motion phase; average, standard deviation, and peak value of the module of acceleration along a motion phase;

For those cues depending on motion trajectories a Lucas-Kanade feature tracker has been employed. A redundant set of 40 points randomly distributed on the whole body has been tracked during each motion phase. Points have been reassigned before the beginning of the following motion phase so that a small and not significant amount of points is lost during tracking. Overall motion cues have been calculated by averaging the values obtained for each trajectory.
An explorative analysis has been carried out on the extracted cues. A decision tree has been built for classifying motion phases depending on the vector of extracted cues with respect to the four basic emotions that dancers tried to covey. The results from the model have been compared with spectators' ratings collected by the psychologists in Uppsala.

7.1. Explorative analysis

At a first stage the extracted variables have been subjected to an explorative analysis mainly consisting in calculating and analysing the descriptive statistics of each of them. When possible, the Analysis of Variance (ANOVA) has been performed with respect to

the four emotional categories. The box-plots have also been drawn and analysed. The results of such a preliminary analysis (part of which can be found in Mazzarino, 2002) are summarized in the following together with a discussion of some emerging aspects.

7.1.1. Quantity of Motion

Quantity of Motion (QoM) has been considered under three aspects: its average, standard deviation and peak value along a motion phase. Since QoM is related to the amount of detected motion, the three variables derived from it are related respectively to the average amount of motion during a motion phase, to the motion "dynamics", i.e., how much the amount of detected motion remained constant or varied in a motion phase, and to the maximum amount of detected motion.
The box-plot of the average of the QoM with respect to the four emotion categories is shown here below.

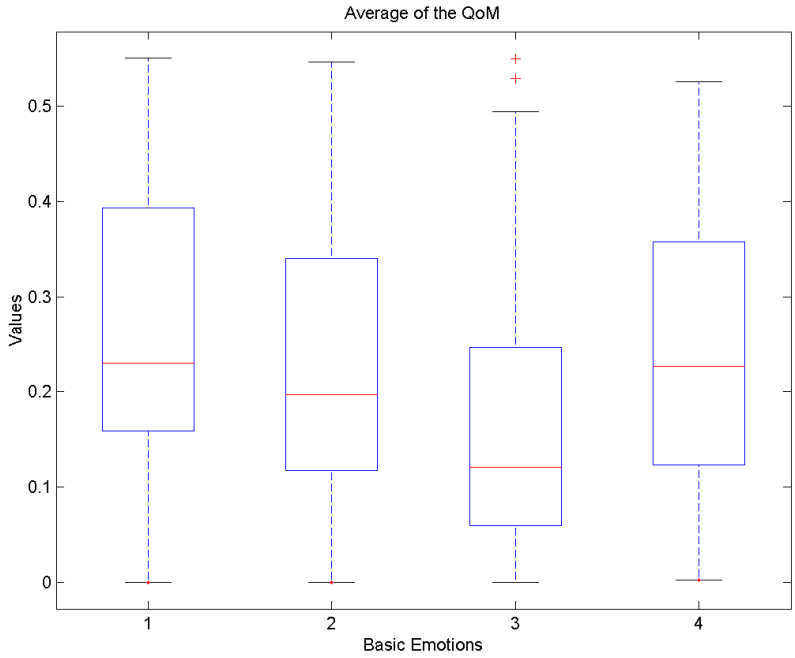


Figure 7.1: box-plot of the average of the QoM along the motion phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy.

The Analysis of Variance has been performed on the average of the QoM along motion phases (since according to the Central Limit Theorem the average of the QoM can be thought to tend to be normally distributed even if the QoM is not normally distributed). Results are displayed in Table 7.1 in the following page.

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Groups	0.53331	3	0.17777	88.852	1.12E-02
Error	66.024	330	0.020007		
Total	71.358	333			

Table 7.1: Analysis of Variance (ANOVA) for the averages of the QoM along motion phases

With a p-value in the order of 10^{-2} the average value of the QoM along motion phases therefore appears to be statistically significant for analysis.
The box-plots for the standard deviation and for the peak value of the QoM are shown in Figure 7.2 a and b respectively.

(a) (b)

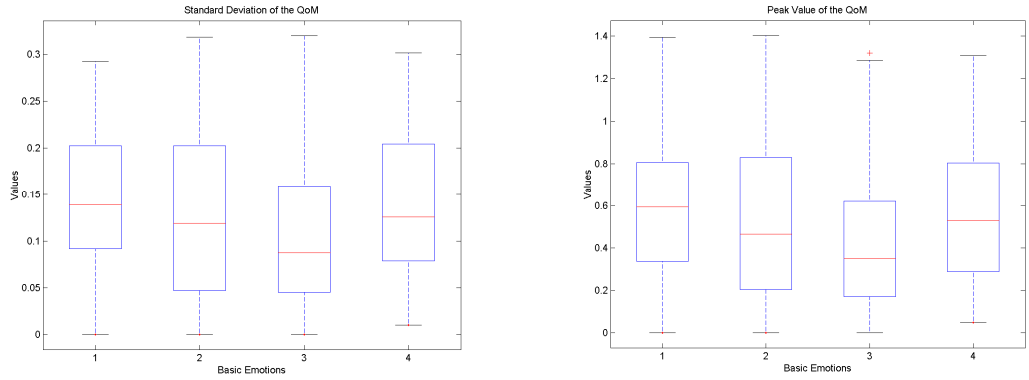


Figure 7.2: box-plots of (a) the standard deviation and (b) the peak value of the QoM along the motion phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy.

Significant differences can be noticed (even if not mathematically proved): for example grief results to have the lower standard deviation, i.e., the lower variation in the amount of detected motion, and the lowest peak value (that could mean a relative absence of impulsive strokes, i.e., a quite low-energy and sustained movement). The highest values are associated with anger and joy that seem to be the two emotions characterized by the highest dynamics along the energy dimension.
An analysis of the average of the QoM along the whole dances (i.e., the average of the QoM for each performance) has been carried out in (Mazzarino, 2002). The results are shown in Figure 7.3 in the following page.
By inspecting such results it is possible to observe that for example the performance of the first dancer satisfies the hypothesis according to which higher energy should be noticed in anger than in fear. In fact, the average of the QoM (along the whole performance) for the first dancer is highest in the anger performance (represented in blue) and lowest in the grief performance (in yellow).

Extending the observation to all the dancers it can be noticed that the lowest value of the QoM average is always associated with the performance conveying grief, but, at the same time, the highest value does not always correspond to the anger performance (this is not the case for dancers number two and four).

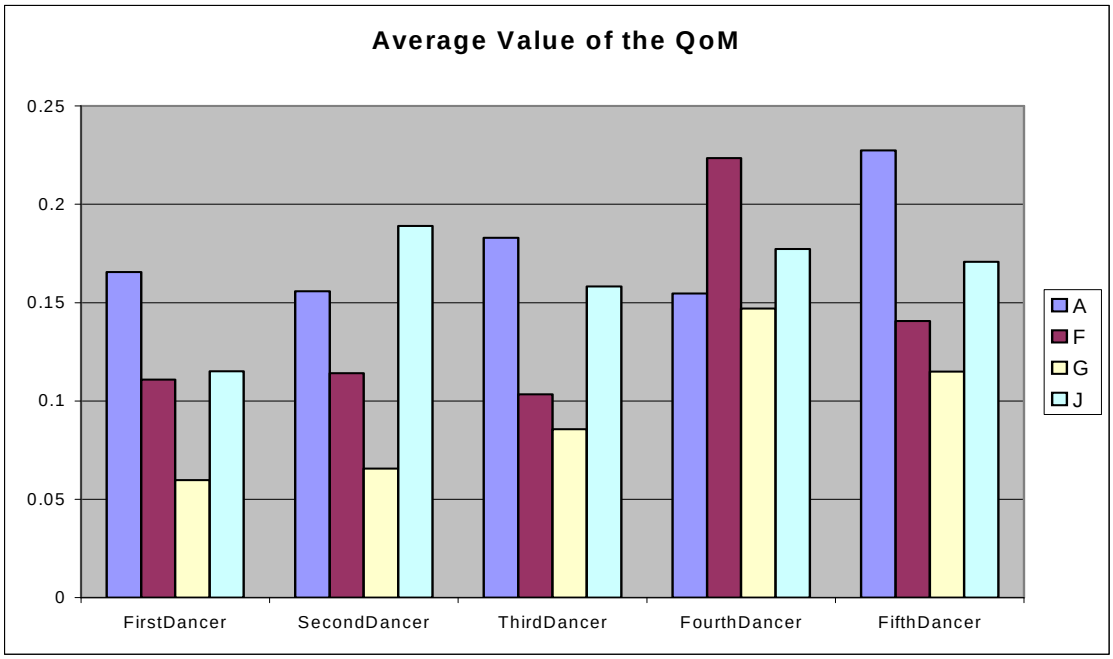


Figure 7.3: the average of the QoM for each performance (Mazzarino, 2002)

As a remark, from the one hand it has to be noticed that since the average along the whole performance is now discussed, this does not mean that the same pattern can be found in each motion phase: it is just a general tendency observed in the whole dance.
On the other hand, the ANOVA on the average of the QoM along each motion phase discussed above seems to confirm that such a general tendency is significant also at the level of each motion phase.

7.1.2. Contraction Index

Contraction Index (CI) has been considered with respect to its average and standard deviation along a motion phase.
Furthermore, the dynamics of the Contraction Index, intended as the duration in time in a motion phase the values of the CI have been above a given threshold has also been computed and measured as percentage with respect to the whole duration of the motion phase under exam. The threshold has been empirically selected in order to maximize the differences between the four basic emotions with respect to this expressive cue: this lead to select 0.7 as threshold for Contraction Index Dynamics.

CI is related to the contraction/expansion of the body with respect to its Centre of Gravity. The three variables derived from it are thus related to the average amount of contraction/expansion during a motion phase, and to how such contraction/expansion evolves during a motion phase, i.e., how much contraction/expansion remains constant or varies in a motion phase (the standard deviation of the CI), and how long the body remains contracted during a motion phase (the CI Dynamics). CI Dynamics can be considered as an extension of "the duration of time arms were kept close to the body" by Boone and Cunningham (1998).
The box-plot of the average of the CI with respect to the four emotion categories is shown here below.

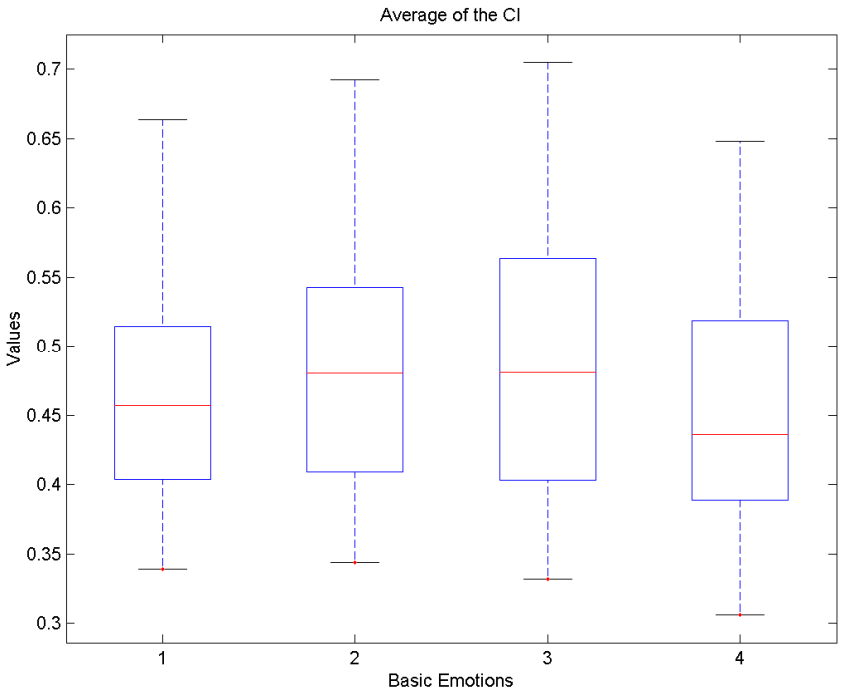


Figure 7.4: box-plot of the average of the CI along the motion phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy.

The Analysis of Variance has been performed on the average of the CI along motion phases. The results displayed in Table 7.2 in the following page show a p-value of 0.0527, i.e. a confidence of 94.73%. This is at the limit of statistical significance (usually a confidence of 95% is required).
In fact, the box-plot seems to indicate that the average of the CI mainly distinguishes among fear and grief (characterized by higher levels of contraction) and angry and joy (characterized by higher levels of expansion): it should be noticed that this result is however consistent with the research hypotheses stated in Chapter 5.

ANOVA Table					
Source	SS	Df	MS	F	Prob>F
Groups	0.0623	3	0.2077	2.59	0.0527
Error	2.64394	330	0.00801		
Total	2.70624	333			

Table 7.2: Analysis of Variance (ANOVA) for the averages of the CI along motion phases

The box-plots for the standard deviation of the CI and for the CI Dynamics are shown in Figure 7.5 a and b respectively.

(a) (b)

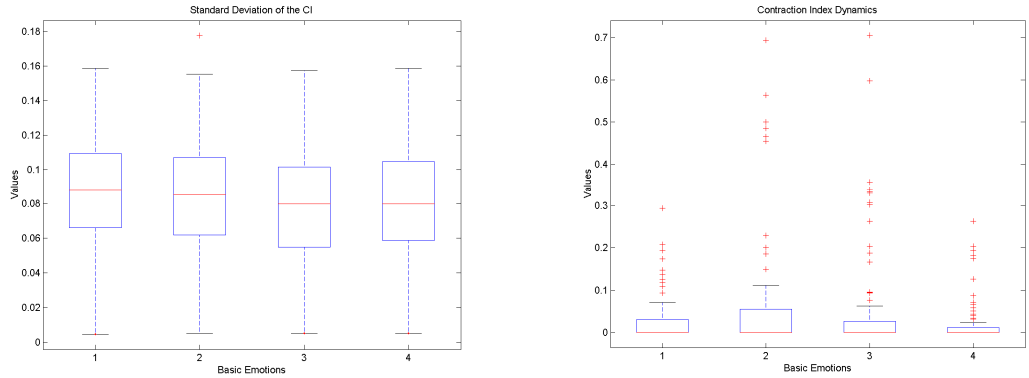


Figure 7.5: box-plots (a) of the standard deviation of the CI and (b) of the CI Dynamics along the motion phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy.

The standard deviation of the CI does not seem very significant: its values are very similar for all the four basic emotions. Contraction Index Dynamics looks more interesting: its values are higher for fear (i.e., the body remains contracted along more time) and lower for joy (whose motion is characterized by more expansions).
As for Quantity of Motion, the mean value of the Contraction Index has also been calculated along the whole performance. The analysis of the results (see Figure 7.6 in the following page) seems to confirm the psychologists' hypothesis according to which fear is characterized by a movement toward to the centre of the personal space (i.e., a contraction). Moreover, as psychologists expected, joyful performances have a low value of contraction index meaning that joyful movements are generally open and "expanding".
In conclusion, Contraction Index seems to be less significant than Quantity of Motion. This could be due to the fact that the choreography was predefined, and therefore contractions and expansions were pre-built in it: the differences that can be noticed would therefore be due to different ways to stress the predefined contractions/expansions depending to the expressive intention. Such differences, anyway, seem to be relevant enough to keep the Contraction Index and its derived cues in the analysis.

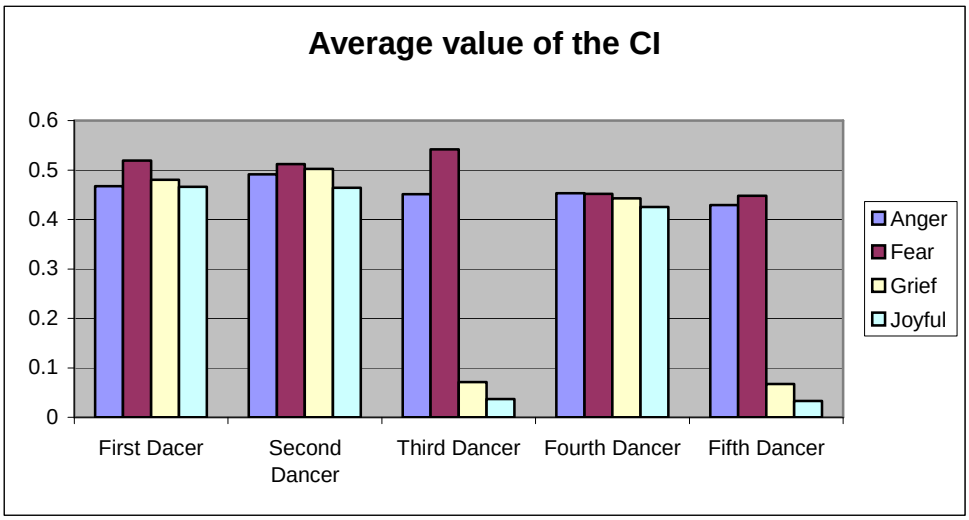


Figure 7.6: the average of the CI for each performance (Mazzarino, 2002)

7.1.3. Kinematical measures

Velocity and acceleration (in their *x* and *y* components) have been computed for the trajectories of 40 points returned by the Lucas-Kanade feature tracker during each motion phase. At the beginning of each motion phase the 40 points have been reassigned in order to avoid loosing too many points during tracking. The symmetric numeric derivative has been used for calculating velocity and acceleration. A low-pass filter has been applied to the result of the numerical derivative in order to reduce the noise introduced by this operation. An overall descriptor has been obtained by averaging the values on the 40 trajectories. The modules of such overall velocity (i.e., speed) and acceleration have then been computed. The average, standard deviation, and peak values along a motion phase of such measures have finally been considered as possible expressive kinematical cues.
The box-plot of the average of the speed along each motion phase is displayed in Figure 7.7 in the following page.
The Analysis of Variance has been performed on the average of the speed along motion phases whose results are displayed in Table 7.3. With a p-value of 0.0042 this cue appears to be statistically significant for the analysis[1].
The box-plots for the standard deviation and for the peak value of the speed are shown in Figure 7.8 a and b respectively.

[1] In fact, this should not be surprising since Quantity of Motion, which is in many aspects related to speed, already resulted significant. It should be noticed however that the p-value for speed results higher that the p-value for Quantity of Motion.

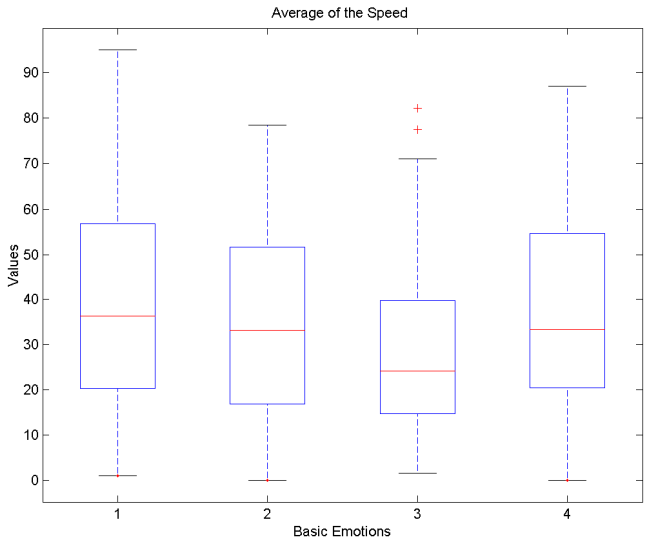


Figure 7.7: box-plot of the average of the speed along the motion phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy. Speed is measured in [pixels/s].

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Groups	5579.6	3	1859.95	4.48	0.0042
Error	137039	330	415.27		
Total	142618.7	333			

Table 7.3: Analysis of Variance (ANOVA) for the averages of the speed along motion phases

(a) (b)

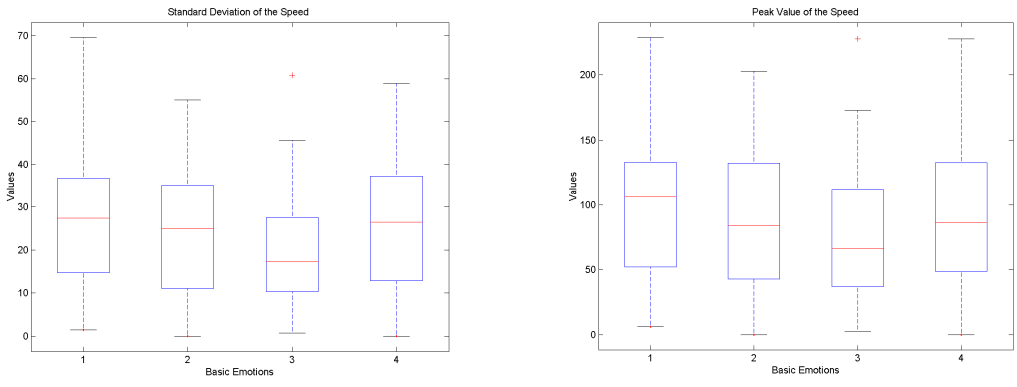


Figure 7.8: box-plots of (a) the standard deviation and (b) the peak value of the speed along the motion phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy. Speed is measured in [pixels/s].

Significant differences can be noticed (even if not mathematically proved): for example grief results to have the lowest standard deviation, i.e., the lowest variation in speed, and the lowest peak value. The highest values are associated to anger that, according to the initial hypotheses, should be characterized by the most impulsive movements.
These observations should be also confirmed by an analysis of acceleration.
The box-plot of the average of the module of the acceleration along each motion phase is displayed in Figure 7.9.

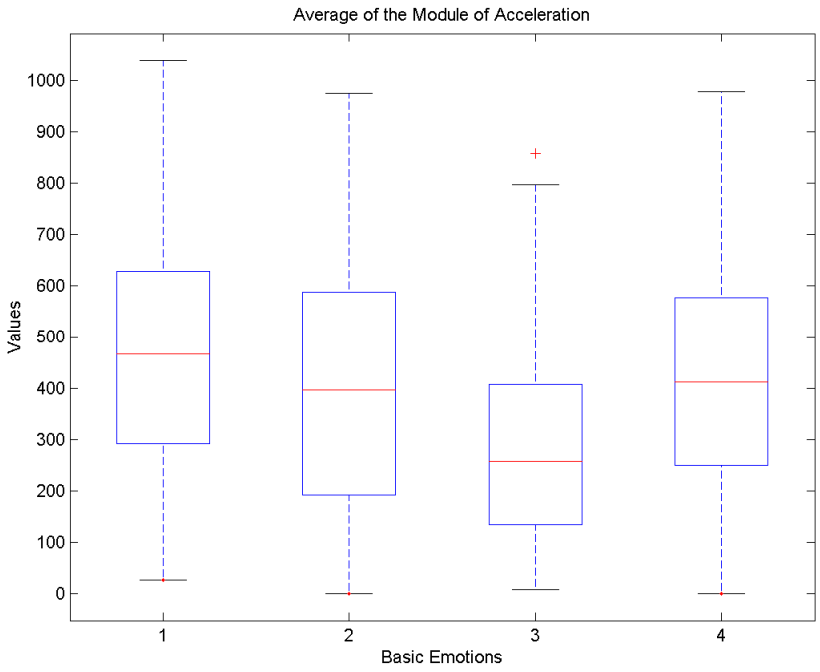


Figure 7.9: box-plot of the average of the module of the acceleration along the motion phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy. Acceleration is measured in [pixels/s^2].

The Analysis of Variance has been performed on the average of the module of the acceleration along motion phases, whose results are displayed in Table 7.4.

ANOVA Table					
Source	SS	Df	MS	F	Prob>F
Groups	$1.40891 \cdot 10^6$	3	469637.7	9.27	$6.6607 \cdot 10^{-6}$
Error	$1.67159 \cdot 10^6$	330	50654.1		
Total	$1.81248 \cdot 10^6$	333			

Table 7.4: Analysis of Variance (ANOVA) for the averages of t of the module of the acceleration along motion phases

With a p-value of $6.6607 \cdot 10^{-6}$, this cue results statistically significant for the analysis.
The box-plots for the standard deviation and for the peak value of the module of the acceleration are shown in Figure 7.10 a and b respectively.

(a) (b)

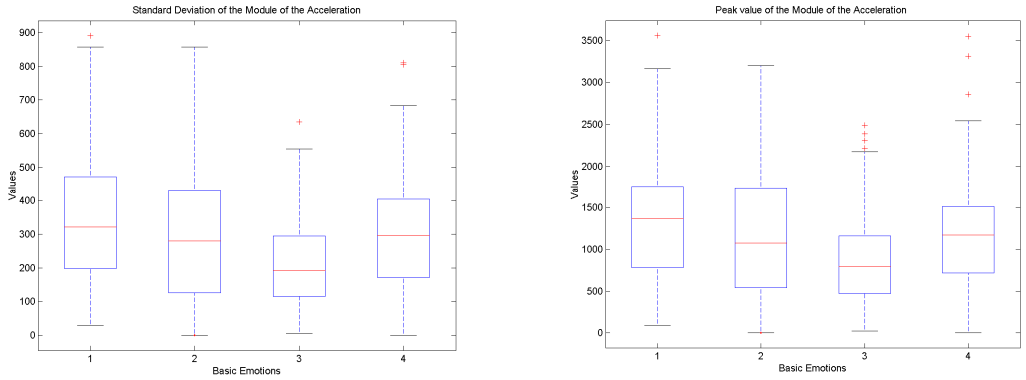


Figure 7.10: box-plots of (a) the standard deviation and (b) the peak value of the module of the acceleration along the motion phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy. Acceleration is measured in [pixels/s^2].

Significant differences can be noticed also in these box-plots, confirming for example that grief performances have the lowest values of acceleration (both in average and as peak value). Grief seems therefore to be characterized by small changes in velocity and therefore by movements sustained in Laban's Time. Conversely, anger seems to be confirmed as the basic emotion having the highest dynamics (i.e., impulsiveness).

7.1.4. Space-related expressive cues

Space has been taken into account by extracting four expressive cues related to it: length of motion trajectories, direction of motion trajectories, Directness Index, and amount of upward movement. These cues have been computed for each motion phase (e.g., the length of the trajectory followed by a point during the whole motion phase has been calculated) on the trajectories of the same 40 points whose velocity and acceleration have been considered above. An overall value for each cue has been obtained by averaging on the 40 trajectories.
Motion length has been normalised with respect to the duration of the motion phase in which it has been calculated. Motion direction has been obtained as the angle of the vector representing the overall motion direction (measured in radians in $[-\pi,\pi]$). Upward movement has been computed as the fraction of time in the motion phase motion direction was in $[0, \pi]$ (and therefore upward movement is in the range [0, 1], being 1 the condition in which motion direction was in $[0, \pi]$ along the whole motion phase).

The box-plots of the four space-related cues are displayed in Figure 7.11a, b, c, and d respectively.

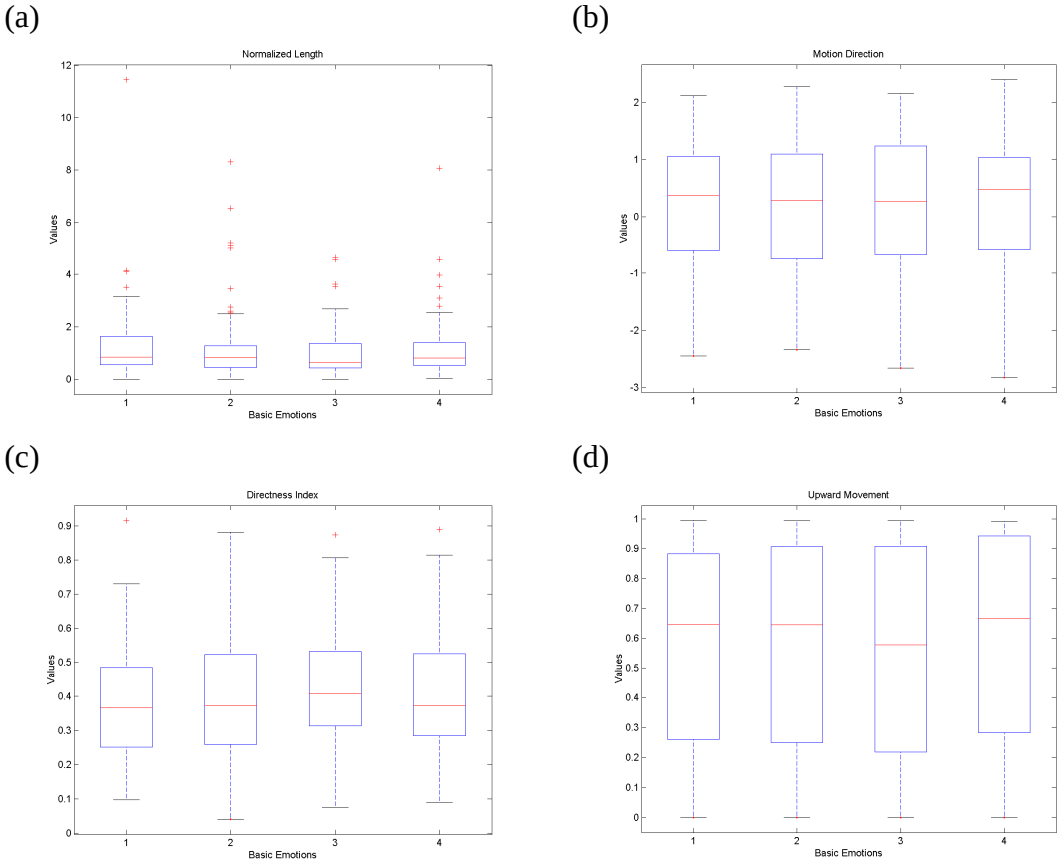


Figure 7.11: box-plots of (a) the normalised length of the overall motion trajectory along a motion phase (measured in [pixel/number of frames]), (b) the overall motion direction along a motion phase (measured in radians in the range [-π,π]), (c) the overall Directness Index along a motion phase, (d) the overall amount of upward movement in a motion phase. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy.

These cues do not seem to be very significant for the analysis. The box-plots show quite similar values of them with respect to the four basic emotions. A possible problem with these cues is related to the duration of the motion phases. If a motion phase has a quite long duration, many motion strokes with different direction and directness can be part of it: if from the one hand the single values of direction and directness can be meaningful for each stroke, on the other hand their average on the whole motion phase can lose its significance since the strokes have too different values. A deeper analysis would therefore be needed on these cues taking into account different time views as it has been illustrated in Chapter 5. It should be noticed however that, as predicted by Boone and Cunningham (1998) upward movement shows a slightly higher value for joy. Even if probably not relevant, these cues have been kept in the analysis since they represent an important aspect of motion analysis.

7.1.5. Time-related expressive cues

Two time durations have been considered in this analysis: the duration of the analysed motion phase and the duration of the immediately previous pause phase. The considered time durations are relative, i.e., divided by the duration of the whole performance. The box-plots for these cues are displayed in Figure 7.12 a and b respectively.

(a) (b)

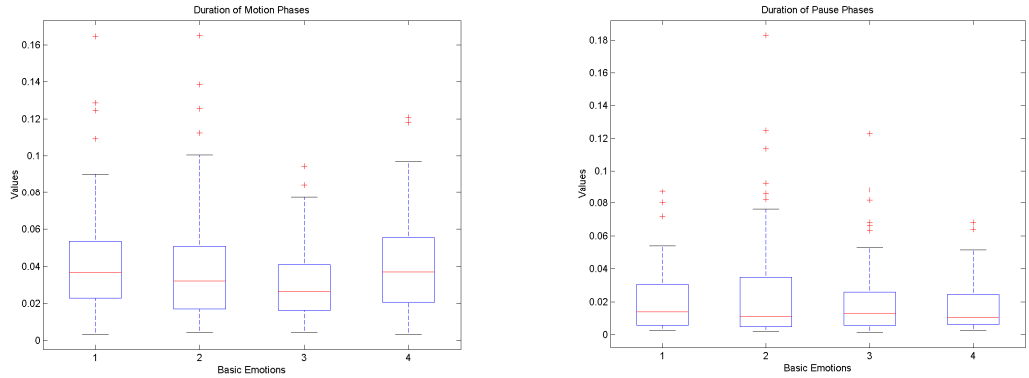


Figure 7.12: box-plots of (a) the relative duration of motion phases and (b) the relative duration of pause phases. The four basic emotions are labelled as follows: 1 – Anger, 2 – Fear, 3 – Grief, 4 – Joy.

The box-plots do not show very big differences even if grief seems to have shorter motion phases and longer pause phases.

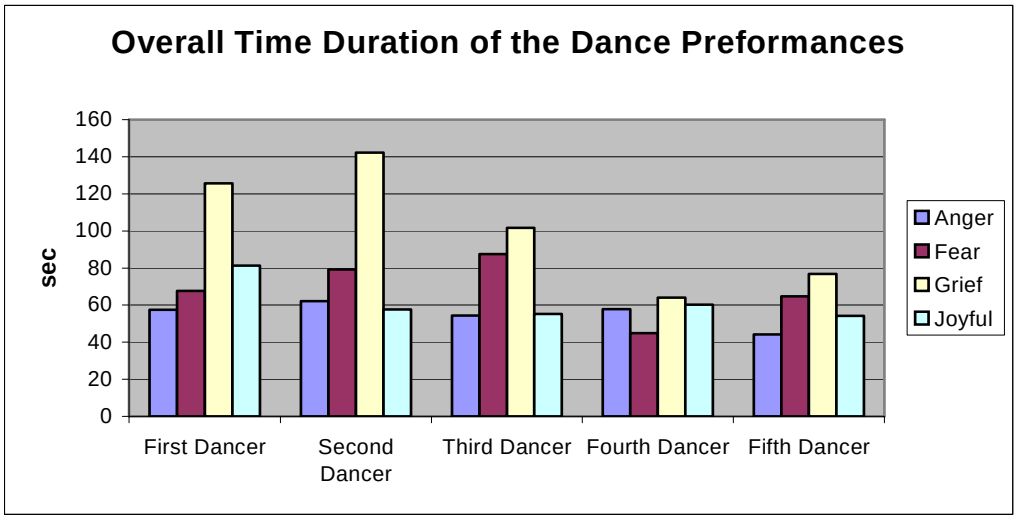


Figure 7.13 the overall duration of the dance performances for the 5 dancers (Mazzarino, 2002)

More interesting information can be obtained from the overall duration of the dances (see Figure 7.13 from Mazzarino, 2002). Psychologists argued that it should be longer for the grief performance and shorter for the anger one. This hypothesis seems to be confirmed by the data at least for grief, while for anger some discrepancies are indeed observed. It is however difficult to use the overall duration of dances for classifying single motion phases.
Another aspect concerns the number of motion phases, the sequence of pause and motion phases and the sequence of the values of the other cues (e.g., QoM and CI) along the motion phases of a dance performance. These aspects are for example related to fluency. As an example of some results in this direction (Camurri, Lagerlöf, and Volpe, 2003), Figure 7.14 and 7.15 show the average values computed for each motion phase of Quantity of Motion and Contraction Index respectively, vs. the index of the motion phases (i.e., as they appear along time).

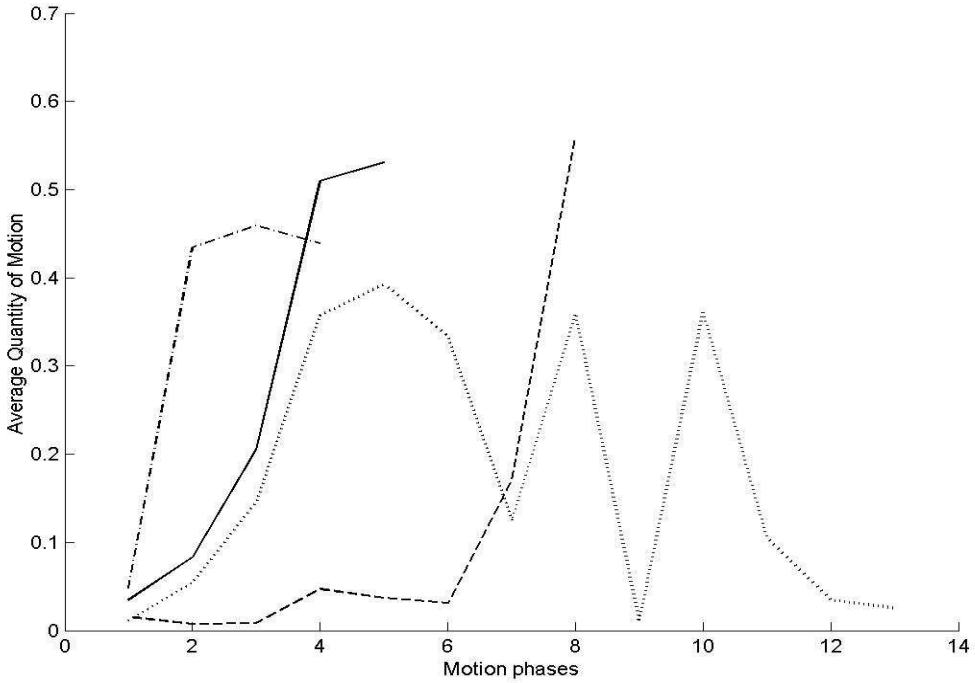


Figure 7.14: average values of the QoM computed for each motion phase (Camurri, Lagerlöf, and Volpe, 2003). The four graphs refer to four performances by the same dancer, each one expressing a different basic emotion: anger – solid line; fear – dashed line; joy – dash-dot line; grief – dotted line. The X axis is the index of the motion phase in which the movement has been segmented (therefore, X is not the time axis).

In each figure the four graphs refer to four performances by the same dancer in which the dancer tried to express the four basic emotions. In the figures line types are associated to emotions as follows: anger – solid line; fear – dashed line; joy – dash-dot line; grief – dotted line.
It can be noticed, for example, that curves representing the average Quantity of Motion for anger (solid line) and fear (dashed line) have a similar trend: i.e., they starts with low

values and slowly increase at the beginning, then they continuously increase with increasing steepness. Fear, however, have much more motion phases than anger indicating a less fluent motion.
Contraction Index for joy (dash-dot line) has quite low values with respect to the other emotions, while fear (dashed line) has quite high values, meaning that the body is often contracted (i.e., limbs are often close to the centre of gravity).
Grief (dotted line) always has a high number of motion phases and a high variance of the average values of Quantity of Motion, meaning frequent transitions between motion and pause phases and very low fluency. Joy (dash-dot line), instead, has few long motion phases indicating a very fluent motion.
It can be also noticed that, while from the one hand each of the four dancers has a particular trend allowing distinguishing between them, on the other hand what it has been observed above holds for all the four dancers, i.e., they expressed the four emotions by acting on the expressive cues in the same way.

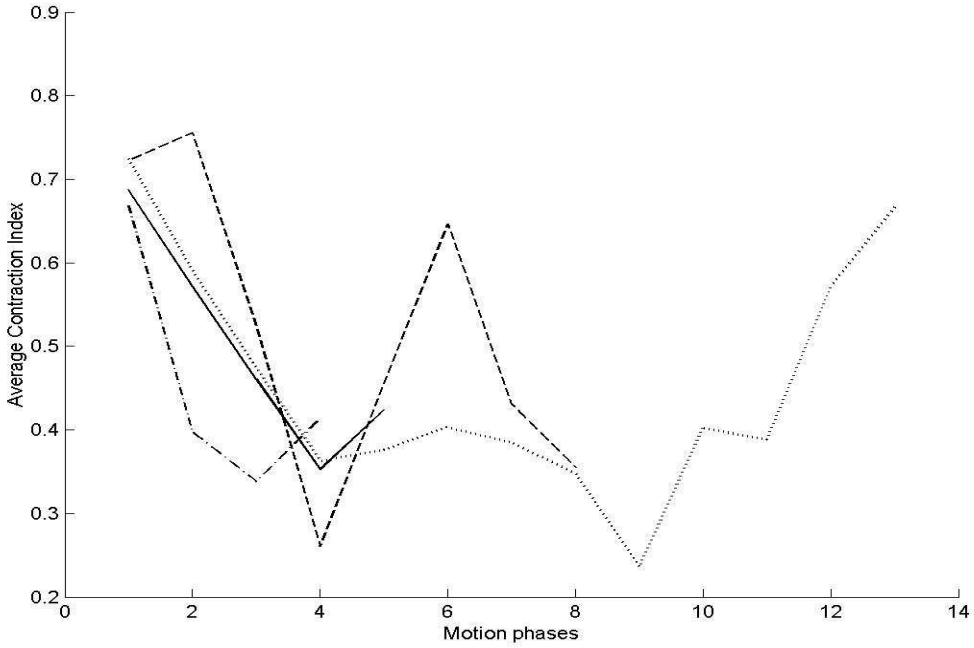


Figure 7.15: average values of the CI computed for each motion phase (Camurri, Lagerlöf, and Volpe, 2003). The four graphs refer to four performances by the same dancer, each one expressing a different basic emotion: anger – solid line; fear – dashed line; joy – dash-dot line; grief – dotted line. The X axis is the index of the motion phase in which the movement has been segmented (therefore, X is not the time axis).

In conclusion, the explorative analysis seems to confirm that at least some of the extracted cues are statistically relevant for the analysis and satisfy the research hypotheses outlined in Chapter 5. Other cues appear to be relevant even if no statistical evidence can be produced about them. Other seems less important for classification purposes. The next section reports an attempt to build a model for performing such an automatic classification.

7.2. Classification of motion phases using decision trees

Segmentation applied to the 20 dances recorded for the described experiment produced 334 motion phases, each one characterized by the 18 variables previously discussed. In order to avoid problems with the range of the variables, they have been standardised.
Decision tree models have been built in order to classify such motion phases in term of the four basic emotions anger, fear, grief, and joy. Five training sets (85% of the available data) and five test sets (15% of the available data) have been extracted from the data set. The samples for the test sets were uniformly distributed along the four classes and the five dancers. Five decision trees have been built on the five training sets and evaluated on the five test sets. The Gini's index of heterogeneity has been used for building the decision trees. Decision trees have been selected for this study since they produce rules that can be used to try to give an interpretation of the results. However, comparison with other classification techniques (e.g., Neural Networks, Support Vector Machines) would be useful and remain as task for possible future work.
The results obtained on the five decision trees are summarized in Tables 7.5 and 7.6 in the following pages (confusion matrices for the training sets and for the test sets respectively).
Two models (model 3 and model 5) fit quite well the data set (the rates of correct classification on the training set for these two models averaged on the four classes are respectively 78.5% and 61.6%). Models 1, 2, and 4 have difficulties in classifying fear (rates of correct classification on the training set for these three models averaged on the four classes are respectively 41.9%, 38.7%, and 36%). Models 2 and 4 have problems also with joy (i.e., they distinguish only between anger and grief).
A similar situation can be observed in the evaluation carried out on the test set: only models 3 and 5 are able to classify the four emotions. Model 1 cannot classify fear; models 2 and 4 cannot classify fear and joy.
The rates of correct classification on the test set for the five models averaged on the four classes are respectively: 40%, 36%, 36%, 26%, and 40%. Thus the average rate of correct classification on the five models is 35.6%. Except for model 4 they are all above chance level (25%). Model 5 can be considered as the best model since it has a rate of correct classification of 40% and is able to classify all the four emotions.
These rates of correct classification that at a first glance seem to be quite low (40% the best model) have however to be considered in relationship with the rates of correct classification from spectators who have been asked to classify the same dances. In fact, spectators' ratings collected by psychologists in Uppsala show a rate of correct classification (averaged on the 20 dances) of 56%.
The rate of correct recognition for automatic classification (35.6%) is thus in between chance level (25%) and the rate of correct recognition for human observers (56%).
Furthermore, if the rate of correct classification for human observers is considered as reference[2], and percentages are recalculated taking it as 100% (i.e., relative instead of absolute rates are computed), the average rate of correct automatic classification with

[2] At the current state of the art of technology in these fields, it is reasonable to consider that machines are still not able to overcome humans in tasks like classification of emotions.

respect to spectators is 63.6%, and the best model (i.e., model 5) obtain a rate of correct classification of 71.4%.

Model 1							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	64	68.75	31.25	44	0	6	14
Fear	60	0	100	30	0	16	14
Grief	86	48,8372	51,1628	17	0	42	27
Joy	74	50	50	18	0	19	37
Model 2							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	64	84.375	15.625	54	0	10	0
Fear	60	0	100	45	0	15	0
Grief	86	70,9302	29,0698	25	0	61	0
Joy	74	0	100	51	0	23	0
Model 3							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	64	79,6875	20,3125	51	4	6	3
Fear	60	71,6667	28,3333	6	43	7	4
Grief	86	81,3954	18,6047	4	0	70	12
Joy	74	81,0811	18,9189	6	5	3	60
Model 4							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	64	68.75	31.25	44	0	20	0
Fear	60	0	100	37	0	23	0
Grief	86	76,7442	23,2558	20	0	66	0
Joy	74	0	100	45	0	29	0
Model 5							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	64	71.875	28.125	46	10	2	6
Fear	60	61,6667	38,3333	15	37	1	7
Grief	86	47,6744	52,3256	10	19	41	16
Joy	74	64,8649	35,1351	13	8	5	48

Table 7.5: confusion matrices for the training set for the five decision trees

By observing the confusion matrix of the best model (both for the test set and for the training set) it can be noticed that fear is often classified as anger. This particularly holds for the test set, where fear is the basic emotion receiving the lowest rate of correct

classification since 6 of the 13 motion phases extracted from fear performances are classified as anger.

Model 1							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	12	50	50	6	0	2	4
Fear	13	0	100	6	0	3	4
Grief	12	66,6667	33,3333	3	0	8	1
Joy	13	46,1538	53,8462	5	0	2	6
Model 2							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	12	91,6667	83.333	11	0	1	0
Fear	13	0	100	7	0	6	0
Grief	12	58,3333	41,6667	5	0	7	0
Joy	13	0	100	9	0	4	0
Model 3							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	12	41,6667	58,3333	5	2	2	3
Fear	13	15,3846	84,6154	8	2	1	2
Grief	12	41,6667	58,3333	3	3	5	1
Joy	13	46,1538	53,8462	4	1	2	6
Model 4							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	12	75	25	9	0	3	0
Fear	13	0	100	5	0	8	0
Grief	12	33,3333	66,6667	8	0	4	0
Joy	13	0	100	7	0	6	0
Model 5							
Class	Total	%Correct	%Error	Anger	Fear	Grief	Joy
Anger	12	41,6667	58,3333	5	3	0	4
Fear	13	30,7692	69,2308	6	4	2	1
Grief	12	41,6667	58,3333	2	0	5	5
Joy	13	46,1538	53,8462	4	0	3	6

Table 7.6: confusion matrices for the test set for the five decision trees

Something similar can be observed also in spectators' ratings (Camurri, Lagerlöf, Volpe, 2003). A more detailed comparison between automatic classification and spectators' ratings for each performance (i.e., for each dancer and each basic emotion) can be found in Table 7.7 in the following page.

Automatic Classification					
	Anger	Fear	Grief	Joy	Total
Dancer 1	50	13.2	70	6.6	29.94
Dancer 2	73.2	10	40	20	39.96
Dancer 3	30	6.6	60	33.4	30
Dancer 4	67	0	33.2	30	36
Dancer 5	70	13.2	50	46.6	41.94
Total	60	9.153846	48.3	27.67692	35.568
Spectators' ratings					
	Anger	Fear	Grief	Joy	Total
Dancer 1	40	33	40	67	45
Dancer 2	93	40	56	81	67.5
Dancer 3	53	75	47	75	62.5
Dancer 4	73	67	31	76	61.75
Dancer 5	44	60	25	53	45.5
Total	60.6	55	39.8	70.4	56.45

Table 7.7: comparison between automatic classification and spectators' ratings.
The table of spectators' rating is taken from (Camurri, Lagerlöf, Volpe, 2003).

The numbers appearing in the table of automatic classification are the average rates of correct classification computed for each performance on the five decision trees.
While anger is generally well classified both by spectators and by the automatic system (60% for automatic recognition vs. 60.6% for spectators), as already noticed quite bad results are obtained for fear (below chance level for automatic classification).
The biggest overall difference between spectators and automatic classification can be observed for joy (70.4% for spectators vs. 27.7%, just above chance level, for automatic classification).
In the case of grief instead automatic classification performs better than human observers (48.3% for automatic classification vs. 39.8% for spectators): in the tables this happens in five cases and mainly for grief.
In seven cases the rate of correct classification for the automatic system is below chance level (and this always happens for fear).
In one case automatic classification did not succeed in finding the correct emotion (Fear – Dancer 4), but spectators obtained 67% of correct classification.
In one case spectators' ratings are below chance level (Grief – Dancer 5), but automatic classification could obtain a rate of correct classification up to 50%.
Dancer 1 obtained the lowest rates of correct classification both from spectators and from the models. Dancer 5 obtains similar rates from both. Dancer 2 is the best classified by spectators and also obtains a quite high rate (with respect to the other dancers) in automatic classification.

7.3. Discussion

From the results of automatic classification sketched in the previous section some issues emerge that are worth to be shortly discussed.
The reduced rate of correct classification for human observers (56%) could be partially due to facial expressions that have been removed for spectators' ratings. Moreover, it is possible that some dancer did not perform at his/her best (this aspect in reflected in the variance of the rates of correct classification obtained for each performance). Some concerns also arise with respect to the methodology and to the aims of the experiment. For example, is it possible to ask a dancer to dance trying to express fear? Fear is usually displayed without the intention of doing it (and more often it is dissimulated in order to avoid to give the impression of feeling fear). Further, how much is dance (or music) able (or is it intended) to communicate a specific emotion? While listening to music or watching a dance performance it is not obvious that a specific emotion is triggered and perceived by spectators. Therefore, if from the one hand, the experiment here described can be a good starting point for research on expressive gesture, on the other hand other kinds of experiments would probably be needed for investigating less specific aspects (e.g., engagement, arousal) that are likely to be a more common component of the emotional experience.
The gap among automatic classification and spectators' ratings could be due to the lack of cues related to the temporal aspects of movement (i.e., rhythm). Such aspects have been only marginally considered in this experiment: in fact, among the 18 variables used for classification only the duration of the motion phase and of the previous pause phase can be considered as (weakly) related to rhythmical aspects. Indexes of fluency and impulsiveness as those sketched in Chapter 6 should also be introduced. Analogies with music could be envisaged, e.g., cues related to articulation: depending on if and how much motion phases overlap each others something similar to music legato and staccato could be taken into account. Enhanced segmentation techniques can be applied in order to shed light on the internal and rhythmical structure of a motion phase (i.e., in analogy with music, its attack, sustain, and decay sub-phases).
By considering only single motion phases in the automatic classification, lot of contextual information has been removed (i.e., no information about previous motion phases is available to the automatic classification algorithm). Information about postures in the previous pause phases has also been removed. Such reduction of information about the context (related to the time perspective illustrated in Chapter 5) may also be responsible of the gap between spectators and automatic classification.
In fact, in comparing spectators' ratings and automatic classification, it should be noticed that spectators observed the whole dance before giving their judgment, while automatic classification is performed on single motion phases. That is, spectators received more information (i.e., overall duration of the dance, information related to the sequence of motion and pause phases, information related to body postures) with respect to the automatic system. Thus, if from the one hand, such a comparison can be useful for evaluating the performance of automatic classification and for indicating possible directions for future work (e.g., by trying to understand the differences among the dances

that received a high rate of correct classification and those that did not), on the other hand it could be misleading and it has to be considered with prudence.
With respect to the conceptual framework discussed in Chapter 3 the automatic classifier can be collocated at Layer 4. It is an example of the kind of techniques that can be employed at this level. As already stated, comparison with other possible techniques could be useful and should constitute a research direction for future work. The decision trees mainly take into account cues located at Layer 2 (and some of them at Layer 3). The classifier works on a vector of such cues. Alternatively, it could be useful to work on a Layer 3 representation of expressive gesture (e.g., in term of energy and rhythm or of the dimensions of Laban's Theory of Effort), for example trying to classify trajectories in expressive spaces. While in music examples of such expressive spaces are available (see for example Canazza et al. 2000), in movement further research is needed in order to ground possible spaces on solid scientific bases.
In conclusion, this experiment can be considered as a first step and a starting point toward understanding the mechanisms of expressive gesture communication in dance. A collection of cues having some influence in such a communication process has been individuated, measured, and studied. A first attempt of automatic classification of motion phases has also been carried out and some results obtained (e.g., an average rate of correct classification not particularly high, but however well above chance level). Some directions for future research also emerged.
A final remark (that will be reconsidered in the conclusion of this dissertation) concerns possible exploitations in concrete applications of the obtained results. It should be remembered that if the scientific focus was on the communicative mechanisms of expressive gesture, from a technical/applicative perspective the goal was to develop techniques enabling the development of novel interaction paradigms for interactive multimedia (especially for artistic performances). From the applicative point of view, what has been presented here can be considered as a first concrete implementation of the conceptual framework described in Chapter 3, although limited to the aspects of analysis of expressive gesture in human full-body movement. These techniques can thus be employed in developing the Expressive Gesture Analysis component of a virtual or mixed subject inhabiting a Multilayered Integrated Expressive Environment. Such a subject would be able to observe movement, measure expressive cues, extract expressive gestures, and possibly classify them according to the conveyed expressive content. If from the one hand, the way toward a subject fully having such skills is still very long and I do not know if it will be ever possible to obtain such a subject[3], on the other hand from this first attempt it is already possible to get some information (e.g., the values of the measured expressive cues) on the expressive gesture the dancer is performing and such information can already be used in design and implementation of interaction mechanisms. In particular, in the field of performing arts the information that the described algorithms already make available can provide artists and designers of interactive systems with a collection of conceptual as well as technical tools enabling them to work in a scenario that technology only makes possible.

[3] Indeed I also don't know if I really would like to obtain such a fully skilled subject able to automatically classify people according to their emotional state…

8. Analysis in the General Space[1]

As already described in Chapter 5, the concepts of "Kinesphere", referred also as "Personal Space", and of "General Space", the whole space surrounding the Kinesphere, come from theories of the researcher and choreographer Rudolf Laban. Personal Space and General Space constitute two different space perspectives along which expressive gesture in human movement can be analysed.
The models and techniques discussed in the previous Chapters mainly dealt with movement in the Personal Space, i.e., they considered overall descriptors of the body movements of one dancer. The analysis of movement in the General Space instead consists in analysing the movement of one or more dancers (i.e., of his/her/their Kinesphere) in the surrounding space. This Chapter introduces some main research issues for analysis in the General Space and discusses a model for it. The model has been implemented as a collection of software modules for the EyesWeb Expressive Gesture Processing Library.

8.1. Research issues

Analysis of movement in the General Space is here addressed with respect to four main research issues. They can be shortly summarised as follows:

(i) *Use of the space.* The objective is to study how a dancer[2] uses the space surrounding him/her and the relationships between use of space and communicated expressive content. The focus is on individuating trajectories in the space and classifying them. Typical and repetitive patterns can also be extracted and further analysed. A set of parameters can be calculated such as the classified shape of a trajectory, the level of utilization of regions on the stage (e.g., occupation rates), the periodicity of repetitive patterns. Notice that at this stage of the analysis the space is considered as "neutral" i.e., without scenery or particular lighting (or, at least, scenery and lighting are excluded from the analysis).

(ii) *Relationship with elements such as lights and scenery.* The expressive content conveyed by the movement of a dancer in the space can widely change depending on elements giving a particular meaning to regions in the space. For instance, if the dancer moves continuously near a wall, the expressive content he/she conveys is very different with respect to a situation in which the dancer stands or moves directly in front of the audience. Mechanisms to associate an expressive potential to

[1] This Chapter is partially taken from Camurri A., Mazzarino B., Trocca R., Volpe G. "Real-Time Analysis of Expressive Cues in Human Movement", in Proc. Cast01 - Conference on artistic, cultural and scientific aspects of experimental media spaces, pp. 63-68, Bonn, September 2001.

[2] Notice that if from the one hand the reference model is here defined with reference to a dance performance, on the other hand it could be applied to the more general case of an object moving in a given space. As already mentioned, dance has been chosen as a test-bed in this work since it is the artistic expression of movement, therefore emphasizing the role of expressive gesture.

regions in the space can thus be developed and trajectories in such expressive maps can be studied. The parameters extracted in conditions of "neutral space" as described in (i) can thus be reconsidered in relationship with the expressive regions of the space (e.g., trajectories repetitively passing through a region with a high expressive potential can assume a particular relevance).

(iii) *Relationship between the movements of two or more dancers.* In the more general situation in which two or more dancers are involved in a performance, their movements in the General Space can be compared. The analysis on more dancers can be carried out both with respect to a "neutral" space and with respect to a space having expressive potentials. The relationships between the movements of each single dancer and the movement of the group can be also investigated.

(iv) *Relationship between parameters related to the General Space and parameters related to the Kinesphere.* The techniques developed for analysis in the General Space are quite general: they can be applied to the analysis of movement in the Kinesphere as well. For example, analysis of trajectories, levels of utilization of particular regions, detection of repetitive patterns can be applied also to the motion trajectories of limbs inside the Kinesphere. Conversely, some parameters that are calculated mainly with respect to the Kinesphere can be reconsidered from the point of view of the General Space (e.g., "equilibrium" with respect to the expressive potentials, ratio between rotational movements and straight movements, use of straight and direct trajectories with respect to smooth trajectories).

8.2. Reference model

The main contribution to analysis in the General Space discussed in this dissertation is the development of a reference model that can be used as a basis for such analysis. The model improves an older model coming from previous studies carried out at the DIST – InfoMus Lab in collaboration with Waseda University, Tokyo (Camurri, Hashimoto, Suzuki, and Trocca, 1998). Extraction and analysis of parameters from this model is still an ongoing work.

In the model, the General Space (considered as a rectangle) is divided into active cells forming a grid (see Figure 8.1). Each time the position of a tracked dancer is detected[3], the corresponding cell is individuated and its indexes *h* and *k* are returned. Discrete potential functions can then be defined on the General Space. A discrete potential function can be represented by a matrix $\Phi = [\phi_{ij}]$. The items in the matrix directly correspond to the cells in the grid: ϕ_{ij} is the value that the potential function assumes in correspondence with the cell having (*i*, *j*) has indexes. Three main kinds of potential functions are envisaged:

[3] Here I do not face the problem of tracking the position of the dancer in the General Space. Depending on the conditions of the stage (e.g., lighting, number of dancers on stage etc.), the solution could be difficult to find and implement. Several techniques can be employed, ranging from computer vision techniques to special purpose hardware localization systems.

(i) Potential functions *not* depending on the current position of the tracked dancer.
(ii) Potential functions depending on the current position of the tracked dancer.
(iii) Potential functions depending on the definition of regions inside the General Space.

Suitable mapping strategies (see Chapter 4) can be developed in order to associate some behaviour to a particular cell or set of cells (e.g., a direct mapping could generate some kind of output when a particular cell is activated by the dancer passing on it).

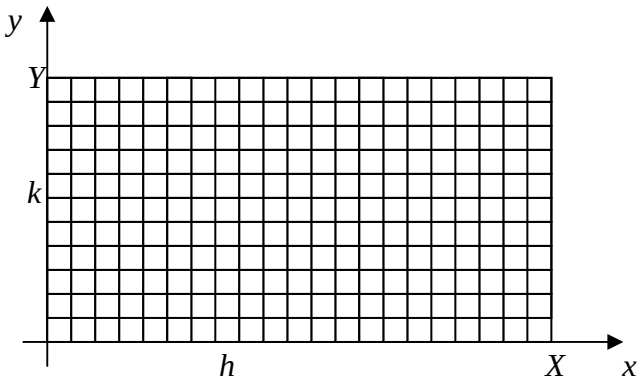


Figure 8.1: the General Space considered as a grid of active cells

8.2.1. Potential functions not depending on the current position of the dancer

ϕ_{ij} is constant with respect to the cell currently activated by the dancer. Consider, for example, lights and fixed scenery: potential functions can be associated to each element of fixed scenery and to the lights that are used in a particular scene. The potential function associated to each element can be represented by a matrix Φ_p. The overall effect can be determined by summing the matrixes Φ_p in an overall matrix $\Phi = \Phi_1 + \Phi_2 + \ ... + \Phi_P$, being P the number of scenery and light elements taken into account. The trajectory of the dancer with respect to such potential function can be studied in order to identify relationships between movement and scenery and suitable mapping strategies can be developed in order to associate outputs to movements performed in relevant places. Nevertheless, the current cell *(h, k)* in which the position of the dancer is mapped has no influence on the elements ϕ_{ij} of the matrix representing the potential function: the values of such elements ϕ_{ij} are in fact calculated only on the basis of the positions of the considered fixed scenery and lights. Note that this does not mean that ϕ_{ij} has to be constant along time: consider, for example, lights that are turned on and off during the performance. Their contribution to the overall potential function can be added only when they are on. In this case, the values ϕ_{ij} change over time, nevertheless the potential function is still independent from the current position of the dancer.

8.2.2. Potential functions depending on the current position of the dancer

$\phi_{ij} = \phi_{ij}(h, k)$ where (h, k) is the cell currently occupied by the dancer. In this way it is possible to define potential functions moving in the space together with the movement of the tracked dancer. Consider, for example, the following potential function:

$$\phi_{ij}(h,k) = \begin{cases} \dfrac{1}{|(i-h)|+|(j-k)|} & \text{if } (i,j) \neq (h,k) \\ 1 & \text{if } (i,j) = (h,k) \end{cases} \qquad (*)$$

The potential function depends on the current cell position *(h, k)* of the dancer and changes every time he/she moves. For example, Figure 8.2 shows the potential function (*) calculated when the dancer occupies respectively the cells (10,10), (40,40) and (60,60) (i.e., the dancer is moving along a diagonal) in a space having 100×100 cells.

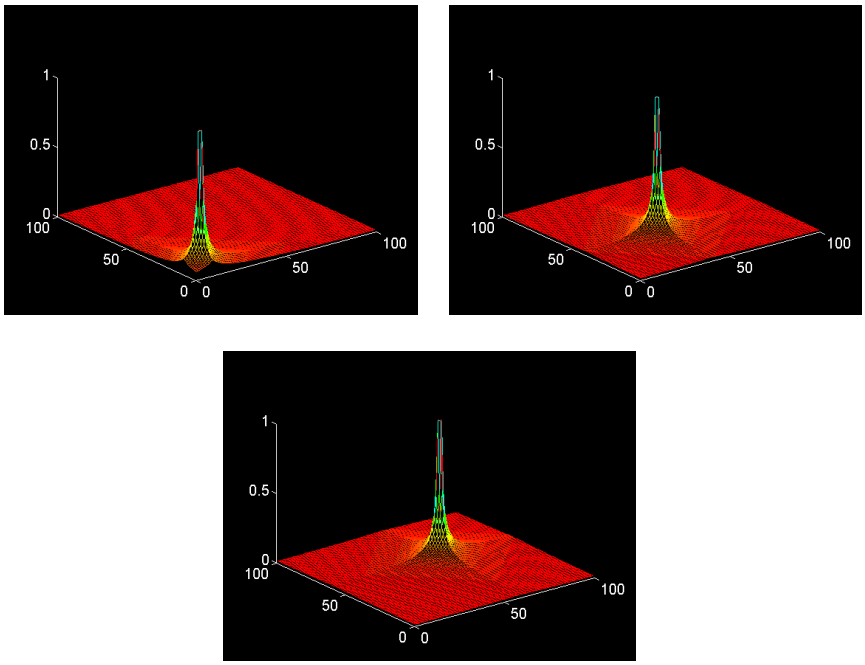


Figure 8.2: the potential function (*) calculated for a dancer moving along a diagonal, i.e., cells (10,10), (40,40) and (60,60) in a space having 100x100 cells

In a more general perspective, it is possible to create potential functions forming a "bell" around the dancer and moving with him: further the "bell" can be modified (e.g., made wider) corresponding to the analysed movement inside the Kinesphere (e.g., a wide "bell" associated to expansive movements of the Kinesphere). Mobile scenery can also be associated to this kind of potential functions.
Another example: suppose that each cell is characterised by an activity level, i.e. a sort of measure of how much the cell has been occupied by the dancer. The activity levels of the cells are stored in an m×n matrix A = $[a_{ij}]$ where *i* and *j* are the indexes associated to a

cell. The matrix A defines a potential function in the General Space. Consider a neutral environment. An increment function I(a_{ij}) and a decrement function D(a_{ij}) are defined. Since, at a first stage, the space is considered as neutral, i.e., no particular meaning is associated to regions in the space, the same increment and decrement functions are applied to all the cells in the grid. Each time the position of the tracked dancer is detected, the corresponding cell is individuated and its activity level is incremented according to the increment function. The activity value of the remaining cells is decremented according to the decrement function. This example can be implemented using a potential function depending on the current position of the tracked dancer defined as:

$$a_{ij}(h,k) = \begin{cases} D(a_{ij}) & \text{if } (i,j) \neq (h,k) \\ I(a_{ij}) & \text{if } (i,j) = (h,k) \end{cases}$$

8.2.3 Potential functions depending on the definition of regions

Regions can be defined on the grid. An hit function H(·) and a miss function M(·) can be associated to each region. The hit function is applied to calculate the potential function for a cell inside a region, each time the cell currently occupied by the dancer is inside that region. Otherwise, the miss function is used. $\phi_{ij} = \phi_{ij}$ (R(i,j)) where R(i,j) is the region to which the cell (i, j) belongs. In particular if N regions R_1, R_2, ... , R_N are defined with the correspondent H_1, H_2, ... , H_N hit functions and M_1, M_2, ... , M_N miss functions,

$$\phi_{ij}(R(i,j)) = \begin{cases} H_p(\phi_{ij}) & \text{if } R(i,j) = R_p = R(h,k) \\ M_p(\phi_{ij}) & \text{if } R(i,j) = R_p \neq R(h,k) \end{cases}$$

Note that, since the hit and miss functions are here defined as functions of the previous value of the potential function in the cell (*i, j*), some kind of memory is involved in this approach.

The previous example concerning the activity level of a cell in a neutral environment can be also implemented by using a potential function depending on the definition of regions in the General Space: in particular, in that case each cell defines a region (i.e., m×n regions are defined) and the same hit function H(·) = I(a_{ij}) and miss function M(·) = D(a_{ij}) are associated to all the regions (cells). Suppose now to consider a stage environment with presence of scenery and lights. The "neutral" values of the activity level of each cell previously calculated are no more valid: there will be some regions in the General Space in which the presence of movement is more meaningful than in others. A certain number of "meaningful" regions (i.e., regions on which a particular focus is placed) can be defined and suitable hit and miss functions can be associated to them. A variation related to the meaning of a specified region is added to the "neutral" evaluation of the activity level, thus obtaining a new activity level taking into account elements of a particular stage environment.

8.3. The EyesWeb Space Analysis Library

The model previously described has been implemented as a collection of software modules for the EyesWeb open software platform (see Camurri, Coletta, Peri, Ricchetti, Ricci, Trocca, Volpe, 2000, and Appendix A), included in the EyesWeb Space Analysis Library. The EyesWeb Space Analysis Library is part of the EyesWeb Expressive Gesture Processing Library (see Camurri, Mazzarino, Volpe, 2003, and Appendix B).
The General Space model consists of four EyesWeb blocks: the first one allows subdivision of an image in cells and returns as output the indexes of the cell currently occupied by a given tracked point (e.g., a dancer moving on stage), the other three allow the definition of (i) potential functions independent from the position of a tracked object, (ii) potential functions depending on the current position of a tracked object, (iii) potential functions depending on the definition of regions in the space.

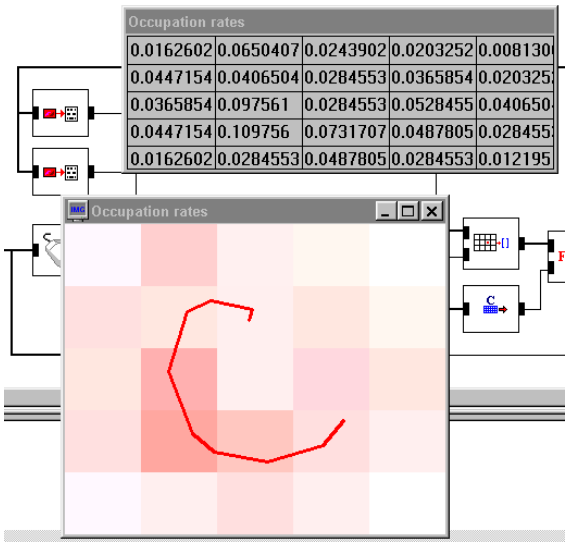


Figure 8.3: occupation rates of cells in the General Space

Motion features can also be extracted from analysis in the General Space. For example, Figure 8.3 shows the occupation rates calculated on a rectangular space divided into 25 cells. After sampling the trajectory followed by a tracked point, the occupation rate is calculated for a given cell as the ratio between the number of samples the tracked point was inside that cell and the total number of available samples. In the Figure the intensity (saturation) of the colour for each cell is directly proportional to the occupation rate of the cell. The trajectory of the tracked point is also displayed.

9. Conclusions

This dissertation introduced Multilayered Integrated Expressive Environments (MIEEs) as environments in which novel forms of artistic performances can be developed, where the performing action takes place in a number of physical as well as virtual connected spaces, inhabited by real, virtual, and mixed subjects. MIEEs have been discussed with respect to their structure and the communication processes taking places among the subjects inhabiting them. A particular focus has been put on expressive gesture as a main carrier of information in MIEEs in which, as it often happens in artistic performances, communication mostly exploits non-verbal mechanisms. An architecture for virtual and mixed subjects inhabiting MIEEs has been presented, and a conceptual framework for multimodal analysis and synthesis of expressive gesture by such subjects has been discussed.
Attention then moved on analysis of expressive gesture in human full-body movement considered as an example of processing a virtual or mixed subject has to carry out in order to accomplish its task inside a MIEE. An experiment has been present analysing expressive gesture in dance performance with respect to the emotions it is able to convey. After reviewing the sources on which research has been grounded, a collection of techniques for real-time extraction of expressive cues from video-captured human full-body movements has been presented. A prototype of decision tree classifier of expressive gestures in term of the four basic emotions anger, fear, grief, and joy has been developed and its outputs have been described.
In conclusion, before shortly discussing obtained results, future works, and possible ethical concerns of this research, two concrete sample applications are presented in this Chapter, related to two different application scenarios: artistic performances and therapy and rehabilitation. These are just examples of the wide possibilities of exploitation of the developed models and techniques in a broad set of application fields, such as for example interactive edutainment, interactive entertainment, applications for culture, museums, and exhibits, tools for performing arts, for the industry of digital music instruments, for music theatre, for therapy and rehabilitation.

9.1. Two sample applications

Two examples of concrete applications exploiting the developed models and techniques are now briefly introduced, the first one in the field of performing arts, the second in therapy and rehabilitation. The two applications have been developed at the DIST – InfoMus Lab in the framework of two EU projects: the cited EU IST project MEGA and the EU-IST project CARE HERE (Creating Aesthetically Resonant Environments for the Handicapped, Elderly, and Rehabilitation).

9.1.1. The concert "Allegoria dell'opinione verbale"

This piece was conceived by the composer Roberto Doati during a workshop at the DIST - InfoMus Lab in June 2000 and performed (first performance) in September 2001 at the opening concert of the season of Teatro La Fenice, Venice, Italy. The concert has been performed again in March 2002 at Auditorium "E. Montale", Teatro dell'Opera Carlo Felice, Genova, Italy. During the concert an actress (Francesca Faiella) is on stage, seats on a stool placed in the front of the stage near the left side. The actress is turned towards the left backstage (the audience therefore sees her profile). A large screen projects her face in frontal view. A videocamera is placed (hidden) in the left part of the backstage, and it is used both to get images of the face of the actress to be projected on the large screen and to acquire her lips and face movements.

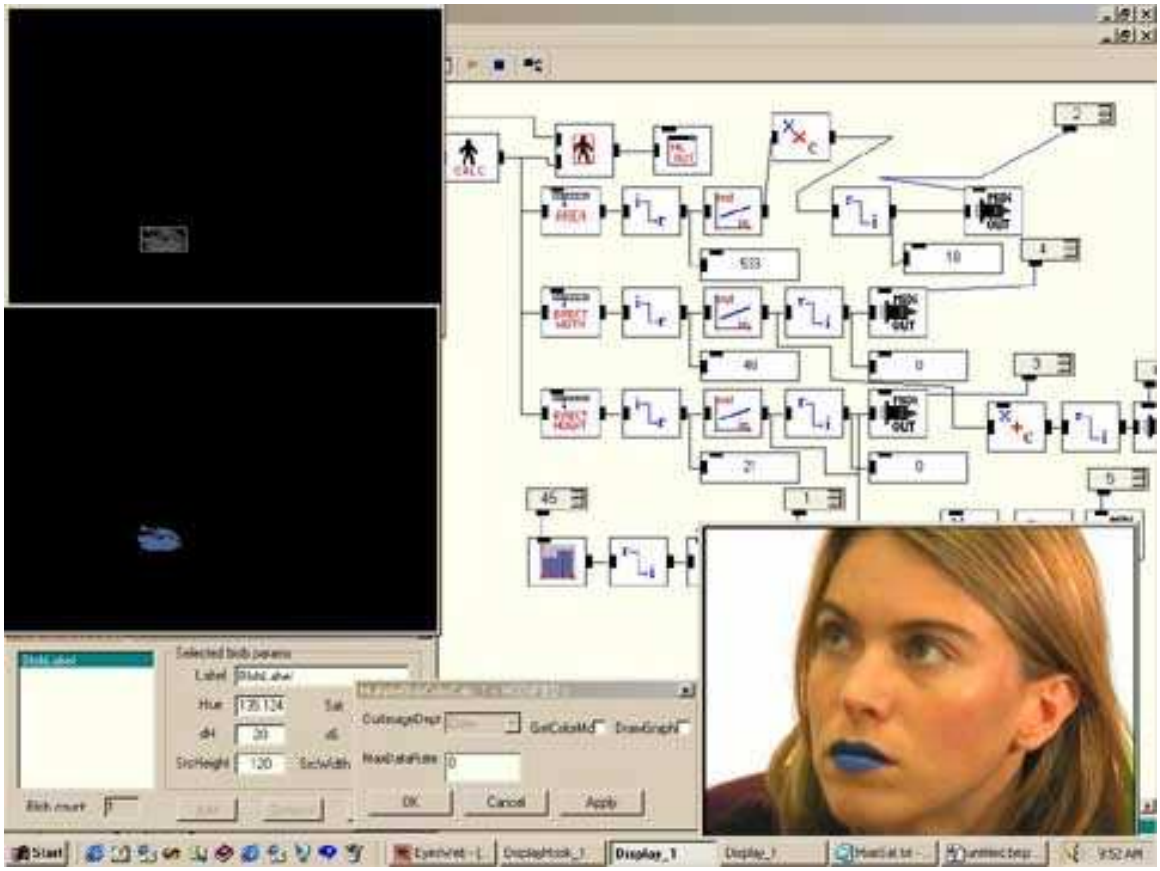


Figure 9.1: the EyesWeb patch for the concert "Allegoria dell'opinione verbale"

The actress plays the text in front of the camera. The EyesWeb open platform and the EyesWeb Expressive Gesture Processing Library are employed to process the movements of actress' lips and face, in order to extract expressive cues (similar to the ones described in Chapter 6) used to record and process in real-time her voice and diffuse spatialised electroacoustic music on four loudspeakers placed at the four corners of the auditorium in a standard electroacoustic music set-up. The signals reproduced by the

loudspeakers are only derived by the actress' voice: former recordings of her voice, real-time recordings, and post-processing in real-time. The audience can observe the movements of the actress' face in the large screen, while listening to the piece and thus perceiving the overlapping and interaction of her movements with sound changes coming from the loudspeakers. Figure 9.1 in the previous page shows the EyesWeb patch employed in the concert.

During the performance held in Genova in March 2002, an experiment was carried out (in collaboration with the Department of Psychology of the University of Uppsala, Sweden) in order to measure and evaluate the reactions of the audience to a concert exploiting interactive technologies. The event was structured as follows:

1) Performance of the piece.
2) Soon after the performance, distribution to the audience of a questionnaire prepared by the psychologists in Uppsala (no explanation at the entrance, only at this point). No introductory words apart from the kind request to fill the questionnaire.
3) Discussion, presentation, explanation by the composer, the actress, and prof. Antonio Camurri of both the aesthetic/artistic and technological issues, including a short live demonstration of how the system works by showing it on the big screen.
4) A second performance of the piece.
5) The audience answers to a second questionnaire.
6) End of the event.

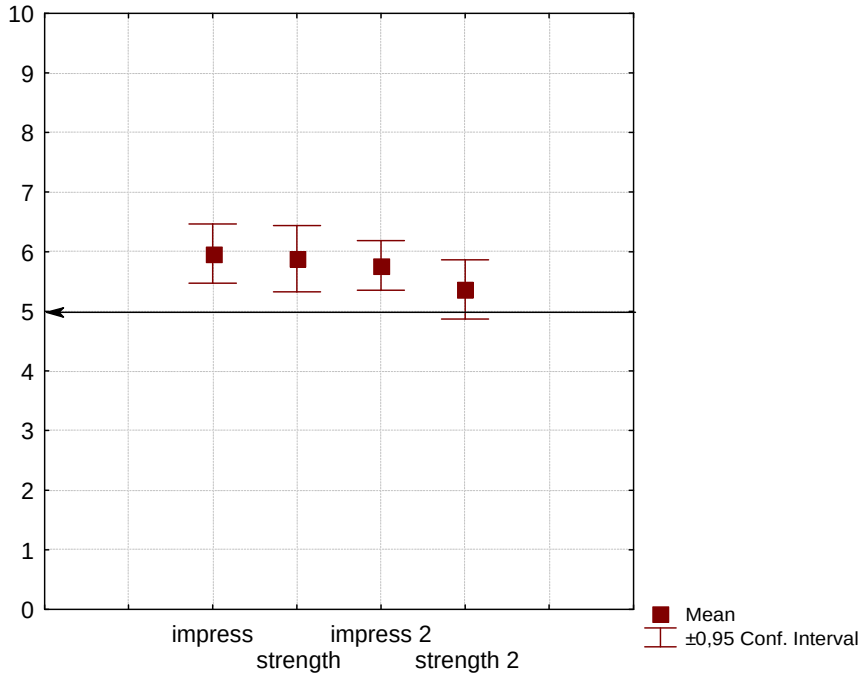


Figure 9.2: mean and 95% confidence interval for the audience's perceived first impression of the performance (first bar, from left), strength of experience (second bar), and to what extent the impression (third bar) and strength (fourth bar) have been changed by the second performance after discussion with the staff.

The audience included 60 people, with a wide variety concerning age, sex and culture. Figures 9.2 shows the responses of the audience respectively after the first and the second performance and with respect to two aspects: the overall impression (positive or negative) and the strength of the experience.
After the first performance the audience immediate impressions were positive (Mean 6.0, on a scale ranging from *very negative* 0 to *very positive* 10, see first bar from the left in Figure 9.2). The second bar in Figure 9.2 shows the perceived strength (mean 5.9) indicating a common strong experience of the performance among people in the audience. After the second performance (and discussion with the staff) the impressions were significantly more positive (mean 5.8) with respect to the audience's first impressions (on a scale ranging from *more negative* 0 to *more positive* 10, third bar in Figure 9.2). The strength of the experience was also affected (mean 5.4, last bar) but not significantly above scale-level 5, which for the two rightmost bars defines no change with respect to the former performance.

9.1.2. Therapy and rehabilitation of Parkinson's patients

The research outputs described in this dissertation have also been used in the framework of the EU-IST project CARE HERE to analyse the body movements of different kinds of patients (Parkinson's patients, severely handicapped children, people with disabilities in the learning processes) and to map the analysed parameters onto automatic real-time generation of visual outputs, attempting to create aesthetic resonance.
The underlying idea of aesthetic resonance is to give patients a visual and acoustic feedback depending on a qualitative analysis of their (full-body) movement, in order to evoke ludic aspects (and consequently introduce emotional-motivational elements) without the need neither of the rigid standardisation required for typical motion analysis, nor of invasive techniques: subjects can freely move without on body sensors/markers.
A pilot experiment carried out in order to test the developed techniques on patients with Parkinson's disease (PD) is described in (Camurri, Mazzarino, Volpe, Morasso, Priano, Re, 2003). The experiment consisted in analysing movement of two PD patients, extracting a collection of motion parameters related to motion energy and fluency, and producing in real-time audio and visual feedback.
For example, Figure 9.3 shows the output of a therapeutic session where patients are engaged in "interactive painting" with their own body. The patient sees himself on a large screen painting in real-time through his/her motion in the space. Previous work in the performing arts field exists where engagement of the audience is obtained in a similar way: see for example the PAGe - Painting by Aerial Gesture system (Tarabella, 2001). With PAGe the user can interact through an interaction paradigm like the MS Paint software, using his hands while standing in front of a large video screen: the user can select a colour or an action with one hand, then he can paint with that colour with the other hand. This therapeutic exercise is slightly different: the interaction is based on some of the movement cues described in Chapter 6. For example, the colour may depend on fluency; Quantity of Motion may be associated to the intensity of the colour trace; pauses in movement (using the segmentation techniques previously described) allow restarting the process and re-assigning/adapting the mappings strategies. In this way, by a

careful choice of colours, e.g., by creating “pleasant” colour associations/mappings with fluent and non-hesitating movements, it is possible to create a sort of visual feedback encouraging improvement of movement in patients. During this exercise the subject looks at the picture painted on the screen and continuously changes it while moving. On another display the researcher analyses the parameters and if needed corrects them in order to tune the exercise on the patient’s needs.

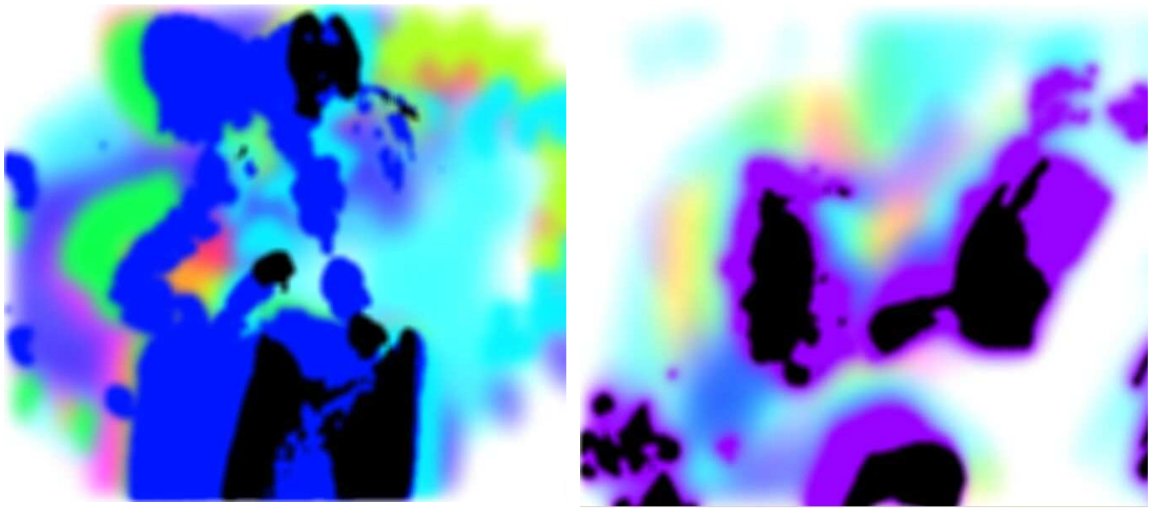


Figure 9.3: example of therapeutic session using a painting exercise. Colour and intensity of the colour trace depend on automatically extracted movement features.

9.2. Research results and perspectives

The work presented in this dissertation can be considered as belonging to a collection of first attempts of understanding the mechanisms underlying non-verbal communication through expressive gesture. Research in such direction is becoming more and more important as demonstrated by the continuous growing of the Affective Computing and KANSEI Information Processing fields in the U.S. and in Japan and by EU-funded projects like the MEGA project.
This work contributes in the development of applications for interactive multimedia scenarios in which such non-verbal mechanisms are the main communication channels. In this perspective, a particular focus has been on performing arts, even if others application domains are also envisaged (e.g., therapy and rehabilitation).
Relevant outputs can be considered the contributions in:

- The definition of a structure for inhabited multilayered environments able to provide a paradigm for the development of distributed interactive performances and giving a framework in which expressive gesture can be studied as the main carrier of non-verbal information among the inhabiting real, virtual, and mixed subject.
- A better (even if not exhaustive) definition of expressive gesture taking into account some of the existing literature on gesture modelling and processing and

encompassing different aspects of expressive gesture i.e., its role in communication, its aesthetical valence (as a main component of artistic performances), its relation with physicality (e.g., expressive gestures of dancers and music performers with respect to computer animated expressive gestures).

- The definition of an architecture for virtual and mixed subjects inhabiting Multilayered Integrated Expressive Environments, that starting from previous works on emotional agents for Multimodal Environments and grounding on the research carried out with partners in the MEGA project provides a way to include in the same framework analysis, synthesis, and mapping of expressive gesture in a multimodal perspective.
- The development of an instance of the analysis side of such architecture in the context of dance performance and as a result of an experiment aiming at understanding emotion communication through expressive gesture.
- The development of algorithms for measuring global expressive cues from human full-body movement, the analysis of such cues, and their use for a first attempt of automatic classification of dance fragments in term of conveyed basic emotions.

Of course, this work cannot be considered exhaustive and conclusive since many research issues still need to be discussed and wide perspectives are open for possible future work. As an example, I just try to list some of them:

- The dissertation focused on the structure of MIEEs and on the communication processes taking place inside them through expressive gesture. Another very relevant aspect has been only marginally faced: the dynamics of MIEEs along time. MIEEs are not static constructions, but they continuously evolve along time depending on the storyboard of the performance for which they are employed. A definition of storyboard would therefore be needed and the mechanisms through which the structure of MIEEs evolves would need a deeper investigation.
- The implementation of a MIEE will employ some kind of hardware and software platform. The discussed structure of a MIEE, its dynamics, the interactions taking place in it would therefore produce requirements for the employed platform. These requirements would need to be stated and analysed.
- The definition of expressive gesture should be further worked out. A comparison (and maybe a distinction) with respect to music gesture and dance gesture would be useful. A unified taxonomy of expressive cues characterising expressing gesture in movement (e.g., dance), music, and visual media should be developed, grounded on analogies between similar aspects in the different modalities.
- Other (and maybe more significant) aspects of expressive gesture need to be investigated. As discussed in Chapter 7, emotions might be too specific (i.e., watching a dance performance and listening to music do not always trigger specific emotions). In this perspective other experiments have been carried out and are still ongoing aiming at studying expressive gesture as responsible of emotional engagement in the audience. An example is the experiment sketched in Chapter 3 on the engagement of spectators exposed to musical stimuli. In this experiment recordings of piano performances (a piece by Skriabin and a piece by Liszt played by professional concert pianist Massimiliano Damerini) in different situations (i.e., in a studio situation without audience and in a performance-like situation with the

audience, both as a the first piece of the concert or after a virtuosistic piece) have been collected. Recordings consist of audio (four microphones, two near the piano and two in ambience) and video (four videocameras: from the front, from the top, from the left, and from the right with respect to the pianist, see Figure 9.4). Furthermore, MIDI data from the piano (a Yamaha Grand Coda Disklavier, rented for the experiment) have been collected. Spectators have been asked to evaluate their emotional engagement while listening to the performances by using techniques for continuous measurements (Schubert, 2001). Such continuous measures from spectators have been compared with extracted motion and audio cues in order to find possible correlations. Preliminary results can be found in (Camurri, Mazzarino, Timmers, Volpe, 2003, and Timmers, Camurri, Volpe, 2003).

- Some aspects in dance performance have only been marginally considered. In particular, aspects related to rhythm should be further investigated. Expressive cues like impulsiveness and fluency should be further worked out. Moreover, perceptual experiments would be needed to empirically validate the extracted expressive cues.
- Multimodal integration should be deeper investigated, i.e., analysis on particular aspects (e.g., expressive gesture in dance and in music) should be better related to the unifying conceptual framework described in Chapter 3. Work on synthesis and mapping strategies is still needed.

Figure 9.4: recordings of a piano performance for analysing spectators' engagement.

As the above incomplete list of possible extensions shows, lot of work is still needed for really understanding the power of expressive gesture in human-computer interaction and for developing systems able to fully exploit it.
Expressive gestures can open a path toward novel forms of artistic performances, in which technology is not just something added to a traditional scenario, but rather becomes a component of the artistic language. They are also a challenge for designer of interactive systems: as in software engineering methods for designing and implementing good software are developed and studied, the designer of interactive systems would need methods to develop and adapt his/her work with respect to the application scenarios and the requirements of the designer of a performance or of an installation. Moreover, analysis of some aspects of expressive gesture can lead to results that might be useful for other research domains. For example, the analysis of engagement in spectators exposed to musical stimuli, or the analysis of the behaviour of visitors in a museum exhibit can lead to the development of models of spectators/visitors and, in more general terms to models of users' behaviour, taking into account information related to the affective, emotional sphere.
Of course, the broad possibilities of industrial exploitation of such techniques also raise ethical concerns. As an example, let's consider the risk related to the availability of techniques able to emotionally classify users according to their behaviour and to convey them suitable emotional messages. Such techniques could allow third parties to control in some way user's behaviour (e.g., as it is already happening on a certain extent in advertising, companies could use such information to control the behaviour of their customers). Moreover, the emotional, affective sphere is related to the most private aspects of individuals' life and techniques able to deal with it must be carefully considered with respect to privacy safeguard. Of course, as it often happens when dealing with technology, models and algorithms are not ethically good or bad intrinsically. Rather, it is how they are used that determines whether they are ethically acceptable or not. This technology has the power to bring big advantages to humans (consider for example the benefits of an enhanced human-computer interaction in term of diminished stress for people working with computers, the potentialities in therapy and rehabilitation e.g., for autistic children, the possibility to improve the learning process by employing a learning-by-playing paradigm). It has some potential risks too. It is also our responsibility, as scientists and technologists, to fully exploit any possible benefit and to be on guard against any possible misuse.

Appendix A. The EyesWeb open platform

The EyesWeb open hardware and software platform (Camurri, Coletta, Peri, Ricchetti, Ricci, Trocca, Volpe, 2000; www.eyesweb.org) has been adopted for the implementation of the gesture processing algorithms discussed in this dissertation and for the development of the applications employing them (as for example the concert "Allegoria dell'opinione verbale" and the therapeutic exercises illustrated in Chapter 9).

EyesWeb is an open hardware and software platform conceived for the design and development of real-time music and multimedia applications. It supports the user in experimenting computational models of non-verbal expressive communication and in mapping gestures from different modalities (e.g., human full-body movement, music) onto multimedia output (e.g., sound, music, visual media). It allows fast development and experiment cycles of interactive performance set-ups by including a visual programming language allowing mapping, at different levels, of movement and audio into integrated music, visual, and mobile scenery.

EyesWeb is the basic platform of the MEGA EU IST project and has also been adopted in the EU IST CARE HERE project on therapy and rehabilitation. EyesWeb is fully available at its website (www.eyesweb.org). Public newsgroups also exist and are daily managed to support the growing EyesWeb community (more than 700 users at the moment), including universities, research institutes, and industries.

The EyesWeb open platform consists of a number of integrated hardware and software modules that can be easily interconnected and extended. The EyesWeb software consists of a development environment and a set of libraries of reusable software components that can be assembled by the user in a visual language to build patches as in common computer music languages.

EyesWeb includes a software Wizard enabling users to extend the system with new modules, data-types, and libraries.

The software runs on Win32 and is based on the Microsoft COM/DCOM standard; it supports Steinberg VST and ASIO; it supports OSC (Open Sound Control).

Two kinds of modules are currently available: "passive modules" (i.e. filters) and "active modules", i.e., modules with an internal dynamics, which receive inputs as any other module but may send outputs asynchronously with respect to their inputs. For example, the Affective Decision Maker module discussed in Chapter 4 has been implemented as an active module.

EyesWeb libraries include:

- Input: support for frame grabbers (from webcams to professional videocameras), wireless on-body sensors (e.g. accelerometers), audio and MIDI input, serial, tcp/ip;
- Math and filters (e.g. pre-processing, modules for signal conditioning, etc.);
- Imaging (processing and conversions of images);
- Sound and MIDI libraries;
- Communication (e.g. MIDI, OSC, tcp/ip, serial, DCOM, etc.);
- Output: visual, audio, MIDI, serial, tcp/ip, etc.

In the particular framework of this dissertation, EyesWeb has been selected since (i) it allows to interactively map motion parameters onto sounds and visual media in a multimedia scenario, (ii) it allows integration of novel analysis techniques as new libraries or extensions to existing libraries, (iii) it allows fast design, development, and testing of multimedia interactive applications, (iv) it can display in real time the analysed expressive cues, (v) it supports different types of sensors (including wireless), one or more videocameras, and can be programmed to perform specific analysis of movement in real-time. To this last task, the EyesWeb Expressive Gesture Processing and Motion Analysis Libraries (see Appendix B) have been developed and employed, including software modules for extraction and pre-processing of physical signals (e.g., video from videocameras), and extraction and processing of motion parameters.
Figure A.1 shows three examples of EyesWeb patches in which visual output is obtained as a result of a direct mapping of expressive cues extracted from movement.

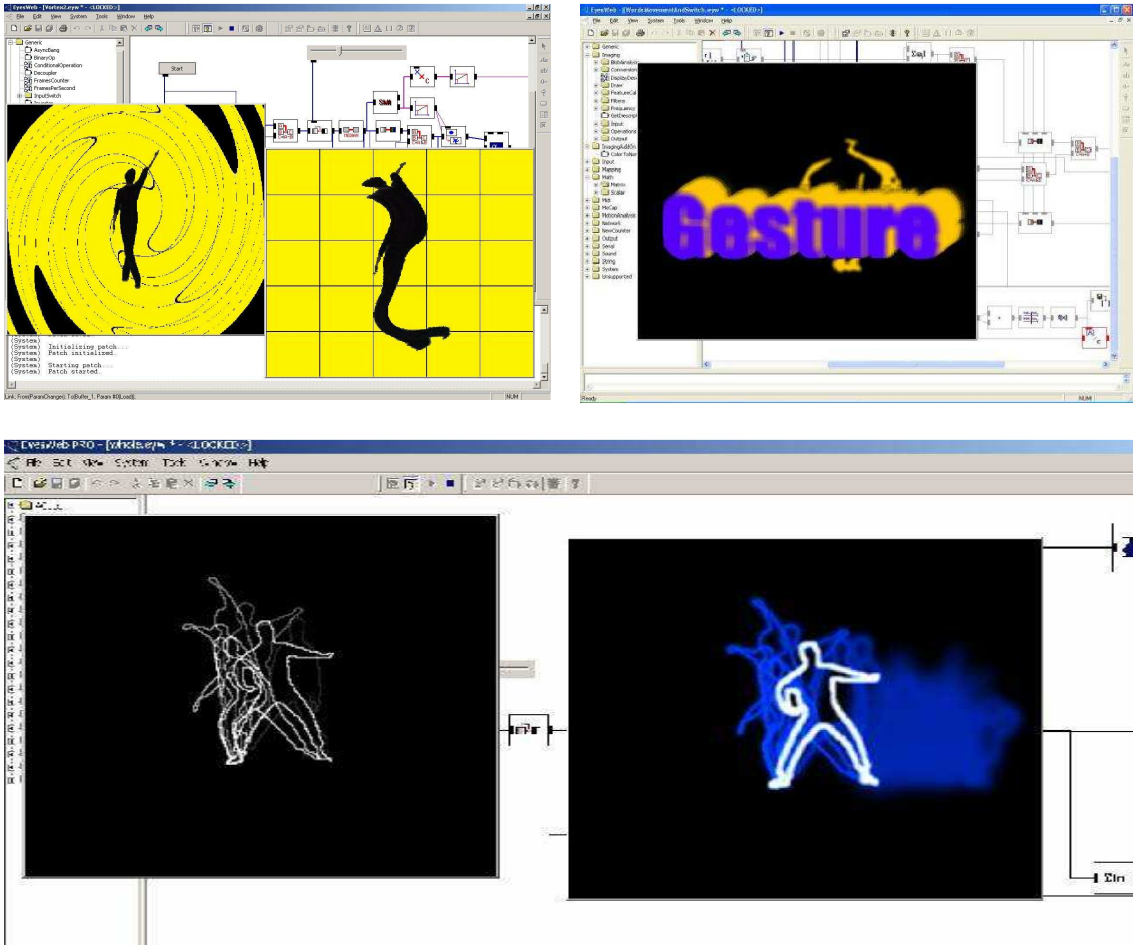


Figure A.1: three examples of EyesWeb patches mapping expressive cues extracted from human full-body movement into real-rime generation of visual content.

Appendix B. The EyesWeb Expressive Gesture Processing Library

The EyesWeb Expressive Gesture Processing Library is a main concrete output of the research discussed in this dissertation. It includes a collection of EyesWeb blocks (software modules) and patches (interconnections of blocks) contained into three main sub-libraries:

- The EyesWeb Motion Analysis Library: a collection of modules for real-time motion tracking and extraction of expressive cues from human full-body movement.
- The EyesWeb Space Analysis Library: a collection of modules for analysis of occupation of 2D (real as well as virtual) spaces.
- The EyesWeb Trajectory Analysis Library: a collection of modules for extraction of features from trajectories in 2D (real as well as virtual) spaces.

The *EyesWeb Motion Analysis Library* applies computer vision techniques to extract expressive cues from human full-body movement. A first layer consists in individuating and tracking motion in the incoming images. Background subtraction is used to segment the body silhouette. Algorithms based on searching for body centroids and on optical flow based techniques (e.g., the Lucas and Kanade tracking algorithm, Lucas & Kanade, 1981) are available. Starting from silhouettes and tracking information a collection of expressive cues is extracted (see Chapter 6). They include Quantity of Motion, Contraction Index, Stability Index (i.e., equilibrium), orientation of body parts, kinematics (speed, acceleration and their average and peak values obtained by using the outputs of the tracking algorithms as inputs to the Trajectory Analysis sub-library), overall motion direction, measures related to the temporal dynamics of movement (e.g., segmentation of movement in pause and motion phases, duration of pause and motion phases, inter-onset intervals as the time interval between the beginning of two subsequent motion phases). A set of modules for posture recognition is also available.

The *EyesWeb Space Analysis Library* is based on the discussed General Space model considering a collection of discrete potentials defined on a 2D space (see Chapter 8). Objects and subjects in the space can be modelled by time-varying potentials (e.g., a dancer on a stage can be modelled as a bell-shaped potential moving around the space). Interactions between potentials can be used to model interactions between (real, virtual, or mixed) objects and subjects in the space. Regions in the space can also be defined. The metaphor can be applied both to real spaces (e.g., scenery and actors on a stage) and to virtual, semantic, expressive spaces (e.g., a space of parameters where gestures are represented as trajectories). The library includes blocks allowing the definition of interacting discrete potentials on 2D spaces, the definition of regions, the extraction of cues (such as, for example, the occupation rates of regions in the space).

The *EyesWeb Trajectory Analysis Library* contains a collection of blocks for extraction of features from 2D trajectories. Such features include geometric measures (e.g., trajectory length, Directness Index) and kinematical measures (velocity, acceleration, curvature). Statistic measures along time (e.g., average, peak values calculated both on running windows or on all the samples between two subsequent commands) and statistic measures among trajectories (e.g., average velocity of N trajectories) are also available. Trajectories can be real trajectories coming from the tracking algorithms of the EyesWeb Motion Analysis Library or virtual trajectories (e.g., trajectories representing gestures in semantic, expressive spaces). The extracted features can be used as inputs to clustering algorithms.

References

Arcos J., Lopez de Mantaras R., Serra X., ”Saxex: a Case-Based Reasoning System for Generating Expressive Musical Performances”, Journal of New Music Research, 27(3), 1998.

Argyle M., “Bodily Communication”, Methuen & Co Ltd, London, 1980

Bahorsky, R. (ed.): “Official Internet Dictionary”, Government Institutes, 1998.

Bartenieff I., Davis M., “Effort-Shape analysis of movement: The unity of expression and function”, in Davis M. (ed.): Research Approaches to Movement and Personality, Arno Press Inc., New York, 1972.

Bates J. “The role of emotions in believable agents”, Communications of the ACM, 37(3): 122 – 125, 1994.

Benford S., Brown C., Reynard G., Greenhalgh C., “Shared Spaces: Transportation, Artificiality, and Spatiality”, in Proc. CSCW'96, 77-85, Boston, Massachusetts, ACM Press, 1996.

Benford, S., D. Snowdon, A. Colebourne, J. O'Brien, and T. Rodden, “Informing the design of collaborative virtual environments” in S. C. Hayne and W. Prinz (eds.): GROUP'97, in Proc. of the ACM SIGGROUP Conference on Supporting Group Work, 71-80, Phoenix, Arizona, ACM Press, New York, 1997.

Benford S., Creenhalgh C., Reynard G., Brown C., Koleva B., “Understanding and Constructing Shared Spaces With Mixed Reality Boundaries”, ACM Transactions on Computer Human Interaction (TOCHI), September 1998.

Bindiganavale R., “Building Parameterized Action Representation from Observation”, Ph.D. Dissertation, CIS Dept., University of Pennsylvania, 2000.

Bobick A.F., Davis J., “The Recognition of Human Movement Using Temporal Templates”, in IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3): 257-267, 2001.

Boone R. T., Cunningham J. G., “Children's decoding of emotion in expressive body movement: The development of cue attunement”, Developmental Psychology, 34: 1007-1016, 1998.

Bradsky G., Davis J., “Motion segmentation and pose recognition with motion history gradients”, Machine Vision and Applications 13:174-184, 2002.

Buck R., “The communication of emotion”, New York: Guilford Press, 1984.

Cadoz C., Wanderley M., “Gesture – Music”, in M. Wanderley and M. Battier (eds.): Trends in Gestural Control of Music, Ircam, 2000.

Camurri A. (ed.): “Proceedings of the International Workshop on KANSEI: The technology of emotion”, AIMI (Italian Computer Music Association) and DIST-University of Genova, 1997.

Camurri A., Coglio A., “An Architecture for Emotional Agents”, IEEE Multimedia, Oct-Dec 1998, 24-33, IEEE CS Press, 1998.

Camurri A., Ferentino P., “Interactive environments for music and multimedia”, Multimedia Systems 7: 32 – 47, Springer-Verlag, 1999.

Camurri A., Hashimoto S., Suzuki K., Trocca R., “Kansei analysis of dance performance”, in Proc. IEEE Intl. Conf. Systems Man and Cybernetics SMC99, Tokyo, Japan, 1999.

Camurri A., Volpe G., “A goal-directed rational component for emotional agents”, in Proc. IEEE Intl. Conf. Systems Man and Cybernetics SMC99, Tokyo, Japan, 1999.

Camurri A., Hashimoto S., Ricchetti M., Trocca R., Suzuki K., Volpe G., “EyesWeb – Toward Gesture and Affect Recognition in Interactive Dance and Music Systems” Computer Music Journal, 24(1): 57-69, MIT Press, 2000

Camurri A., Coletta P., Ricchetti M., Volpe G., “Expressiveness and Physicality in Interaction”, Journal of New Music Research, 29(3): 187-198, Swets & Zeitlinger, 2000.

Camurri A., Coletta P., Peri M., Ricchetti M., Ricci A., Trocca R., Volpe G., “A real-time platform for interactive dance and music systems.” in Proc. International Computer Music Conference ICMC2000, 262-265, Berlin, Germany, 2000.

Camurri, A., De Poli G., Leman M., “MEGASE - A Multisensory Expressive Gesture Applications System Environment for Artistic Performances”, in Proc. Cast01 - Conference on artistic, cultural and scientific aspects of experimental media spaces, 59-62, GMD, St Augustin-Bonn, Germany, 2001.

Camurri A., Mazzarino B., Trocca R., Volpe G., “Real-Time Analysis of Expressive Cues in Human Movement”, in Proc. Cast01 - Conference on artistic, cultural and scientific aspects of experimental media spaces, 63-68, GMD, St Augustin-Bonn, Germany, 2001.

Camurri A., De Poli G., Leman M., Volpe G., “A Multi-layered Conceptual Framework for Expressive Gesture Applications”, in Proc. Intl. Workshop on Current Research Directions in Computer Music, 29-34, Barcelona, Spain, 2001.

Camurri A., Trocca R., Volpe G., “Interactive Systems Design: a KANSEI-based Approach”, in Proc. NIME-02 Intl. Conference on New Interfaces for Musical Expression, 155-161, Dublin, Ireland, 2002.

Camurri A., Mazzarino B., Timmers R., Volpe G., “Multimodal analysis of expressive gesture in music and dance performances”, V Intl. Gesture Workshop, Genova, 2003.

Camurri A., Mazzarino B., Volpe G., “Analysis of Expressive Gestures in Human Movement: The EyesWeb Expressive Gesture Processing Library”, in Proc. XIV Colloquium on Musical Informatics (CIM), Firenze, Italy, 2003.

Camurri A., Lagerlöf I., Volpe G., “Emotions and cue extraction from dance movements”, International Journal of Human Computer Studies, in press, 2003.

Canazza S., Rodà A., “Adding Expressiveness in Musical Performance in Real Time”, in Proc. AISB99 Symposium on Musical Creativity, 134-139, Edinburgh, UK, 1999.

Canazza S., De Poli G., Drioli C., Rodà A., Vidolin A., “Audio Morphing Different Expressive Intentions for Multimedia Systems”, IEEE Multimedia, 7(3): 79 – 83, 2000.

Chi D., Costa M., Zhao L., Badler N., “The EMOTE model for Effort and Shape”, in Proc. ACM SIGGRAPH’00, 173-182, New Orleans, 2000.

Cirotteau D., De Poli G., Mion L., Vidolin A., Zanon P, “Recognition of musical gestures in known pieces and in improvisations”, V Intl. Gesture Workshop, Genova, Italy, 2003.

Clarke E., “Generative principles in music performance”, in J. Sloboda (ed.): Generative processes in music, Oxford, Clarendon Press, 1988.

Clarke E., Davidson J., “The Body in Performance”, in W. Thomas (ed.): Composition - Performance - Reception. Studies in the Creative Process in Music, Aldershot: Ashgate Press, 74-92, 1998.

Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., Taylor J., “Emotion Recognition in Human-Computer Interaction”, IEEE Signal Processing Magazine, 1, 2001.

Cruz-Neira C., Sandin D., De Fanti T., Kenyon R., Hart J., “The CAVE: Audio Visual Experience Automatic Virtual Environment”, Communications of the ACM, 35(6): 65-72, 1992.

Dahl S., Friberg A., “Expressiveness of musician’s body movements in performances on marimba”, V Intl. Gesture Workshop, Genova, Italy, 2003.

Damasio A.R., “Descartes’s Error: Emotion, Reason, and the Human Brain”, Gosset/Putnam Press, New York, NY, 1994.

De Meijer M., “The contribution of general features of body movement to the attribution of emotions”, Journal of Nonverbal Behavior, 13: 247-268, 1989.

De Meijer M., “The attribution of aggression and grief to body movements: The effects of sex-stereotypes”, European Journal of Social Psychology, 21: 249-259, 1991.

De Poli G., Rodà A., Vidolin A, “Note-by-note Analysis of the Influence of Expressive Intentions and Musical Structure in Violin Performance”, Journal of New Music Research, 27(3): 293-321, 1998

Dittrich W.H., Troscianko T., Lea S.E.G., Morgan D., “Perception of emotion from dynamic point-light displays represented in dance”, Perception, 25: 727-738, 1996.

Dixon S., Goebl W., Widmer G., “The Performance Worm: Real Time Visualisation of Expression Based on Langner's Tempo-Loudness Animation”, in Proc. of the International Computer Music Conference (ICMC'2002), Göteborg, Sweden, 2002.

Friberg A., “A Quantitative Rule System for Musical Performance”, Doctoral Dissertation, Royal Institute of Technology, Sweden, 1995.

Friberg A., Sundberg J., “Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners”, Journal of the Acoustical Society of America, 105(3): 469-1484, 1999.

Friberg A., Sundberg J., Frydén L. “Music from motion: Sound level envelopes of tones expressing human locomotion”, Journal of New Music Research, 29(3): 199-210, 2000.

Friberg A., Colombo V., Frydén L., Sundberg J., “Generating Musical Performances with Director Musices”, Computer Music Journal, 24(3): 23-29, 2000.

Gabrielsson A., “Expressive intention and performance”, in R. Steinberg (ed.): Music and the Mind Machine: the Psychophysiology and the Psychopathology of the Sense of Music, 35-47, Springer Verlag, 1995.

Gabrielsson A., Juslin P., “Emotional expression in music performance: between the performer’s intention and the listener’s experience”, Psychology of Music, 24: 68-91, 1996.

Guinn I.C., Biermann A., “Conflict Resolution in Collaborative Discourse” in Computational Models of Conflict Management in Cooperative Problem Solving, Workshop Proceedings from the 13th International Joint Conference on Artificial Intelligence (IJCAI), Chambery, France, 1993.

Hashimoto S., “KANSEI as the Third Target of Information Processing and Related Topics in Japan”, in Camurri A. (ed.): Proceedings of the International Workshop on KANSEI: The technology of emotion, AIMI (Italian Computer Music Association) and DIST-University of Genova, 101-104, 1997.

Hu M.K., “Visual pattern recognition by moment invariants”, IRE Transactions on Information Theory, IT-8: 179-187, 1962.

Hummels C., Smets G., Overbeeke K., “An Intuitive Two-Handed Gestural Interface for Computer Supported Product Design”, in I. Wachsmuth and M. Fröhlich (eds.): Gesture and Sign Language in Human-Computer Interaction, 1998.

Ishii H., Ullmer B., “Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms”, in Proc. of CHI '97, 1997.

Johansson G., “Visual perception of biological motion and a model for its analysis”, Perception and Psychophysics, 14: 201-211, 1973.

Kendon A., “Gesticulation and speech: Two aspects of the process of utterance”, in M.R. Key (ed.): The Relation Between Verbal and Nonverbal Communication, 207-227, Mouton, 1980.

Kilian J., “Simple Image Analysis By Moments” Open Computer Vision (OpenCV) Library documentation, 2001

Kurtenbach, G., Hulteen, E., “Gestures in Human Computer Communication”, in Brenda Laurel (ed.): The Art and Science of Interface Design, 309-317, Addison-Wesley, 1990.

Laban R., Lawrence F.C., “Effort”, Macdonald&Evans Ltd., London, 1947.

Laban R., “Modern Educational Dance” Macdonald & Evans Ltd., London, 1963.

Lagerlöf I., Djerf M., “Communicating emotions in dance performance”, Department of psychology, Univerity of Uppsala, manuscript under revision, 2002.

LeDoux J., “The emotional brain”, Simon & Schuster, New York, NY, 1996.

Leman M., Lesaffre M., Tanghe K., “A toolbox for perception-based music analysis”, (http://www.ipem.rug.ac.be/toolbox), IPEM - Dept. of Musicology, Ghent University, 2001.

Leman M., Vermeulen V., De Vooght L., Taelman J., Moelants D., Lesaffre M., “Correlation of gestural audio cues and perceived expressive qualities”, V Intl. Gesture Workshop, Genova, Italy, 2003.

Liu Y., Collins R.T., Tsin Y., “Gait Sequence Analysis using Frieze Patterns”, European Conference on Computer Vision, 2002.

Lucas B., Kanade T., “An iterative image registration technique with an application to stereo vision”, in Proc. of the International Joint Conference on Artificial Intelligence (IJCAI), 1981.

Machover T., Chung J., “Hyperinstruments: Musically intelligent and interactive performance and creativity systems”, in Proc. International Computer Music Conference 1989 (ICMC89), 186-190, 1989.

Mann S., “’Smart Clothing’: Wearable Multimedia Computing and ‘Personal Imaging’ to Restore the Technological Balance Between People and Their Environments”, in Proc. ACM Multimedia 96, 63-74, Boston, MA, 1996.

Mazzarino B., “Analysis of Expressive Gesture in Human Movement”, Master Thesis, University of Genova, Faculty of Engineering, 2002.

McNeill D., “Hand and Mind: What Gestures Reveal About Thought”, University Of Chicago Press, 1992,

Milgram P., Kishino F., “A Taxonomy of Mixed Reality Visual Displays”, IEICE Transactions on Information Systems, E77-D(12), 1994.

Milgram P., Takemura H, Utsumi A., Kishino F., “Augmented Reality: a Class of Displays on the Reality-Virtuality Continuum”, in SPIE Vol. 2351: Telemanipulator and Telepresence Technologies, 1994.

Naimark M., “Elements of realspace imaging: A proposed taxonomy.” in SPIE Vol. 1457, Stereoscopic Displays and Applications II, 1991.

Ortony A., Clore G.L., Collins A., “The Cognitive Structure of Emotions”, Cambridge University Press, Cambridge, MA, 1988.

Pérez-Quinones M., Sibert J. L., “A Collaborative Model of Feedback in Human-Computer Interaction”, in Proc. Conference on Human Factors in Computing Systems (CHI'96), 1996.

Picard R., “Affective Computing”, Cambridge, MA, MIT Press, 1997.

Pollick F.E., Paterson H., Bruderlin A., Sanford A.J., “Perceiving affect from arm movement”, Cognition, 82: B51-B61, 2001.

Pollick F.E., “The Features People Use to Recognize Human Movement Style”, V Intl. Gesture Workshop, Genova, Italy, 2003.

Rinman M. L., “Forms of Interaction in Mixed Reality Performance – A study of the artistic event Desert Rain” Licentiate thesis, Royal Institute of Technology (KTH), Stockholm, 2002.

Rowe R., “Machine Musicianship”, Cambridge MA: MIT Press, 2001.

Rowe R., “Interactive music systems: Machine listening and composition”, Cambridge MA: MIT Press, 1993.

Russell J.A., “A circumplex model of affect”, Journal of Personality and Social Psychology, 39: 1161-1178, 1980.

Russell S., Norvig P., “Artificial Intelligence: A Modern Approach”, Prentice-Hall, 1995.

Schaeffer P., “Traité des Objets Musicaux”, Second Edition, Paris, Editions du Seuil, 1977.

Shi J., Tomasi C., ”Good Features to Track”, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR94), 1994.

Schubert E., “Continuous measurement of self-report emotional response to music”, in J.A. Sloboda and P. Juslin (eds.): Music and Emotion, Oxford University Press, 2001.

Sloman A., “Damasio, Descartes, Alarms, and Meta-Management,” in Proc. IEEE Intl. Conf. Systems, Man, and Cybernetics (SMC98), 2652-2657, IEEE Computer Society Press, Los Alamitos, Calif., 1998.

Sundberg J., Friberg A., Frydén L., “Common Secrets of Musicians and Listeners - An analysis-by-synthesis Study of Musical Performance”, in P. Howell, R. West & I. Cross (eds.): Representing Musical Structure, London: Academic press, 1991.

Sundberg J., Friberg A., Frydén L., “Music and locomotion. Perception of tones replicating force patterns of walking”, in A. Friberg et al. (eds.): Proceedings of the Stockholm Music Acoustics Conference 1993, 136-141, 1994.

Suzuki K., Hashimoto S., “Modeling of Emotional Sound Space Using Neural Networks”, in Camurri, A. (ed.): Proc. Intl Workshop on KANSEI: The technology of emotion, 116-121, AIMI and DIST-University of Genova, Genova, 1997.

Tarabella L., Bertini G., “Wireless technology in gesture controlled computer generated music”, in Proc. Workshop on Current Research Directions in Computer Music, Barcelona, Spain, 2001.

Timmers R., Camurri A., Volpe G., “The Expressive Functioning of two Acoustic Cues in Three Performances of a Skriabin Etude”, in Proc. XIV Colloquium on Musical Informatics (CIM), Firenze, Italy, 2003.

Trocca R., " KANSEI Information Processing: Toward Analysis of Expressive Human Movement", Master Thesis, Faculty of Engineering, University of Genova, 2001.

Volpe G., "Interazioni fra razionalità, emozioni e comportamento in agenti per applicazioni museali e teatrali: un'architettura software real-time", Master Thesis, Faculty of Engineering, University of Genova, 1999.

Wachsmuth I., "Communicative Rhythm in Gesture and Speech", III Intl. Gesture Workshop, Gif-sur-Yvette, France, 1999.

Walk R.D., Homan C.P., "Emotion and dance in dynamic light displays", Bulletin of the Psychonomic Society, 22: 437-440, 1984.

Wallbott H.G., "The measurement of human expression", in von Rallfer-Engel W. (ed.): Aspects of nonverbal communications, 203–228, Swets & Zeitlinger, 1980.

Wanderley, M., Battier M. (eds.): "Trends in Gestural Control of Music." (Edition électronique), Paris: IRCAM, 2000.

Wanderley M., "Quantitative Analysis of Performer Non-Obvious Gestures", IV Intl. Gesture Workshop, London, UK, 2001.

Weiser M. "The Computer for the 21st Century", Scientific American, 94-104, September 1991.

Wilson, A. D., Bobick A. F., Cassell, J., "Recovering the Temporal Structure of Natural Gesture", in Proc. 2nd Intl. Conf. on Automatic Face and Gesture Recognition, 1996.

Zhao L., "Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures", Ph.D. Dissertation, University of Pennsylvania, 2001.