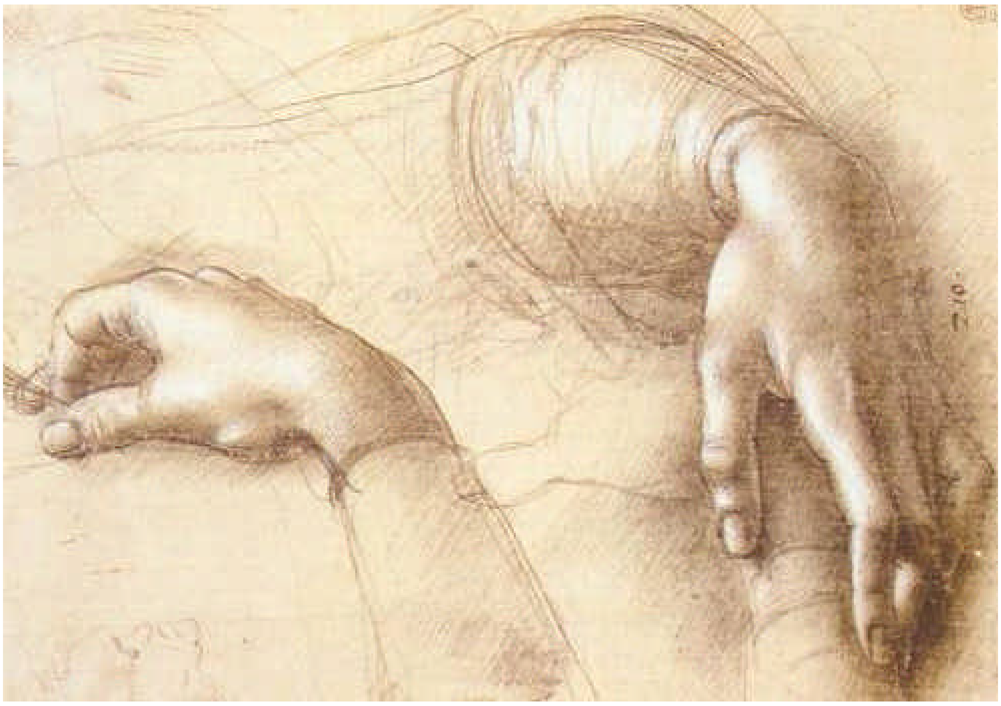
Alessandro Valli

Notes on
Natural Interaction



http://alessandrovalli.com
http://naturalinteraction.org

Contents

Introduction

People naturally communicate through gestures, expressions, movements. Research work in natural interaction is to invent and create systems that understand these actions and engage people in a dialogue, while allowing them to interact naturally with each other and the environment. People don't need to wear any device or learn any instruction, interaction is intuitive. Natural interfaces follow new paradigms in order to respect human perception. Interaction with such systems is easy and seductive for everyone.

With natural interfaces common people can explore ancient Rome, become a character in an old movie, play with virtual fishes, meet Socrates, and all this without the sense of interacting with a machine. The expressions ‘natural interfaces’, ‘natural interaction’ or ‘natural interactivity’ have been used by many researchers in a vague or general way, or from a partial perspective, often to describe something somehow different from common interfaces; the author is persuaded that this concept is a key issue for the next decades, and investigates new forms of relation between people and computers, trying to define what is naturalness and how it can be achieved; this book is one of the results of this effort.

This text is organized as a collection of short chapters that can be read independently. However it is a ordered collection, and the reader that will walk through all the chapters will take advantage of the overall structure. The purpose of this text is to describe the fundamental properties of natural interaction issues, methods and systems, using mainly research prototypes as examples. The aim of this text is to introduce the novice to a new, challenging world, and to give useful hints to the experienced reader, that will take advantage of many qualitative asserts and observations. Formulas and details about specific methods or themes will be made available in the references.

PART ONE: FOUNDATIONS

Goal

The success of a natural interaction system depends on how it influences people experiencing it. Persons should be amazed, they should have fun, get satisfactory (and maybe unpredicted) answers to their questions, have the illusion they are dealing with something that is alive, experience a kind of magic under their control. The term 'experience' is preferred here to the term 'use' since it has a broader sense. A natural interface activates the cognitive and cybernetic dynamics that people commonly experience in real life, thus persuading them they are not dealing with abstract, digital media, but with physical, real objects. This results in a reduction of the cognitive load, thus increasing the amount of attention on content. Of course the key to achieve this goal is the synthesis of a number of aspects, like non obtrusive sensing, visualization, response times, and cognitive load.

In this perspective, the technologies used don’t define the nature of a system. Technology is just a tool in the hands of who creates a communicative space or artefact. In comparison with early interfaces, nowadays interfaces go in the direction of a greater coherence with human perceptive characteristics, but these are greatly limited (the first reason for this is the lack of adequate input technologies). The main feature of the proposed approach is to step further along that way, accepting no compromises in terms of people experience.

Persons experiencing such systems are not necessarily active or willing users, they can be simply passing by and enjoy passively the encounter; that's why in this text the term 'user' will be replaced by 'people' or 'person', or similar terms that maintain in the meaning the richness of factors present in any human being. What are the elements that make an interaction successful? The rest of the chapters will give some ideas about possible answers to this question.

Framework

In order to better explain the particular aspects involved in natural interaction, it is better to depict an overall scheme first, in which all these aspects will take place in the rest of the dissertation. The proposed framework is based on three main functional elements (sensing, intelligence, and presentation) and two methodological elements (influence and cognition).

Hardware and software involved in sensing must be able to deal with common human behaviours, in a strict sense. Devices must be absolutely non obtrusive, and ideally disappear in the environment. Data coming from this module is injected into the intelligent core of the system, whose purpose is to engage humans in a real and convincing dialog. The word 'intelligence' here defines the capability of this module to manage high level communication and convey the illusion of life in the audience, the illusion of dealing with an entity that is not a mere unanimated tool; messages sent to the persons are both content and functional stimuli that enable satisfactory interaction dynamics. Communication happens by means of actuators (e.g. speakers, projectors) following precise presentation rules that permit a better acceptance by humans than common user interfaces (e.g. no windows, no menus, and no scrollbars, just to enumerate some differences).

All these elements will be discussed later in separate chapters, but it is important to remember that they represent different faces of an entity that is one: every element deeply influences each other, and cannot be designed and coded separately. Mutual influence between the single parts is thus another key element. The last and most important factor in the framework is related to cognitive psychology. Cognition observations control the design of every module. The human factor decides what gestures and expressions will be recognized, how narration will be shaped, how information will be rendered and presented. Every detail can impact persons' experience, thus expanding designer's interest to the whole surrounding environment. The experienced reader could argument that this is an almost traditional interaction framework, common to present methodologies and realizations; the novelty is in the goal, as stated above, and in the proposed solutions to push towards this goal.

Natural

Thousands of years of human evolution and the first years of life of every person define what for people is natural and what is not. This definition, although subjective, justifies the following sentence: it is natural to manipulate coloured balls, but it is unnatural to use a computer keyboard. On one hand natural activities are those that humans are made for, implicitly written in the structures of minds and bodies (e.g. the attitude to direct two-handed manipulation of relatively small objects); on the other hand other activities are here referred as 'natural', activities that have a cultural origin, but that are so ordinary in real life that are often considered as ancestral (e.g. deictic gestures or some simple symbols). In some sense ‘natural’ is used here as the opposite of ‘abstract’.

Once there is a stimulus, the brain searches for the corresponding schemes of action, and activates the scheme that is less expensive in terms of effort; if a system induces simple schemes, the interaction is more straight forward and not fatiguing; the higher is the level of abstraction, the higher is the cognitive effort required for mere interaction. It is also common sense that humans use practical, physical, real life metaphors to handle complex, abstract problems; natural interfaces can help also in this direction (e.g. it is easier to think in front of a notepad than in front of a computer screen). ‘Natural’ is thus also a synonym of ‘usual in real life’.

Interaction

The mutual or reciprocal action or influence between entities. These entities can obviously be unanimated objects or persons. From a physics point of view, the interaction consists in the forces that arise between the bodies that come into contact. Interaction between people is the sum of the activities of the involved subjects, like speaking, gesturing, listening, watching. If a system simulates physical objects or characters, and is able to perceive human actions and intentions, people can interact with them using the ordinary schemes that they use in real life, and this is the key to a new level of satisfaction.

It is a commonly accepted concept, as stated among the others by Bateson, Maturana, and Varela, that there is no perception and knowledge without interaction with the environment. The role of body motion as a mean of expression and a tool for exploration is a forgotten dimension in western culture. It is even less considered in computer interfaces, also because the devices and sensors that one has to wear reduce heavily the pleasure of moving into space. The paradigm to rely on is natural interaction of people with other people and with objects of common use.

Leonardo

Natural interaction research is a point of intersection of computer science, mechanical engineering, creative design, cognitive sciences, art, architecture, just to report some disciplines. The multidisciplinary approach is a must in such research, and cannot be productively managed by experts of the different fields working together: anyone must learn the fundamentals of each discipline, to have a common ground to build on.

Leonardo da Vinci represents a model of researcher across different disciplines, showing the benefits of cross-contamination between different fields. He was an artist, a scientist and a technologist and at the same time. Flavia Sparacino wrote: "The European renaissance has given birth to two typologies of intellectuals: the scientist type, incarnated by Galileo, who first established rules for scientific experimentation and scientific method and the artist-engineer, incarnated by Leonardo, involved in a creative research equally informed by art and science. [...] While the Galileo-scientist type has been predominant in western culture since after the renaissance, the emergence of digital media favours the reappearance of the artist-engineer, equally versatile in artistic creation and engineering abilities". Only a complete figure like Leonardo can handle creatively the complexity of natural interaction research, without considering the divisions among the different fields.

Natural interaction is thus a leonardian science: human has to be considered in its unity, experience in its relation with all the factors.

Cognition

One of the main effects of natural interaction is the reduction of the cognitive load on the subject. What does this term from cognitive science mean? In order to answer this question a short insight is needed. Scientists classify memory in three main areas: sensory memory, working (or short term) memory, and long term memory. Working memory purpose is thinking, reasoning, learning; data from sensory or long term memory is copied to working memory in order to be processed. Working memory is extremely limited: it can contain an average of seven elements at a time in adults (the range is five to nine, varying from person to person). Single elements are often clustered into groups based on meaning or other characteristics. Sensed data of different nature is processed in parallel (e.g. visual, audio and linguistic).

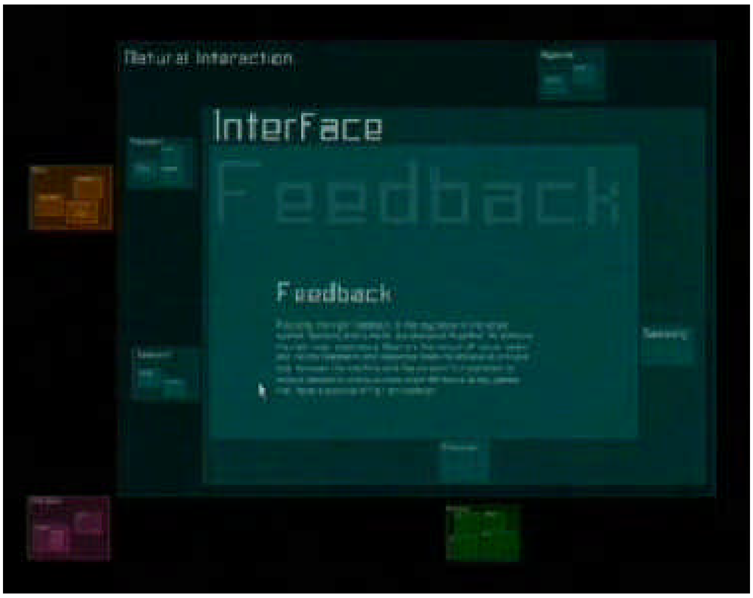
Cognitive load is the total amount of activity imposed on working memory at a moment. The major factor that contributes to cognitive load is the number of elements that need to be attended to. Cognitive load is the sum of an intrinsic part and an extraneous one. Intrinsic load depends on the content itself, while extraneous load is due to the form of the representation. The strategy proposed here consists in representing high level information (especially context data) through elementary sensory stimuli, so that it does not require a large amount of mental activity to be processed.

Another relevant element about cognition is visuospatial perception: humans’ ability to process and interpret visual information about where objects are in space. It represents the relation between physical space around the person and what the person sees. As interaction between human and machine moves from the computer screen to the environment, this aspect becomes fundamental, and can be exploited by mapping content and relational information to the space around the person. The next chapters will propose solutions based on this observation.

Context

This text refers to the term ‘context’ in two ways. The first one is related to the context in which interaction takes place. Humans are incredible perceptors: every detail in the area covered by senses influences the way people discover and interact with the focus of their attention. For this reason it must be clear to designers that it is not the system to communicate, it is the full space. Systems must be able to perceive human context in order to modulate communication depending on what is happening. Context awareness can help in interpreting people actions and intentions. Architectural and interface design can help limiting the context the system must be able to deal with;

The second dimension of the term ‘context’ is related to the information world. Infoscapes can be much more complex than landscapes is used to perceive; for this reason, it is necessary to simplify the bulk of information by grouping and hierarchy. People need, while focusing on detail, to perceive the whole informational context, in order to be able to understand the collocation of the detail and to move to similar or related subjects. By properly representing this relational structure, cognitive load is reduced, since the abstract information map must not be actively created and kept in memory.



Attention

From a psychology point of view, attention can be defined as an ability to focus and maintain interest in a given object; it is the condition of reasoning and learning. Systems must be able to get and hold people attention: stimuli must be strong enough to cause this. Attention on a given particular should be released when another stimulus gets the focus. The center or focus of attention is the detail one is thinking about, and is managed by working memory. Periphery of attention is the processing of the rest of the stimuli one can perceive, performed by sensory intelligence.

Mark Weiser wrote: “There is no less technology involved in a comfortable pair of shoes, in a fine writing pen, or in delivering the New York Times on a Sunday morning, than in a home PC. Why is one often enraging, the others frequently encalming? We believe the difference is in how they engage our attention. Calm technology engages both the center and the periphery of our attention, and in fact moves back and forth between the two”. Interface should be as invisible as possible, leaving attention on content, thus reducing fatigue and distraction. Moreover, it should allow smooth passages between center and periphery, something that is absolutely missing in common realizations.

The same medium can at the same time render information for the focus of attention and for the periphery, increasing usability, as far as it is possible to seamlessly move the focus on every piece of information.

Periphery

The role of periphery of attention requires an additional insight, since it is a dimension completely unknown to common interfaces. William Mitchell wrote: “Peripheral information is by no means unimportant; in fact, it plays a crucial role in establishing the character of a place and sustaining your relationship to it. When a room has a window, for example, it provides a continuous flow of information about the external environment - the cycles of day and night, the movement of sunlight and shadows, the succession of bright and cloudy moments, and the alternation of dry and rainy patches. You rarely pay explicit attention to all this, but you are peripherally aware of it, and you feel uncomfortably isolated if you are cut off from it”.

In order to be effective, peripheral information has to be coded into elementary messages that can be processed by sensory intelligence. This leads to the use of colour, shape, sound characteristics, size, volume, agitation as variables expressing what is going on and the state of context. Ambient displays are devices whose purpose is exactly this: conveying non critical (i.e. not requiring immediate attention) information peripherally.

Properly represented stimuli can make people aware of a lot of data without requiring continuous and intensive polling of the sources of information. Weiser wrote: “What is in the periphery at one moment may in the next moment come to be at the center of our attention and so be crucial. First, by placing things in the periphery we are able to attune to many more things than we could if everything had to be at the center. Things in the periphery are attuned to by the large portion of our brains devoted to peripheral (sensory) processing. Thus the periphery is informing without overburdening”.

Social

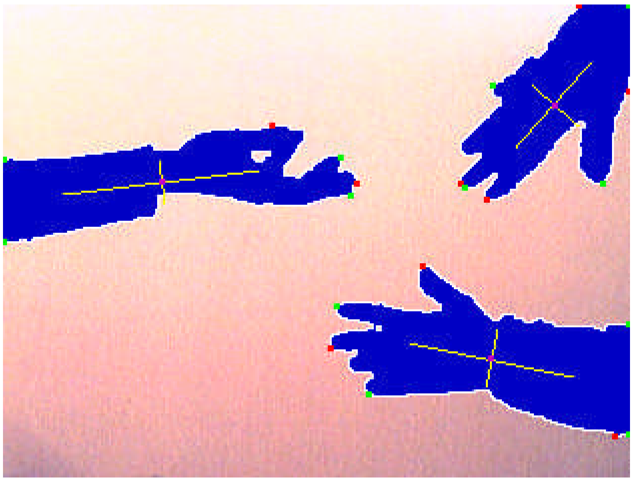
Interaction with computers is usually demanding a total separation between the person that uses the machine and the rest of the world. All the attention of the subject is drawn by the screen in front of him. The social dimension is destroyed. A natural interface allows normal relation between the persons in (active or passive) contact with the system. This is achieved in many ways. Persons can move freely inside or around the interactive space, with no (apparent) restrictions. Design of the space takes in account this, not imposing the subjects to sit in front of a screen or to wear devices they would not wear in normal conditions.

The system plays as one of the actors between other people, and can interact with different persons at a time. Moreover, there are no modal constraints like ‘begin’ or ‘end’ phases of interaction: one can ask the system for something, speak with another guy, go away, come back and find the system ready to interact again. These spaces should respect the cultural, social and organizational contexts that host them; this is not a matter of additional functions: it is more likely a problem of accurate design. Depending on context, such systems should allow normal conversation among the participants; allow them to behave as they would normally do, allow multiple simultaneous interactions and collaboration. These should also be able to disappear completely as needed. Only this can actually make the system disappear and enrich the environment.

Competition and collaboration

As computing moves to extended environments, the problem of dealing with more persons at a time arises. Multiple simultaneous interaction is thus a fundamental issue in interactive spaces: reducing interaction to one person at a time is not well accepted by the audience, that prefers more flexible solutions. In presence of multiple persons with different goals conflicts can happen, since system's resources (e.g. physical space, effectors, and screen surface) are limited. The problem is critical especially concerning audio, since it is difficult to localize this kind of medium.

A useful approach is represented by seamless multi-user spaces that allow both collaborative and competitive behaviours. It is the same scenario as a common workbench: depending on persons' actions, collaboration can occur, and people work together towards a common goal; otherwise, participants will compete, and media resources will be split between the different tasks. Offering a unique setting that fluidly changes its attitude is a means of adaptation to context.



Aesthetics and emotion

There is no doubt that people are attracted by beauty. Every natural interface should appear as a (good) art installation, no matter if it is an intelligent timetable in a railway station or a map viewer in a town hall. The cure for aesthetics influences the quality of the shown media content, the overall setup and its integration with the environment surrounding it, the feedback to people actions and intentions. The reason for this is that it has to move from a research laboratory to a public space, to a museum, to a theme park, to a house; it has to be an element that aesthetically enriches the surroundings.

Aesthetics design should consider the overall space and people in it as a whole, orchestrating physical and virtual elements with multiple human behaviors. Minimal, essential setups usually work better than complex ones, highlighting the role of content.

Emotional aspects are usually ignored in computing environments; in natural interfaces, instead, these play an important role: amazement, fear, admiration influence human perception and learning. An interactive space is seductive by nature, and has a power of conveying emotion greater than any standard computer setting. Correct use of light, video and audio can tune emotional engagement of people. At the same time, sensing algorithms can estimate emotional responses from the public, such as excitement and attention, from movement and audio measurements; these estimations can be used to tweak communication.

Movement

Humans have a special ability in perceiving movement through vision (it is a gift of evolution). Consider a person that wants to draw someone's attention on a particular; he or she will agitate, move his or her hand around or in direction of that particular. Humans associate motion to life; it is impressive how an object or scene in motion can attract attention for even long times in comparison with a static or repetitive scene. Motion is the sign that something is changing, that a novelty is coming; this is why it is a so important element in narration and interaction.

In these systems this dynamic can be used, for example, in two opposite ways: on one side, the system can attract people attention by moving visual objects, on the other side the system can be able to detect and recognize voluntary movements whose aim is to get 'attention', and behave consequently. Movement is thus a key element (in all its forms) for the development of successful interaction engines. This concept, under different points of view, will be detailed later.

In real life, each status change is accompanied by a transformation, movement that takes place in a finite but not infinitesimal time interval. The use of movement in computer systems can communicate to the user the nature of the change in a manner that is easily perceived.

Freedom

People in naturally interactive spaces must feel free; they must not feel they have to follow a thousand rules. There is a trade off between this requirement and a limitation of the interaction context that is needed to make interpretation of people behaviors easier. The solution is to hide the constraints so that these are accepted spontaneously. It is better to place a good looking architectonical barrier than a sign saying “stop here”. Proper implicit stimuli, either physical or digital, are the best elements to induce behaviors and restrict the scope of acts that have to be recognized.

Moreover, interaction must not be mechanical. Persons using the system must not feel as part of a mechanism. Inclusion of a sufficient number of degrees of freedom (although easy to explore) ensures the possibility for people to play, discover, even fail and learn from errors. There is obviously a trade off between complete control over the interface in every moment and successful narration. On one side the risk is that of repetition of acts and fragmentation of communication; on the other side there is the risk of rigidity: people feel they are in front of a movie, having little control over what is being told to them.

In gaming and artistic scenarios the solution is simple: media content is governed by a set of rules (that has to be sufficiently broad), and interaction is the exploration of such rules. In narrative scenarios the problem is harder to solve, since there is the requirement to communicate content that has a fixed evolution in time; this will be addressed later.

Embodiment

The purpose of a natural interaction framework is to remove any level of mediation between the person and the object of his action; the medium becomes an active element of interaction. Paul Dourish wrote: "Embodiment refers to the way in which interactive resources are manifest in an interface. It does not refer simply to physical reality, but denotes a participative status. It points to the ways in which we interact as involved participants rather than detached observers. [...] It strikes to make computation (rather than computers) directly manifest in the world so that we can engage it using the same sets of skills with which we, as embodied individuals, encounter an embodied world. So, it exploits our physical skills, the ways in which we occupy and move around in space, and the ways in which we configure space to suit our needs. Embodiment, for this side of the research activity, explores the relationship between the environment and the task in hand".

Not only physical relations are embodied, also a conversation is an example of embodiment. Relation with the digital world takes the form of human to human and human to objects relations, preserving social and cultural dimensions. Instead of applying natural dynamics in an unnatural setting, digital media and information are moved to the natural world, inducing natural interaction schemes. This way, the computer disappears, and is not perceived as a barrier between people and information. Abstraction is substituted with experience.

A large screen is a large screen, but in an interactive space it should always behave like a window, or like a table, or like another artefact from common life, and depending on its role, it induces different spontaneous behaviors in people. In the same way, physical or media objects functional to interaction can be shaped and play different roles, suggesting multiple interaction modalities.

Architecture

The physical space where a system or installation is placed is part of the system itself, since it influences people. In this sense, the architectural dimension is very important and is much relevant to the overall experience of the person. A plenty of powerful tools in now coming available to space designers and architects, expressive tools that enable new ways of communication and interaction between the physical space and people; William J. Mitchell, from the School of Architecture and Planning of the MIT, wrote: “Architecture is no longer simply the play of masses in light. It now embraces the play of digital information in space”.

The problem is that the traditional figures that design spaces don’t have all the knowledge needed to use these tools; Flavia Sparacino wrote: “...technology is not simply hardware or software that the space designer and the media artist add to their projects to make them work. It is really not sufficient to wait for technologists to develop new modalities of interaction and man-machine communication in their laboratories, to later incorporate these in space design, as software that one buys at the store”. Space design is like a language; if a traditional designer knows only the letters A to L, he will express with a language made up of that letters; only who masters an alphabet going from A to Z will be able to valorise all the letters and build a successful phrase.

Physical space

It is not the single element that communicates with the person, is the overall environment (comprising all the other persons in it) to communicate; and this implies that the design must take care of all the elements as a whole. Since interaction is no more restricted to a narrow scenario (e.g. screen and mouse), the interface invades large physical spaces (e.g. by means of projections on walls and floors), and thus space becomes a very important element influencing visual feedback nature and sensing technology. Simon Greenwold introduced the interesting concept of spatial computing; he wrote: “Spatial computing is human interaction with a machine in which the machine retains and manipulates referents to real objects and spaces. It is an essential component for making our machines fuller partners in our work and play. […] It is not enough that the screen be used to represent a virtual space—it must be meaningfully related to an actual place”.

Human mind naturally locates information and concepts spatially, and the opportunity to work with full environments allows a strict mapping between physical space and abstract information, favouring human perception and interaction. Physical space can be enriched by digital information, and digital information can be made more accessible and understandable by a mapping to physical space.

Integration

A natural interaction system is commonly made up of many components: physical objects, projected light, screens, computers, mathematical models, cameras or other sensing devices, speakers, digital media... all these pieces must be integrated efficiently, in order to disappear to people as single elements and favour the illusion of interacting with an entity that is a whole. Hardware and software optimization and integration is an issue in the realization of such systems. State of the art real-time sensing, interpretation, behaviour simulation and information rendering is necessary to convey the sense of this illusion.

From a purely technical point of view, very low response times are crucial to this end. Latencies on any part of the feedback loop can easily make the interaction unfeasible and frustrating. For example, for discrete commands, like a pointing gesture, a response time of at most 600 milliseconds can be tolerated; continuous commands, like drag and drop, must usually guarantee a maximum response time of 150 ms. The exigency of processing multiple video streams from the cameras, simulating system's behaviour, and rendering video and audio at the same time raise the problem of a overall optimization of all the processes. Experience, methods and tools from the computer game industry are a good starting point to obtain the needed performance.

Currently, Microsoft Windows XP, DirectX and Flash MX probably provide the best platform to build natural interfaces due to the plenty of software tools and libraries available to interface with cameras and other sensing devices and to exploit graphics and audio hardware. In addition, this platform is not expensive and is well know worldwide to common people (e.g. museum staff). System integration is a complex problem since it is a delicate trade off between many factors, but a good configuration is also a great added value to people experience.

PART TWO: SENSING

Perception

Machine capabilities to understand inputs define the possible levels of interaction. Prof. Alex Pentland wrote on Scientific American: “The problem, in my opinion, is that our current computers are both deaf and blind: they experience the world only by way of a keyboard and a mouse. Even multimedia machines, those that handle audiovisual signals as well as text, simply transport strings of data. They do not understand the meaning behind the characters, sounds and pictures they convey. I believe computers must be able to see and hear what we do before they can prove truly helpful. What is more, they must be able to recognize who we are and, as much as another person or even a dog would, make sense of what we are thinking”.

Sensing technologies should not be considered as external peripherals added to the system, but as part of the overall intelligence; their functioning is heavily influenced by the rest of system components: high level processing, presentation and physical space. Sensing algorithms must be absolutely robust in order to provide predictable outputs in a public environment. Uncertainty about system’s perception leads to person’s dissatisfaction and ruins interaction; reliability is thus a key issue.

What kind of behaviours should these systems be able to detect? Presence, being able to understand when a person arrives or leaves is the most basic capability, and permits basic interactivity. Location, where a person is is relevant to communication, both because it allows localized information broadcast and because it can be a means of information selection. Interest, since attention plays a fundamental role in a dialogue. Selection, through which a person can specify his interest in a particular (i.e. by pointing or touch). Relation between objects or places, and dragging, to move objects. Full manipulation of physical artifacts. In artistic or entertainment contexts even theatrical gestures can play an important role.

Gestures

Gestures are a primary way of communication with the machine, and under some conditions can be recognized reliably. Following Turk's definition, "Gestures are expressive, meaningful body motions – i.e. physical movements of the fingers, hands, arms, head, face, or body with the intent to convey information or interact with the environment". The first function, communication, has been defined as 'semiotic', while the second one, 'manipulation', can be defined as 'ergotic'. Even static gestures, like poses or postures, express an activity, a mood, an intention, an interest. Humans have two different spaces in which there are natural or spontaneous gestures: the manipulation space (i.e. the space can be reached with the hands or feet), and the external space.

Most of human motions are gestures that accompany speech (gesticulation) and meaningless motions for comfort and equilibrium. Other categories are much more relevant to the purposes of interaction; a short classification of simple and useful motions is reported here. Deictic gestures, pointing actions that refer to objects inside or (more likely) outside the manipulation space. Pathic gestures, that represent a path or a direction (i.e. the gesture for 'walk to your left'). Mimic gestures, actions that copy some other motion, like waving arms like a bird. Ergotic gestures, natural inside the manipulation space of each person, to move or rotate an object or to open a box. Even the mere position and orientation of a person in a room can convey a lot of information about his momentary interest.

Deictic gestures can be induced by foreground objects that are easily distinguished by the background, and that a 'promising' appearance, i.e. represent a particular content or concept. In the same way, ergotic gestures can be induced by physical or virtual objects at hand. Mimics should be suggested by the interface, and can have a great potential in storytelling scenarios, where people can play the role of some character inside the story. More abstract motions, like symbolic gestures or sign languages, are less feasible for natural interfaces, unless the context justifies their use; so are all kinds of gestures that presume a training phase.

Gesture recognition

The goal of gesture recognition is not to measure metrical parameters of a motion, but to recognize the intention that the action signifies. The same action can mean different things in different contexts. To make machines able to recognize purposeful motor activities, processing algorithms need to deal with the great variety of shapes and different styles a gesture can assume. Moreover, purposeful motions are hidden in a weave of movements of other nature and expressive gestures addressed to others; the same gesture can be slow or fast. There are several strategies that can be put into action to overcome these difficulties.

A priori knowledge about the context in which gestures take place can be used in order to build effective heuristics or machine learning algorithms. Analysis should focus on invariants in gesture characteristics, thus allowing easier detection and recognition. Gestures chosen for interaction should not be too similar, in order not to require extremely precise motions. The role of feedback is fundamental: visual control (although undesirable as an additional level of mediation) allows easier interaction, and makes people aware of what the machine is understanding. Proper feedback loops can enable spontaneous processes of disambiguation.

Control through gestures must be based on simple and intuitive motions, in order not to increase the cognitive load of the subject, and must be used when really needed, not to physically or psychologically tire the subject. In mixed reality setups, where a person sees his image on screen, with superimposed virtual objects and graphics, allow visual control loop, and move the perceived manipulation space from the surroundings of the body to the screen space. In any case, the recognition process must be executed in strict real-time, and must provide feedback to the user instantly, because this ends the expressive action of the person, releasing attention.

Among the gestures that can be usefully recognized two main categories can be determined: those signifying discrete commands, and those allowing continuous control through positioning, normally implemented with visual control feedback.

Computer vision

Digital cameras are good input devices for a number of reasons: are inexpensive, off-the-shelf products, follow standard protocols, are easily interfaced to computers, are small, can be placed far away from people, out of reach, allow non obtrusive sensing, are robust. Real-time computer vision techniques can be implemented in order to process live video streams and extract information useful to interaction. Among the variety of possible approaches, the proposed methodology exploits colour (or brightness) based segmentation and clustering since it is a fast, robust and reliable processing technique, due to its integral nature.

In order to manage the (huge and noisy) streams of data provided by cameras, efficient software algorithms must be implemented; such procedures incorporate statistical methods and a priori knowledge to enable successful interpretation. Sensing intelligence is moved from the hardware device to software code. Interactive spaces can be observed by multiple cameras that can be placed in order to get the most informative views on the scene. Optimized code allows to process multiple video streams and manage graphics and audio for the interface in real-time (e.g. 30 fps for vision algorithms with two cameras, 85 fps for screen rendering) on a single personal computer, providing an inexpensive and reliable platform.

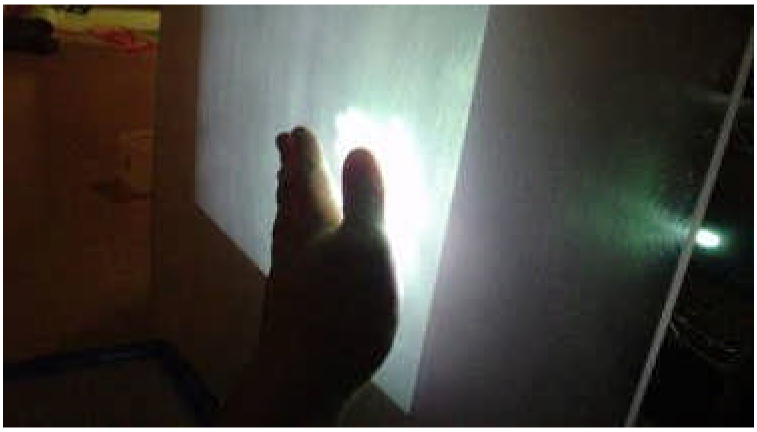
As computation leaves the world of clean zeros and ones and moves towards reality the problem of noise arises. A variety of mathematical models has to be implemented in order to deal with camera noise, human behaviour modelling, visual control dynamics etcetera. The real world is so noisy that most of the overall processing resources are spent to manage it.

Visible and infrared

Human eyes can see light in a spectrum ranging from red through violet: the visible spectrum. The spectrum above violet is ultraviolet, and the one under red is infrared, both invisible. Sunlight comprehends all these components. Deep infrared frequencies can be used for thermal images (i.e. to sense heat), while the part of infrared light close to visible is called 'near infrared' (NIR), and behaves almost like visible light. There are illuminators that emit invisible light in the NIR spectrum, and common CCD or CMOS sensors are sensible to NIR: changing a band blocking filter from the camera enclosure provides a powerful and reliable (achromatic) sensing apparatus.

Visible light allows tracking and interpretation based on colour information (e.g. skin detection, chromakey, detection of coloured objects), but requires more constraints in setup, in order to limit variations in external illumination, since strong and sudden changes could prevent correct functioning. Near infrared imaging does not provide chromatic information, but can overcome the problem of sensibility to external light, since it allows the use of powerful illuminators (of invisible light), that make the system insensitive to external light.

In many cases both solutions can be feasible, and the advantages and problems of each method should be examined deeply case by case in a preliminary design phase. Of course infrared imaging allows even to work in conditions of complete darkness, preserving almost complete freedom in terms of architectural requirements.



Early processing

The first processing phase consists in classifying each pixel of the input image according to some criteria. The colour of a pixel is represented with one to three numbers, depending on the colour space (e.g. RGB, YUV, greylevels); it is possible to perform this classification measuring the color distance between the pixel and a reference pixel, by means of linear or quadratic metrics. Depending on setup, image differencing can be implemented in a number of ways. If the reference image model is the mean of the last processed frame, the classifier will have a motion detection behaviour, and all moving parts will be extracted. If a single colour is used as reference, a colour segmentation behaviour occurs, extracting image parts that possess (or not) specific colour characteristics. If the reference image model is computed as a mean of frames recorded with an empty scene, the result is background subtraction: people in the scene are extracted from a static background.

To deal with camera sensor noise, simple mathematical models must be applied, modelling image evolution in time by means of median filters and covariance matrices. To reduce the negative impact of shadows cast by people, chromaticity measures can be included in the metrics, such as colour components normalized respect to luminance, to exclude from the foreground areas pixels that seem just (a little) darker than the reference.

Thresholding can be performed either with a fixed threshold or with adaptation; in this case CCD camera noise becomes an important resource, since it can be measured counting the number of isolated foreground pixels, thus enabling a sliding threshold dynamically set just above noise level. In addition to colour information, also spatial information can be exploited for classification. Topological filters (erosion and dilation) reduce camera noise effects by analysing neighbourhoods instead of single pixels.

Image analysis

Once pixels are labeled, they have to be clustered. Useful information is not related to single pixels, but to image regions. Two grouping strategies are presented here: grid based and blob based. The first one consists simply in dividing the processed area in rectangular portions (e.g. 1 x 1, 8 x 6), assigning each foreground pixel in a portion to the same region. The second one is more complicated and is based on the detection of connected regions: blobs. For each region a series of features is computed: area, centroid position, bounding box, contour points, colour statistics, extreme points along various directions, and moments.

Image moments are statistical descriptions of the morphological properties of a blob: central moments of the second order contain information about displacement respect to centroid: main axis orientation, extension of the main axis, extension of the orthogonal axis; third order central moments contain shape information, from which Hu moments can be computed, that are invariant to (small) rotations. From these numbers, for example, it is possible to recognize the pose of a hand or classify full body postures.

Extracted features are referred to the image plane, i.e. to the camera view. It is often required to refer these measurements to physical space or to screen plane. A geometrical mapping between the two spaces is needed. This is solved using mathematical mapping functions, based on affine or perspective view models (and that can consider even lens distortion effects), governed by a set of parameters. The calibration process is the estimation of such parameters, given a sufficient number of mapping examples (e.g. ordered couples of source space points and destination space points); after calibration, the mapping algorithm generalizes the solution to all the points in the required area.

Interpretation and accuracy

Once low level features are extracted, high level information has to be estimated, depending on context. This is the processing phase that requires more personalization and adjustment, depending on setup and desired interaction. Public intentions are recognized by estimation of people positions, hands and heads trajectories, spoken words, verbal and mimic expressions.

After initial calibration, the system must adapt to single persons' behaviours and styles, i.e. dynamically create temporary parameters' adjustments that will be lost as the person leaves. This can be accomplished in a number of ways, exploiting feedback (e.g. a visualized button with central symmetry will induce people to select its center; if the sensed selection point is slightly different, the measured displacement vector can be used to correct system calibration; similar tricks can be applied in different contexts). Adaptation affects geometrical, temporal and behavioural parameters.

The concept of accuracy is highly ambiguous, since it depends on the context. In natural interaction systems, that are meant to bring relation with machines at the same level of human to human communication, the accuracy needed is the same that humans have. To explain this concept, an example is useful: a person points at a particular far away from him; there is no need for the system to be more accurate than the estimation that a second person observing the pointing gesture could perform.

Computer vision techniques, working with inexpensive and off-the-shelf hardware, have a linear accuracy of less than a centimeter, in setups where the object is two meters away from the camera(s): more than what is needed. The estimation of metrical parameters is not used for replication of movements or mere measurement; it is used for motion interpretation, thus allowing less precise tracking. Accuracy can be increased by coupling inputs from different sensors, and by proper context modeling.

Stereo vision

Combined use of multiple cameras, i.e. views, allows depth computation, thus enabling full three dimensional motion estimation. This is done by merging features' information from (both) cameras, and using specific calibration algorithms to transform multi dimensional information to the desired three dimensional representation, by means of minimization. In contrast to classical approaches (e.g. disparity maps) that extend stereo computation to large sets of features, and later select useful information, the proposed solution performs labelling and thus selection of features on the two dimensional source views; triangulation is then applied to ordered couples of points or lines, reducing computational effort; possible ambiguities can be solved with the 3D estimations. This symbolic approach is made possible by correct setup design and successful modelling of a priori knowledge.

Stereo vision is needed also in setups where interaction occurs with two dimensional presentations; a simple example is pointing at screen, that is a relative act whose meaning is defined by the three dimensional location of the eyes and the pointing hand or finger of the person.

Audio analysis

Audio is another important source of information. Traditional speech recognition doesn't work well in noisy environments, and often requires microphones carried close to the speaker's mouth. In order to deal with situations in public and social spaces, two approaches can be implemented. The first is the use of array microphones, i.e. audio capture devices that contain many microphones and a digital signal processor that allows to detect the dominant audio source, eliminating the other ones; its functioning is based on the relative latencies between signals taken from different microphones. The second solution consists in the use of a directional microphone oriented towards the user; user position is usually predicted by setup constraints (i.e. user stands in front of a screen).

The author experimented audio analysis techniques alternative to traditional speech recognition applications. Real-time audio capture provides a series of chunks of digital data representing small intervals of the audio stream. Discrete Fourier Transform is applied, and frequency descriptors are compared to a given database. This way command recognition (on small vocabularies varying depending on context) working in noisy environments occurs. Through the same technique useful 'cry' and 'chat' detectors have been implemented, and a microphone attached on a screen has been used to easily detect knocks on the surface. Audio analysis becomes a powerful tool specially when it accompanies computer vision: combined interpretation provides reliable results.

Feedback

Action is guided by feedback. By properly modulating visual, audio and tactile feedback, a system can enhance interaction, and induce person’s expressions that can be easily interpreted; moreover, it can help disambiguating unexpected behaviors. Cybernetic loop plays a fundamental role to this end. Visual control happens when visual feedback is continuous. In this case people move referencing to media space instead of physical space (e.g. with mouse and pointer on screen). It is always preferable to limit feedback elements, since these stimuli could impair content perception; their role should be as minimal as possible, providing it is sufficient to make interface fully accessible and efficient.

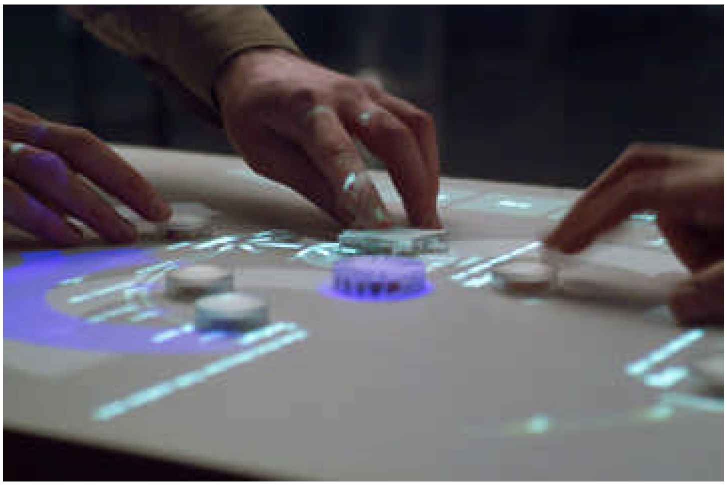
Efficient feedback is a matter of design, of predictions and latencies of milliseconds, of proper smooth filtering. Content and functional stimuli should be merged in a seamless way; since the person is waiting for the end of the task he is performing (e.g. pointing, activating a function, etcetera), immediate feedback should be provided as the system has recognized a given command. This can even convey a sense of action completion, where the system predicts what the one wants and accommodates accordingly.

Interface has to be transparent, people have to sense that media behavior is under their control; feedback should always provide information about the internal state of the system, that thus becomes externally self evident. Lack of feedback can induce frustration and unnatural expressive behaviours, leading to unpredictable results.

Stimuli

Every interaction, of any kind, has a beginning. If a person knows the purpose of an object (e.g. a hammer) and how to use it, there is no problem: if he needs it, he can interact with it successfully. If the object is an artefact unknown to the person the problem arises: how can I interact with it? The interface must contain all the hints necessary to allow a satisfactory interaction. The hardest work is to convey the initial stimulus, the hint that causes the first voluntary action of the person towards the system. Once interaction is engaged, it will be easier for the person to learn the additional interaction capabilities of the system.

Another role of such stimuli is to make people express their intentions in ways that the computer is able to understand (e.g. if the person sees a painting displayed three or four meters away from him, in a position he cannot reach, it will be straightforward for him to move his arms in the direction of the painting, to see if something happens, but if some visual cues appear on the painting, suggesting some particulars, he will probably point at some of the particulars with a finger). The more these hints are absolutely hidden in content the more these are acceptable to the audience. There are two dynamics that can be put into act to favour interaction without giving instructions or manifest indications: intuition and imitation.

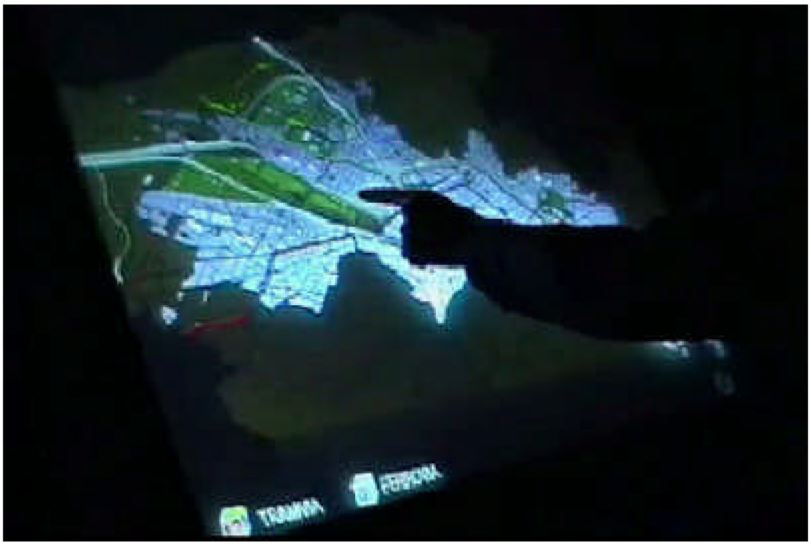


Intuition

The intuitiveness of a natural interface is a must. These interfaces are mostly oriented to public environments, with people passing through that moved by curiosity try to interact with the system for a short time. The encounter can last for a few seconds or a few minutes, so people don't want to get bored learning how to interact. The more the interface is transparent, i.e. allows to easily finding out what is going on behind the surface, the more the person will be able to learn how to interact intuitively.

Proper hints can be weaved in the narrative content in order to suggest interaction modalities. These hints can be explicit messages like 'try to touch me', but will more likely be subtle suggestions, like animated or static media objects inside the direct manipulation area of the subject, that is the space that the person can reach with his hands or feet, that will respond to touch immediately, suggesting more complex forms of interaction.

This relies on the concept of affordance. Weiser wrote: "An affordance is a relationship between an object in the world and the intentions, perceptions, and capabilities of a person. The side of a door that only pushes out affords this action by offering a flat pushplate. The idea of affordance, powerful as it is, tends to describe the surface of a design".



Imitation

The best way to learn how to use something is to see another one using it. In public spaces there is often a continuous flow of people, favouring this dynamics. Humans learn through imitation very well; it is an implicit training phase: while some persons wait their turn to enjoy the media space, they observe other people interacting with the system, and learn while enjoying the content presentation. Also the power of the 'theatrical' scene of the system with people using it is a factor to be taken in consideration: interaction should also be designed to be interesting to see from other people. At the same time, people using the system should feel at ease being watched from strangers, and space design should allow them to feel comfortable in this situation, since they don't know how to master the system, and could be scared from other people looking at them while trying (maybe with a lot of uncertainty) to interact.

Human capability to learn by examples is impressive. An example is worth a thousand words; imitation is the key to inform people about the possibilities of interaction with a system in a seductive and challenging way. Note that while the person is learning he is also discovering the content of the media space. Space should be designed in order to let audience enjoy the whole experience, thus allowing imitative learning.

Natural interfaces in public spaces are usually running 24 hours a day (or during opening times). The system evolves according to public's behavior, there is no start phase for each user; a person will find the interface as his predecessor left it. Consider the passive public waiting to interact: they will see an experience with continuity, and will be able (at their turn) to continue what they were looking at.

PART THREE: PRESENTATION

Style

The particular nature of the proposed class of systems suggests a break from the rules that govern standard graphical user interfaces. Traditional metaphors (e.g. scrollbars, windows, pointers, multiple views, buttons, icons, and toolbars) hardly find their place in media spaces that have to be aesthetically effective, somehow immersive, and coherent to human perception. Natural interfaces are closer to modern computer games interfaces than to operative systems' GUIs; the reason for this is that like games, these interfaces are dedicated to a restricted domain, and must be immersive: in order to leave people attention on content, functional elements can't be too invasive.

Digital media can present information in different sizes and shapes that can be arranged dynamically. If a system is able to sense and interpret contextual information, the presentation style can adapt to different situations: a large screen can communicate to dozens of persons that are far away or to a single person standing in front of it. Interfaces have to be seamless in space and time, have to make digital media behave like real objects and to enrich reality with digital information, have to be fluid, minimal and modeless, have to be direct. The next chapters highlight the foundational elements of a new class of (natural) interfaces. The presentation of information is inexorably linked to sensing and to the core intelligence of a system.

There is a huge gap between the real-life world, made of people, places, nature, objects that can be physically manipulated, and the digital world, made of bits, of characters, of electronic messages and data flows. The gap, for a perceptual point of view, lies mainly in the way people interact with the objects present in the two worlds. To reduce this gap, two opposite approaches can be taken into account: to add physical constraints to the digital world and to augment, enrich the physical world.

Constraints and augmentation

Physical objects obey to the laws of physics. Virtual objects don't. This is of course a plus in most cases, but not when real people access to digital information. For example, common interfaces are a sequence of visual breaks; objects can disappear and reappear in another far location of the screen one hundred times in a second, something that is hardly accepted as natural by human perception.

David Ungar wrote: "User interfaces are often based on static presentations—a series of displays, each showing a new state of the system. Typically, there is much design that goes into the details of these tableaux, but less thought is given to the transitions between them. Visual changes in the user interface are sudden and often unexpected, surprising users and forcing them to mentally step away from their task in order to grapple with understanding what is happening in the interface itself".

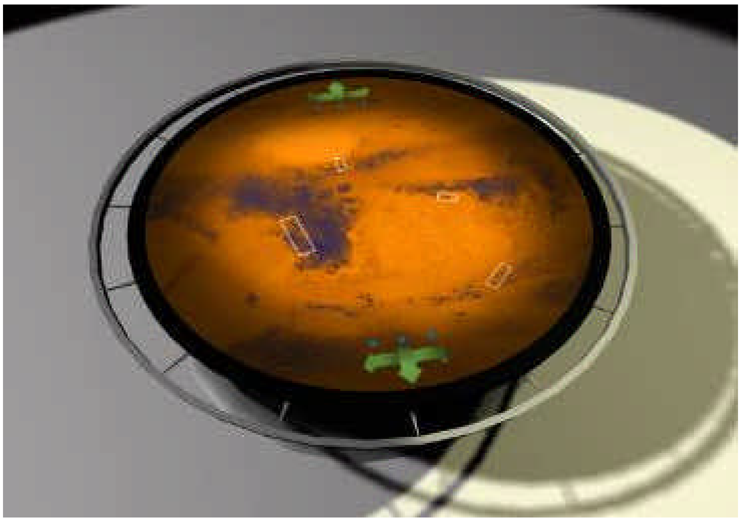
Natural interfaces introduce (simulated) physics constraints to control transformations of digital objects or pieces of information. This way, humans can track the changes and understand better what is going on. This can be achieved by assigning smooth motions and zooms (with accelerations and decelerations), without allowing objects to abruptly appear or disappear, or to penetrate thoroughly, or to cover each other. Zoom is an important tool to move between context and detail, and will be discussed later. The constraints proposed here are strict enough to assure the required result in terms of cognitive load, but still allow full expressiveness in terms of communication.

The opposite action consists in augmenting reality with virtual data. Augmentation is the integration of physical objects and environments and computational media. By means of video projections and directional speakers designers can visually overlay information and localize audio contents, thus enabling real objects and places to 'speak' to people. The encounter of bits and atoms can create new experiences that retain the best of both worlds.

Physics

Kinematics and dynamics mathematical models are a powerful tool for interface animation. The application of the desired constraints requires that objects are made solid, with a mass. This means that these objects will not move with linear velocity. Their speed will increase from zero to a certain amount, and then decrease to zero again, as any real object does. This is known as ‘slow in’ and ‘slow out’ effect. Fade in and fade out (on object’s transparency) should follow the same rule. Thus acceleration (with its consequences on velocity and position) is the first behavior considered from physics.

Ungar proposed to implement interfaces following criteria form cartoon animation: movements and dynamics are exaggerated in order to make them more comprehensible. So objects, before moving in a certain direction, move a little in the opposite direction, and vibrate slightly after a sudden stop. All these behaviors are based on energy issues. The same happens for the organization of objects based on certain semantics. Each object attracts and rejects other objects following gravitational forces, thus allowing a fluid and clear presentation that avoids overlapping. All these behaviors can be enabled in 2D or 3D scenarios.

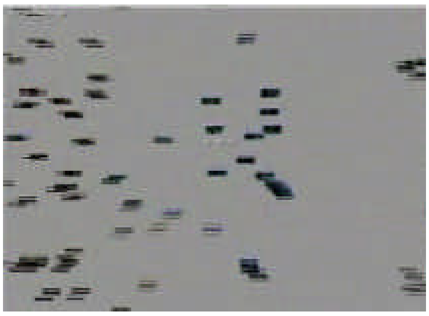


Spatial organization

Spatial information, i.e. the information represented by where an object is, is a powerful means of content organization. The hypertext navigation paradigm is based on an abstract series of jumps from one piece of information to another, with no spatial reference. All the relations between objects must be actively visualized in one's memory, increasing cognitive effort. Establishing a semantic relation between the meaning of data and its representation can bring relational visualization in front of the eyes of people.

To increase content accessibility, high level information should be represented by means of elementary cues; this way these can be processed by sensory or perceptual intelligence, making the subject aware of a variety of data and relations. Similar concepts are expected to be near, and hierarchical constraints are well communicated by a direct mapping on objects' displacements on the interface. Colours and shapes (features that are processed in parallel by human brain) are useful to represent conceptual grouping and classification. Size is a natural hint for importance; agitation is a hint for urgency. Objects may contain other objects, just as sections contain subsections.

Simple representation rules have to be chosen, and these must govern the whole information space, so that people can orient themselves. Bidimensional models, such as planar spaces, are the most used views; 3D models include isometric, weak and full perspective.



Seamless

Avoiding breaks, in space and time, is a key feature in a successful system. Hypertext and standard user interfaces propose a model that is a sequence of visual breaks (e.g. from a page to another one, from a dialog to the next one). The seamless nature of natural interfaces allows smooth navigation between different details, by completely removing visual breaks. Continuous motion is thus used to change point ov view or topic. Continuous zoom is a powerful medium to make people aware of the position of each detail in the whole context, and to allow navigation from detail to detail and from detail to higher contextual levels. Continuous panning is used to move the view window on large content spaces. Fade in and fade out represent the same concept in audio domains.

Commands in common interfaces are discrete (e.g. click, keystroke); in gesture based interfaces the event is fired either by particular mimics (e.g. by moving the finger forth and back) or by persistence (e.g. by keeping the pointing finger stuck for a certain amount of time). The latter solution is preferable, since it is more spontaneous; the problem is that the person must be made aware of what is going on while the persistence time needed to shot the event is passing, i.e. the interface must be transparent to progress information. The proposed solution consists in a fluid information layout that (perceptually) removes the concept of event. This leads to seamless behaviour in time. The transition between two states doesn’t happen abruptly, but becomes a continuous transformation in time: only at the end it will become irreversible, during the transition phase, if the person stops the gesture, the interface will continuously return to the original state. It is important that the person is precisely aware of the progress level of the transformation.

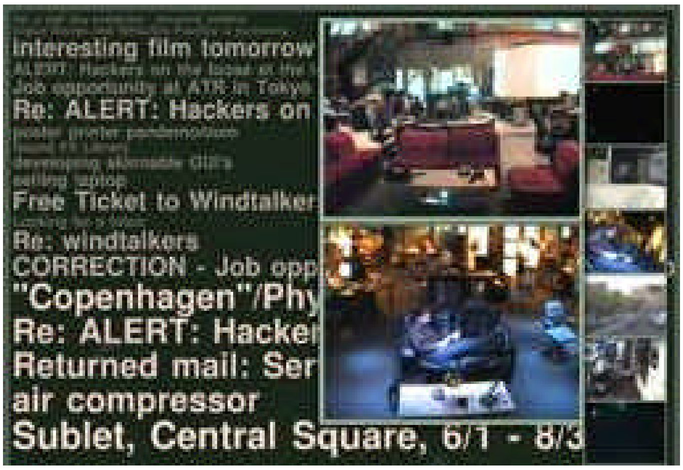
Of course these are not dogmas but rules of thumb; continuous behaviour in space and time (between context and detail, focus and periphery, pre command and post command) is useful to help data accessibility. Traditional interfaces always put rigid space divisions; natural interfaces propose fluid, flexible divisions, governed by mathematical models.

Less is more

Content presentation should be as clean as possible; a minimal presentation style leads to easier exploitation of the interface. The best solution is thus single view, with the manipulation space coinciding with the media space; reduction of the graphical elements, fonts, colours; full screen; appearance separation between content and functional elements, and between different kinds of information; correct splitting of content between different channels (e.g. video, pictures, superimposed text, voice, sound). All these factors influence the cognitive load of the subject.

Natural interfaces have to be minimal also in abstraction. There should be no mediation between person and experience, action and object, meaning and means. Direct manipulation of (media) objects should be preserved. Icons and symbols should be substituted with the object itself or its simplest representation (e.g. thumbnail pictures). Media space and manipulation space should coincide. Directly. This is a core concept in natural interaction.

Interface should always show a unique face. People should always expect what is in front of them, in order to have the sense that the situation is under their control. For this reason natural interfaces are modeless, i.e. their behaviour does not depend on modalities or running tasks; interruptions are not allowed. A seamless transition should be provided for any couple of modalities requested, thus making a continuous variety of modalities.



interesting film tomorrow
ALERT: Hackers on the loose at the
Job opportunity at ATR in Tokyo
Re: ALERT: Hackers on
poster printer pandemonium
selling laptop
Free Ticket to Windtalker
Re: windtalkers
CORRECTION - Job opp
"Copenhagen"/Phy
Re: ALERT: Hacke
Returned mail: Ser
air compressor
Sublet, Central Square, 6/1 - 8/3

Media

An accurate use of each kind of media is needed in order to orchestrate a successful communication. 2D can be used to show bidimensional content (e.g. paintings); photos, videos, images, drawings can be arranged and moved on screen to obtain the desired effect. Interposition of different layers (using color keying and dissolve effects) is a powerful solution to integrate visual data; zoom and pan effects can effectively express changes and motion from context to detail. 3D can be used to show objects or environments that don't exist, and depending on the level of interactivity requested, it can be prerendered or real-time. Prerendered 3D is technically a movie, and can reach impressive levels of realism, such as cinematographic special effects; prerendered shots have to be arranged so that the passage between them is seamless. Rendering can also be performed in real-time, thus providing full interactivity: the aspect of each frame is affected by input data from users; the problem is that the quality of the rendering is lower, due to the necessity to work with a number of primitives that the machine can manage in real-time, and to the lack of powerful authoring tools (e.g. Maya can't be used efficiently to author content for real-time engines).

Audio information can enhance or substitute visual experience; recorded voice messages can allow the system to actually speak to people. Synthesized speech is still insufficient to give people an illusion of naturality. Functional sounds provide feedback about recognized actions, thus releasing user attention on his current task (e.g. selection, dragging). Audification is the process through which an abstract value is represented by sound properties, and can inform about the values of specific variables. A successful audification is easily understood by untrained people. Auditory icons are sounds taken from the real world that are used to mark warnings, events, and activated functions. Sounds can be arranged in soundscapes, where it is possible to perceive background and foreground elements. Instead of presenting information through a single sense, it is better to separate the content, so that some is presented visually and some auditorily; this way learning is enhanced. This is true also for different types of visual (e.g. text, graphics, movies, and 3D) and audio (e.g. voice, music, auditory icons) media. Separating different levels of information and coding these into different clear representations is the best way to convey meaning.

Tangible

To allow people to naturally interact with the system, it is necessary not to introduce devices that the person must wear or use in order to be understood by the computer. For example, instrumented gloves and head mounted displays introduce unnatural mediation between people and media. In some cases, however, physical tools and objects can favor interaction. Tangible interfaces propose graspable objects as input devices, providing a mapping between such objects and digital media or meaning and enabling tactile sense exploitation and spatial reasoning.

Grasping a rigid body provides six degrees of freedom input, and more objects can give control on representation of complex spatial relationships. More people can simultaneously manipulate the media space with both hands, a very useful means of control for complex real-time simulations. Abstract entities can be represented by physical ones, thus enabling efficient cognitive schemes. Tangible tools strongly express affordances; suggesting and guiding action (see Intuition). Ishii and Ullmer wrote: “Our vision is […] about awakening richly-afforded physical objects, instruments, surfaces, and spaces to computational mediation, borrowing perhaps more from the physical forms of the pre-computer age than the present”.

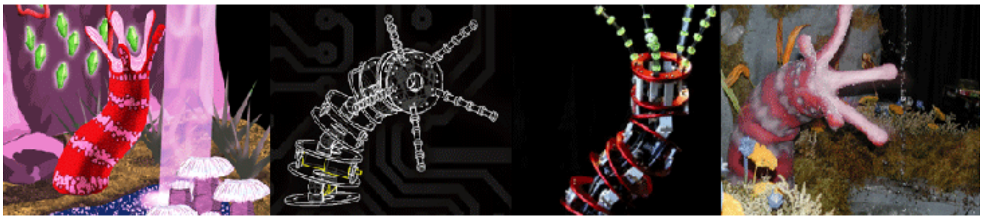
Ishii proposed the concept of phicons (physical icons), physical objects that have a meaning in the digital work. Data and functions become so graspable, and can be computed by manipulating the phicons that refer to them. A wooden ring can thus become a magic lens on a map screen, and a shop product, placed on a desk, can be surrounded by projections that tell everything about its properties and attributes.

Tools and characters

How will the interface look like? The role the interface plays defines the level of interaction it induces. Tool-like interfaces (those acting like a common tool and those acting like a multimedia application) will induce operational behaviors, with an emphasis on manipulation and command (the purpose of a tool can be various: access to information, media object manipulation, measurement and computation, etcetera); character based interfaces will induce social behaviors, emphasizing non verbal (like gesticulation) and verbal communication.

Characters can be represented by real-time or prerendered 3D animated models of realistic, abstract or cartoon-like creatures, real video shots or 2D graphics. Digital beings open new possibilities to interaction designers: the author tested a virtual aquarium, displayed on a large wall screen, with two fishes. The first was shy, while the second was confident and curious. Consequentially, motion algorithms and animations were implemented, and the effect on the public was unexpected: these simple behaviors created the illusion that the fishes were much more intelligent.

Characters can be either anthropomorphic or not, and either organic or not. This influences people attitude with them, as the reader can imagine. The purpose of characters is to make the machine more similar to people, even on its external representation. A character-like interface is generally harder to code than a tool-like one: much simulation is necessary for the creation of digital beings, the more evolved is the life form, the more complex and various the algorithms have to be.



Magic

As the system disappears in the environment, and it shows it is able to understand human expressions, a sense of amazement arises in the public (a visitor once told me “it’s a kind of magic!” after she had tried an interactive space). The technological aspect should not be exasperated (it is just a tool), it should not draw people attention: this would induce people to behave in strange ways. The contrast between a state of the art computer system and the ruins of a castle or abbey, (hidden) technology and tradition; this is the perfect mixture, maximizing impact on people.

As the system behaves in a social way, integrated in the context, the illusion of life and intelligence arises, making people think of it and interact with it as a living interface. The author is convinced that humans and animals are deeply different. Human is marked by something commonly referred as ‘soul’, making it unique and somehow inimitable. Animals are just complex natural machines, whose ability to adapt to circumstances is intelligence, an intelligence that can be reproduced. When robots, interfaces, machines, will achieve a comparable level of adaptation, the distinction between animal (natural) intelligence and artificial intelligence will become just a matter of source. Simple intelligent behaviors can convey to humans the illusion of life.

The magic dimension of natural interfaces is something that must be preserved. Showing the intelligence beneath an interface’s surface would prevent its goal: natural interaction.

Game

A playful dimension is always present in natural interaction. A completely unusual way to interact with a machine is a game by itself; the entertaining factor is the discovery of the behavior of the system by attempts, trying to obtain a desired reaction. The interface should be easy enough to be intuitive, but could be complex enough to hide some advanced features, that can be discovered by chance. The system has to answer to people's requests, but it can do it in unexpected ways, thus including an entertainment factor. Interface should always be new to peoples' eyes.

There is a class of interfaces that are based on gaming: entertainment systems. People can play the role of an ancient warrior and control a character using their own full body motion, or see themselves (actually) immersed in a (overlaid) world of monsters that they can smash or kick. More persons can join a virtual world and play together (maybe one against the other). A small jump can become ten meters high in the game (mixed) reality.

Games have the power of gain people attention so that they don't care about fatigue and don't get bored. Think about the possibilities in rehabilitation. One could stay on a stepper for an hour while virtually trying to walk in a wonderful desert. By delivering the right feedbacks, a system can enhance a player's control on his own body: it is sufficient that it encourages correct behaviors and motions by assuring higher scores or better control over the game character.

Human and language

Human factor is obviously the element at the center of the proposed framework. Human behaviour modelling allows dynamic tweaking of interaction depending on the sensed data. Modelling of body kinetics allows detection and interpretation of purposeful and expressive motion. Such systems must respect humans, being able both to understand and to express through natural language. Too often interfaces impose languages that are coherent with the machine but require extensive adaptation from the person. The goal is not to substitute humans, but to provide them powerful tools to express and access information. This results in an enrichment of the person's capabilities, making complex information awareness and access direct and easy.

The intention of selecting a particular can be expressed by a person in many ways. For example, he could select it by pointing at it persistently, by moving his finger back and forth in that direction, by drawing a circle around it with his hand, by knocking the screen surface in that position, and he could reinforce his selection with voice. The system should be able to respond to any of these actions, by analyzing sensed inputs in parallel to detect the various modalities. This conveys to the public a strong sense of freedom, since allows them to express as they are used to. In this sense, multimodality is more the capability to deal with the richness and variety of human expression than the combined processing of multiple input channels.

All the expressions of a person (verbal or not) constitute a language, a human language not only for communication, but also for discovery and use. Most interfaces ask the user to learn the machine language, and in some contexts it is even a plus: learning a new language is a factor of enjoyment. However in natural interfaces it is the machine that learns human language. The attention must be drawn on content, the enjoyment aspect is always present, since people are not used to machines that understand common gestures.

Narration

Most systems in public spaces don't have to simply show digital media. Their purpose is to communicate a meaning about that media. They have to narrate some kind of content during time, depending on people intentions and behaviors. Glorianna Davenport wrote: "Over the centuries, stories have moved from the physical environment (around campfires and on the stage), to the printed page, then to movie, television, and computer screens. Today, using […] sensing technologies, story creators are able to bring digital stories back into our physical environment".

As people come around the system, it should motivate them to begin interaction. The graphics, audio or physical layout of the installation could be sufficient to assure this; otherwise, the system itself could directly address the visitors, asking them to play with him. After active interaction has started, the system should be aware of people level of attention, and act consequently (e.g. visual attention is fundamental to narrative dynamics, and can be estimated by people movements; another measure of attention is the classification of sound generated by the public).

Of course a system must first of all respond to user commands; the risk is that it could make interaction mechanical and discontinued. The critical point is the passage between narrative elements. Narration is segmented into small media pieces that can be arranged in a variety of ways, respecting some constraints. A stop command should not be executed abruptly, but a feedback must be provided immediately. Narration continuity can also be ensured by means of transition elements that cover the breaks between different chunks, both in audio and video.

Dialogue

Media must adapt to people different interests: a person could be very interested in details, another could want just a general idea about a subject; the problem is that it would require additional controls in order to tweak the communication. A solution proposed by the author is represented by pyramidal texts. As information is requested, the system presents it in 15 seconds; then it goes into deeper and deeper detail, without repetitions. This way, the user will be able to stop the explanation as he gets bored, going back to the context.

As sensing capabilities and context awareness grow, narration becomes dialogue. People commands are not only voluntary actions, but also unconscious behaviors. The system gets a certain level of proactivity (that actually is just reactivity to a wider range of inputs, with memory). The system could address people as they get near, or modulate communication depending on feedback and attention. Communication with an user could allow parallel communication to newcomers, in order to provide them context information and welcome.

Consider a documentary film: it shows one hour of information about a given subject. A person will probably be attracted by ten minutes of content. The solution is not to provide menus to access particular information, but to build high level semantic organizations that provide multiple paths over content, allowing to establish links between distant objects. Each piece of media should be presented in the overall context, showing semantic links that freely change depending on people behaviors. A documentary film, a map viewer, a photo browser, a computer game and a virtual character represent the boundaries of an area in which a new form of (natural) interface can develop.

PART FOUR: APPLICATIONS

Website and workshop

Many ideas, projects and prototypes relevant in some aspects to the issues exposed in this text arise from the research community. Often these items come from groups interested in different fields, and are not perceived as elements of a unique path; speech understanding researchers refer their work to the speech recognition community, researchers in computer vision, psychologists and interface designers do the same; but all of them can boost research in natural interfaces.

To reinforce the concept of natural interaction, and to help the creation of a community of people interested in this topic, a website was created. The website naturalinteraction.org is structured as a collection of resources that anyone can contribute; each resource is accompanied by a title, a short description and a link to the web. The best way to get the point about a concept is to see it (even partially) implemented; the website, with its hundreds of references, is thus a good starting point to explore the world of natural interaction.

On April the 2nd, 2004, an international workshop was held in Florence, Italy. Speakers from Europe and the US and 160 people could meet about 'Natural Interaction'. Invited speaker Flavia Sparacino opened the meeting with a lecture about her works. The author (who also served as Associate Chair) could present his prototypes of natural interfaces, ranging from perceptual spaces to visualization and ambient displays. It was a great chance of confrontation with well known researchers and practitioners. There was a large participation of young researchers and enterprises.

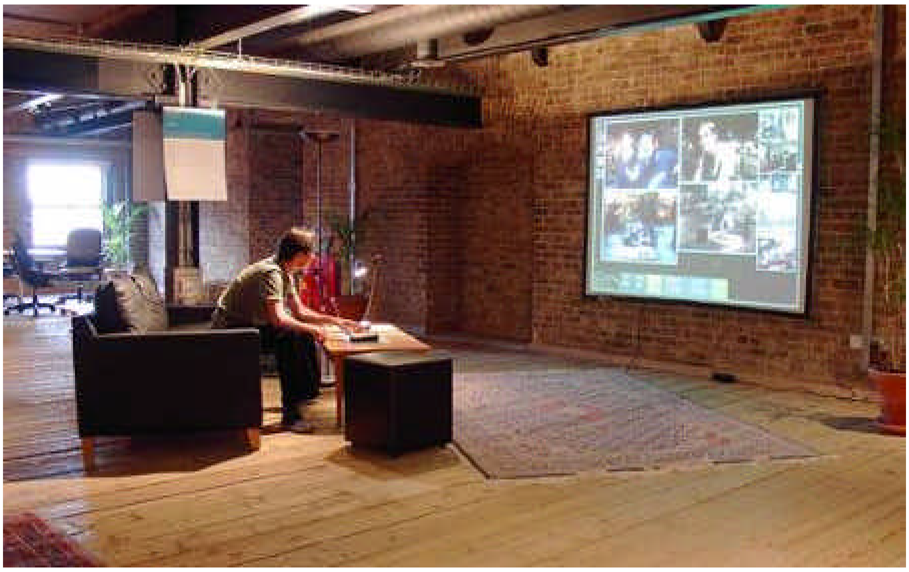
Workshop on Natural Interaction

APRIL 2 | FLORENCE, ITALY

Research

As natural interaction is the result of the contamination among different disciplines, research in such a field must be taken on following non traditional ways. Standard research methodologies fail, bringing no significant contributions. Prof. Neil Gershenfeld wrote: "I found that one of the best predictors of a student's success working this way was their grades: I look to make sure they have a few F's. Students with perfect grades almost always don't work out, because it means they've spent their time meticulously follow classroom instructions that are absent in the rest of the world. Students with A's and F's have a much better record, because they're able to do good work, and also set priorities for themselves. They're the ones most able to pose - and solve - problems that go far beyond anything I might assign to them".

The author's lab doesn't look like an office or a common computer science lab. It is full of computers, but it is also full of lamps, glasses, mirrors, stands, cables and balls, tools to model wood and metal, scissors, paint. No need to say that academic mentality is often sceptical about such mixed approaches. In such a discipline, the right questions are much more important than the answers. It is time for optimism, risk, fantasy, creativity; it is time to go beyond division between disciplines: something largely extraneous to the academic world.



Observing people

To find out how humans spontaneously interact with digital media a good test is to prepare a simple installation with no real functionalities, like a large vertical screen or a floor projection, and populate it with media objects (videos, images) or physical objects (balls, wands). Inviting common people to interact with the setup (children and elders are the best candidates) and observing what they do is the best way to get ideas for the design phase. They will try to activate movies, to push virtual buttons, to enlarge photos, to scroll through a stripe, with a not so large vocabulary of gestures and expressions, and they will provide unpredictable ideas about what they would expect from an interface of that kind, what behaviours they would expect from each media piece. The design phase should follow these indications strictly, because these are the key to a successful natural interaction.

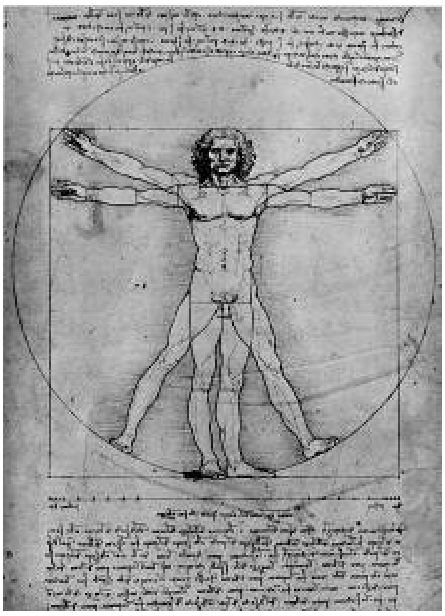
The same practice should be used after the realization of a prototype, when the system works (in the lab and with the creators), to test again that what has been implemented is a good work. Observing how humans act, perceive, discover, communicate is a unique source of useful information. The goal is to copy such interaction (another good test is to observe human to human interaction), not to invent abstract languages to do the same things. At a first look, this could seem a limitation of the expressive potential of a system; on the contrary, being able to communicate with humans at a so natural and immediate level opens new possibilities of engagement, relation, and vehiculation of messages and contents.

A good practice consists in adding logging code to installations. It is easy to detect relevant events, such as presence detection, strong changes in environment appearance, selections, long hesitations… All these events can be documented storing date, time, variables values, and images from the cameras at given instance. All this data can also be sent automatically to remote e-mail addresses or be used to generate specific alarms.

Personalization and perfectionism

Due to the artistic nature of natural interaction systems, it is always necessary to personalize the solution depending on the context. The major variables that require this are the integration in the surrounding physical space and the type of communication wanted. Natural interaction is a matter of details. A single detail can induce a negative interaction experience (e.g. if a system requires the user to stand in a precise position, this has to be marked clearly, maybe with a coloured circular platform slightly above the floor level). Even a wrong stylistic choice in the architectural setup may deeply reduce engagement in the public.

Institutions and privates commissioning a system often tend to be quite rough in providing indications. They think a sensing technology can be integrated into another interface and be hosted in a space designed by a third architect. This is not always true, and often results in a strong reduction of the appeal of the overall solution. Only who masters all the aspects that influence interaction can guarantee a successful solution. Setup is an art, coding is an art, interaction (or experience) design is an art: personalization requires perfectionism, the perfectionist attitude of who masters the whole technological and artistic background.



Public spaces and events

Institutional communication in public spaces has to be effective and attractive. A system must occupy part of a large environment and must attract people and convey a message. New interfaces, that can be used by everyone (even those that have never touched a computer), are a powerful means of communication. A lot of content can be accessed in a short time and selectively by a large public. Interface design must take care of both the active public and the passive one. Public events require immediate communication that can be provided by these interfaces, where the user becomes part of a performance that engages the whole public.

Setup is an art. Setup solution has to be aesthetically interesting and robust, allowing a reliable functioning. Apart from public contexts, even domestic environments represent successful settings for natural interfaces. Diffused commercial solutions include computer vision (in mixed reality) to control game consoles. The public is impressed by the novelty of an interface, and is thus induced to appreciate better the content.

Museums and exhibitions

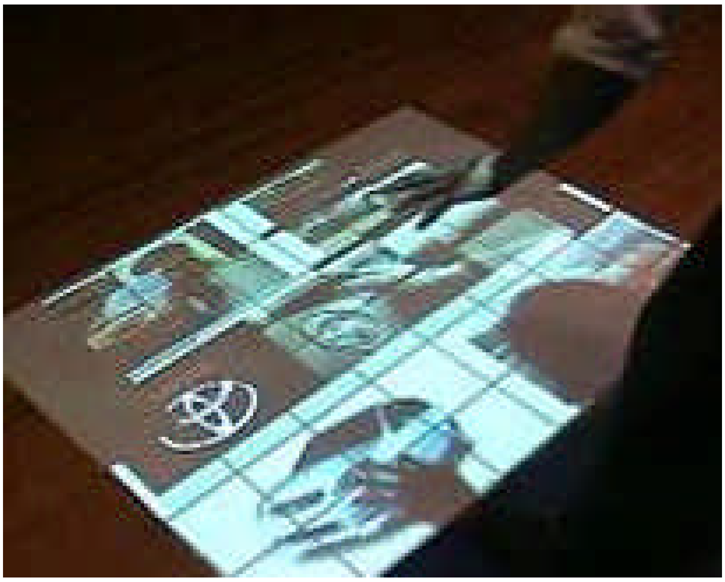
Common museums and exhibitions don't offer compelling experiences to those that are not experts in the particular topic presented. Through natural interfaces, sculptures, paintings and photos can dialogue with people, and personalized experiences can be provided to the public. Computers in museums are often used as web or presentation browsers, a level of interaction that is completely different from the rest of the visit. New interfaces that are integrated in the overall exhibition architecture can allow the exploitation of the expressiveness of digital media in the whole visiting experience.

Museums are now thus the best context for natural interfaces. The contrast between ancient places and high technology that is hidden in the environment is a plus for such systems. Cultural heritage material can thus be enriched while being left free of any technological element. Welcome staff plays a crucial role on the satisfaction of visitors; they should suggest interaction rules when needed, but should also leave people free to try and enjoy the experience.

Showrooms and advertising

Each advertising message is meant to be read from a certain distance (e.g. the back cover of a newspaper usually hosts advertising that is meant to be read from about two meters). Digital media allows to adapt visualization to the context (e.g. multiple people, someone near and someone far away); the same could happen on content. A natural interface in a shop could attract visitors and interactively show them products and services; a desk could provide information about the products deployed on it. The effect on the public is impressive, and attention on single data can be logged.

Advertising solutions include large projections that interactively show product features, billboards that react to people standing in front of them, and banners that can speak to invite people to watch them.



Theme parks and entertainment

The same observations about museums and exhibitions apply to theme parks, places where entertainment is the rule to follow and experiences are generally fast-paced, emotional and not very analytic. Computer vision allows robust installations, since there are no functional mechanical parts that can be broken. Single and multiple users can face remote or virtual players in innovative games, mixed reality and real-time compositing can offer entertainment modalities based on people live images, such as merging the user silhouette and virtual worlds, inhabited by digital data and characters.

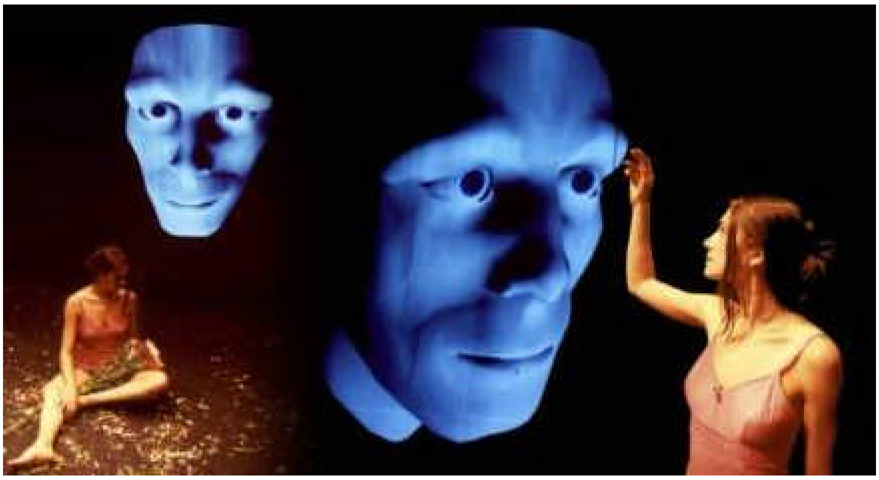
Players can control virtual beings through body movements, causing the character to perform complex motions and interact with the rest of the virtual world. The same can be applied also to driving vehicles and exploring media spaces.

Interactive art and performances

Art is a product of human expressivity. A traditional art piece is subject to passive admiration of the public. Interactive artefacts and spaces enable the additional factor of public's expression and action to the artistic medium. The work of art can change and communicate reacting to people behaviour and feedback. Artistic space becomes alive by means of projected light, sensors and custom software. Physical sculptures can react to touch and voice, visual art can change its appearance over time and react to presence and gestures.

These possibilities not only impact the public. Artists are provided with a new generation of tools that can enhance their message. As an example, many solutions have been proposed for the creation of music by means of gestures or disposition of objects on dedicated media tables; music composition thus becomes an opportunity for everyone, since a simpler mapping between sound properties and their representation is proposed through metaphors.

Even live performances such as theatre and dance are enriched, since media content can be controlled in real-time by performer's expressions. Projections on stage can visually surround the performers, magnifying their expressions and gestures; moreover, performers can interact with digital actors, whose behaviour is controlled by context data.



Remotely

Shared environments allow reducing the distance between different places. Online virtual spaces overcome the problem of spatial remoteness; these systems must assure symmetry, reciprocity and consistency to convey a sense of togetherness to the different users. Two questions arise: How to render the remote user? How to render the environment? The more the view is realistic and coherent, the more the system is convincing; that's the role of natural interfaces in remote collaboration. A screen could represent a window looking at the remote place, or a glass between the two places where media can be displayed and manipulated. The most successful system is the one that transparently maps the two environments together.

Another use of networked natural interfaces is just a provocation: what is or what could be a natural internet? First of all, natural internet should have a strong reference to real places, instead of abstract concepts, spatial content organizations (note: not just one) and interface design not related to printed paper but to real world imagery. People should perceive the other net surfers around them, and should be able to receive media even when they are not actively requesting data; a picture in a living room could show a live Polynesian landscape, with superimposed news and contents. And what about a naturally interactive operating system?

Traditional internet is a useful tool to prolong, extend an interactive experience attended in a public space; once at home, people can deepen their knowledge about the content proposed by the system, at the event website.

Tables and walls

A table that is also a large computer screen is a useful solution for many reasons. It offers a reference for direct manipulation of physical and digital objects (gestures in free 3D space, apart from the simplest ones, like pointing, are an unnatural form of communication); physical objects can be deployed and moved on the screen surface; multiple users can share display space, and media can be presented along any direction (people can walk freely around the table).

The aim of an interactive wall is to provide a big visualization. A wall section that is also a large display is a low cost unencumbered immersive solution. The system can be able to deal with people presence nearby the wall, touch of the wall surface, and pointing gestures from a distance. Multiple wall sections can be added, to cover large areas.

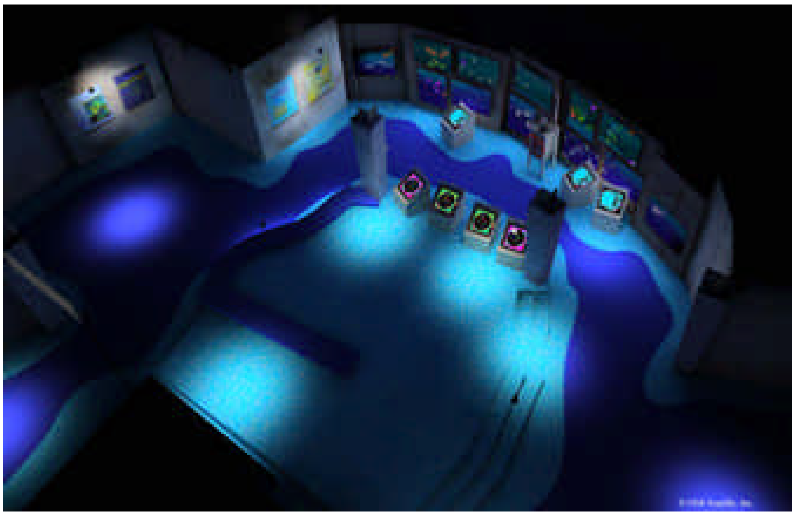
Display solutions range from LCD screens to plasma displays, to video projectors, the latter being the less expensive and the most versatile solution. There are two main factors about projectors: focal length and luminosity. The focal length depends on the lens, and fixes the distance between projector and screen necessary to obtain a certain diagonal. Installations outdoor or in places with much light require more powerful projectors. Screens for front projection are usually white and opaque. Screens for rear projection are partially transparent and can be made in a variety of materials, from glass to plastics.

Screen or display proportions are usually 4:3 and 16:9, the latter being useful for immersive content if the visualized area is very big. Other screen shapes can be obtained from the rectangle: for a circular table with a diameter of 3 meters a display area of 4 x 3 meters is needed, and the undesired regions will be black coloured. Path from projector to screen can be made longer in a small space by using mirrors. Special mirrors, called ‘first surface mirrors’ are needed, since common mirrors, that have the silvered surface separated from the first glass surface, create diffraction effects that reduce projection quality.

Floors and rooms

Digital media projections on floors represent a powerful interface layout. People can walk on the interface, stepping on active areas and using pointing gestures to activate functions. The metaphor is particularly addressed to contexts in which a spatial or geographical information organization is needed. A map projected on a floor can allow visitors to walk over Europe or India, people can use their feet to discover hidden objects or kick digital beings. Large projections can cover several meters, creating effective immersive mediascapes. System hardware, such as computers, projectors and cameras, can be placed out of reach, on the ceiling, preventing malfunctions derived from physical contact. First surface mirrors can be used to adjust projector beams and camera views; infrared illuminators and lenses are used in order to segment people from the background.

In order to sense large environments, hyperbolic mirrors (or a simple silvered light bulb) can be used to make wide camera views. Distortion problems are solved by using simple mapping functions on extracted features. Visitors can be tracked in order to personalize media presentation and interaction. The author tested a solution to track multiple people in a medium (8 x 8 meters) with a common digital cameras, providing position and orientation information. People collisions on image plane are solved through prediction and memory implementation.

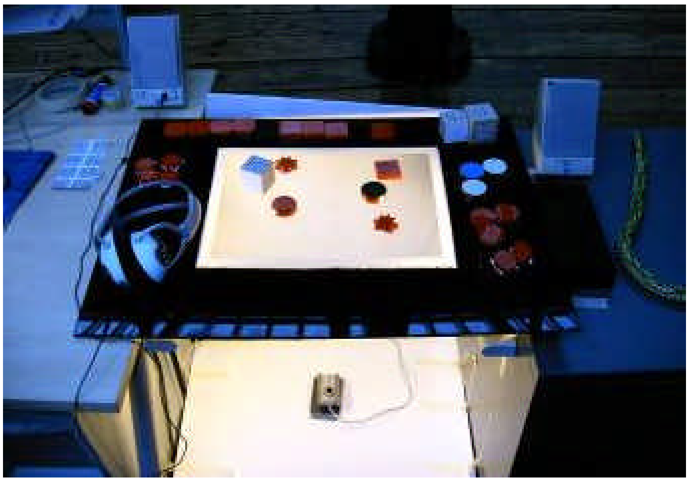


Artifacts

Graphical algorithms allow to adapt projected images to every non planar surface, thus enabling to enrich almost any kind of object or space with digital media (other algorithms can adjust colour properties). Projectors can so visualize information in the real world, near the television, on a sofa, on a tablecloth. Specific physical objects can be designed in order to take life through projected light.

Consider a scale model of a building, white coloured and cut so that a surface represents a inside view; the model could be animated with inhabitants, animals, wall textures, furniture. This way the informative value of a tangible model is added to the value of live digital media, that can make something static alive. At the same time interface presentation and control objects could be visualized near the user, in a position easy to see and to reach. Integration with auditory media is obviously simpler.

The link between a physical reality and related interactive media content is a powerful means of communication, and allows the creation of unusual and impressive solutions. The possibility to use all the physical space as a medium to convey information is the maximum in terms of freedom, since people can manipulate the whole space and what it contains.



Projects

Here are briefly reported the various prototypes the author created in order to push towards the goal of natural interaction. Each of these projects is interesting for a particular aspect. Natural interfaces should integrate all these particulars into a single experience. Videos and additional information is available on the author's http://alessandrovalli.com website.

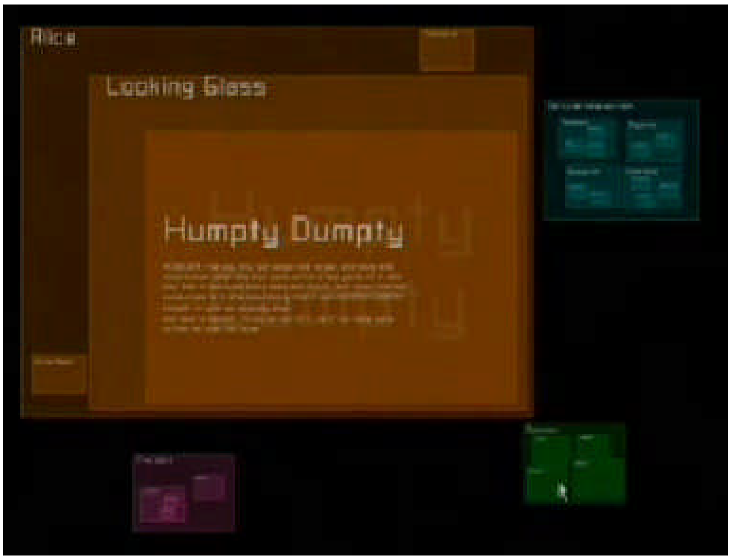
PointAt
An adaptive and robust hand pointing system. Two of such systems have been installed in Palazzo Medici Riccardi, the third museum in Florence, in Lorenzo De Medici's bedroom. Users interact with natural deictic gestures with digital reproductions of the beautiful frescos by Benozzo Gozzoli, hosted in the museum.

tangiTable
A tangible interface for children. Players can interact with maps projected on the table moving wooden physical objects, such as cylindrical markers and round lenses. It is a very robust collaborative tangible playground. More tables can be connected remotely through the internet.

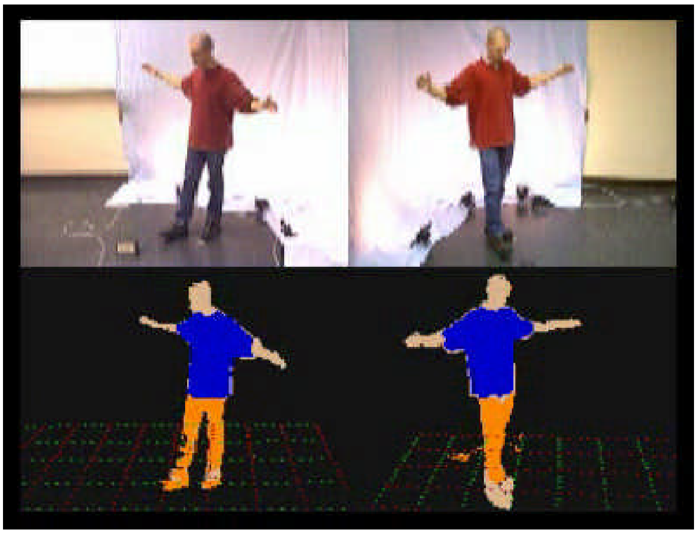
minorityMap
Interaction with scenes and objects visualized on a large rear projection display or a multi-viewpoint autostereoscopic plasma display. Interaction is based on user gestures and voice commands. This system has been hosted in Palazzo Vecchio, Florence, showing the changes that the city is going through (new train station, new underground tracks…).

golem
A real-time full body motion tracking and humanoid animation system. Users don't need to wear any special device, and can animate 3D cartoon characters live. The system is MPEG-4 compliant.

voxbox
A real-time desktop voxelization and broadcast system. The application builds a voxel representation of the user hands and tools and sends it to local or remote computers for real-time 3D visualization. It provides interaction with physical and virtual objects in a seamless way.



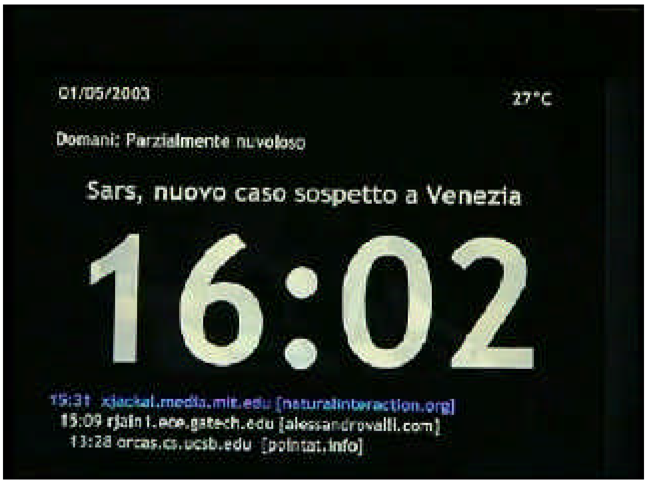
zooi
A context preserving zoomable user interface. This flash application uses non homogeneous zooming in order to show details and context at the same time optimizing the screen surface. It uses a 2D space browsing paradigm instead of the common hypertext paradigm.



golem 2
A full unencumbered body tracker based on a model matching approach in order to allow tracking of a large number of degrees of freedom. A collaborative shared virtual environment based on these systems has been tested too.

ygdrasil vision
Interfacing a vision module to the shared networked environment system ygdrasil to provide natural 3D navigation through deictic gestures and body motions. Exposed in Palazzo Medici Riccardi and Palazzo Vecchio, Florence.

iconoclastBody
Controlling an arcade computer game with intuitive gestures. The game is inspired by Marcel Duchamp. An untrained user can play walking around on the stage and performing smashing gestures.



01/05/2003 27°C

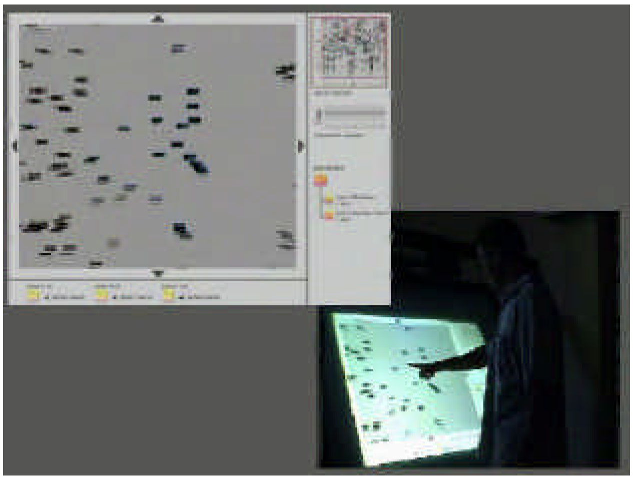
Domani: Parzialmente nuvoloso

Sars, nuovo caso sospetto a Venezia

16:02

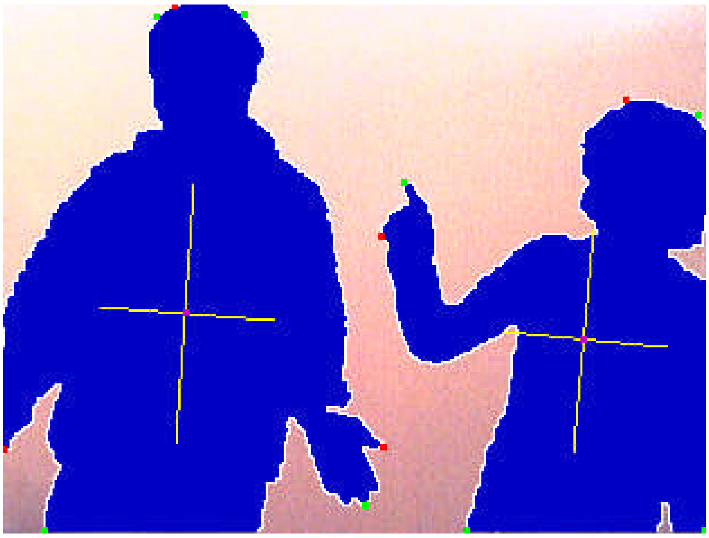
15:31 xjackal.media.mit.edu [naturalinteraction.org]
15:09 rjain1.ece.gatech.edu [alessandrovalli.com]
13:28 orcas.cs.ucsb.edu [pointat.info]

ambientClock
An experiment about ambient displays and peripheral perception. An ordinary wall clock that gives non critical information such as websites accesses, news and weather forecasts associated with sound icons. The system was on 24 hours a day for some months in the lab, and provided useful information when needed, while not disturbing and interrupting the researchers.

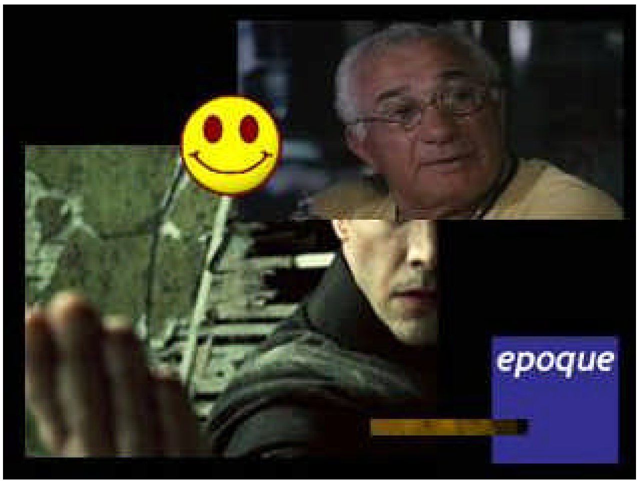


visualBrowser
An automatically segmented and annotated movie or a collection of images can be browsed using a multiresolution interface that arranges thumbnails on screen in a number of ways depending on user needs. An active search paradigm in an ordered universe is used instead of the common query paradigm. The vision system is based on near infrared light.

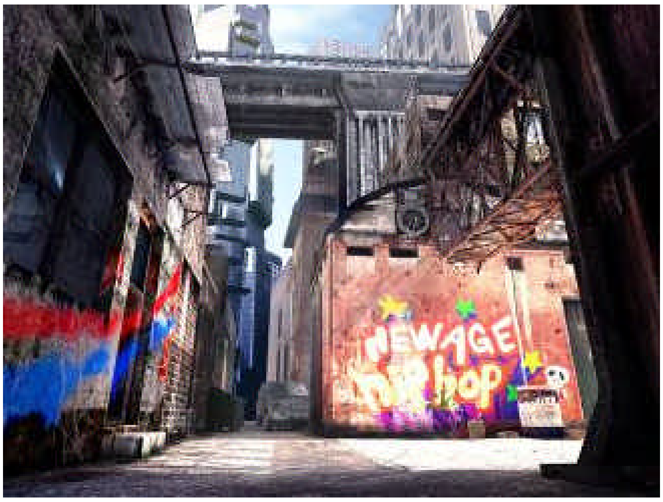
flySwatter
In this installation, exposed in Palazzo delle Papesse museum, Florence, multiple users could smash digital flies projected on the walls using real flyswatters. The flies represented powerful men in history.



Retina
A lightweight, robust computer vision module for interactive natural interfaces. Performs early processing and sends features through sockets or MIDI to other applications (C++, Java, Flash MX, MAX/MSP...). Downloadable from the website.



Epoque
A high performance 2D audio video library for C++. It can be easily interfaced to Retina, and allows to build video based natural interfaces in a very short time. Downloadable from the website.



Carmen
A high performance 3D library, based on the Irrlicht engine. It can be easily interfaced to Retina, and allows to build 3D based natural interfaces in a very short time.

Acknowledgements

The author wish to thank gratefully Prof. Alberto Del Bimbo for allowing this research activity; Flavia Sparacino and all former researchers of the Vision and Modeling Group at MIT Media Lab, for their inspiration; Prof. Stan Sclaroff from Boston University, for his encouragement and his suggestions; all the speakers and the participants to the Workshop on Natural Interaction, for that great day; all the researchers at Media Integration and Communication Center, for their patience and support.

Conclusion

Innovative and comprehensive research visions are rare stuff. Researchers often struggle to solve very specific problems, focusing on minimal enhancements of existing approaches and methods. Fast and continuous development of new technologies provides a variety of expressive means that is largely unexplored. It is important to investigate the meaning and goal of such technological possibilities, in order to make them impact in human life and to drive innovation. The author is persuaded that the topic of natural interaction is fundamental to this end. This text is a tentative to depict an approach by means of a series of glances to different aspects; this should give the reader a comprehensive idea of the whole: sensing or presentation cannot be taken away from the whole framework, due to the strong mutual functional influences and relations.

Natural interaction systems are literally exploding on the market, due to the enormous impact on the audience and to their seductive nature. Open issues include a higher level of computer awareness of the context in which interaction takes place, a deeper understanding of the role of affordances and embodiment in interaction with digital media, a serious investigation of the relations between the single components of a system, eventually seen as a whole. As far as new sensing technologies and algorithms will become available, additional work will be needed in order to give new expressive and relational capabilities to media spaces. The author is working on these subjects, creating prototypes to demonstrate the validity of a unique framework.

A. Valli

References

The references of this text are proposed in a slightly unusual way. Web resources are provided instead of bibliographical ones, and a short description is provided, in order to allow faster consultation.

Perceptive Media
An interdisciplinary initiative to combine multimedia display and machine perception to create useful, adaptive, responsive interfaces between people and technology.
http://www.cs.ucsb.edu/%7Emturk/Papers/PerceptiveMedia.pdf

Embodied Interaction
Embodiment reflects both a physical presence in the world and a social embedding in a web of practices and purposes. The outline of a new foundation for the design and analysis of interactive systems is presented.
http://www.dourish.com/embodied/embodied99.pdf

Spatial Computing
Spatial computing is human interaction with a machine in which the machine retains and manipulates referents to real objects and spaces. It is an essential component for making our machines fuller partners in our work and play.
http://acg.media.mit.edu/people/simong/thesis/SpatialComputing.pdf

Interactive art and entertainment installations
This paper presents a brief summary of body tracking tools and interfaces, and explains how they have been applied to a variety of interactive art and entertainment projects.
http://web.media.mit.edu/%7Eflavia/Papers/Flavia_isea2000.pdf

From Cartoons to the User Interface
User interfaces are often based on static presentations, a model ill suited for conveying change. Events on the screen frequently startle and confuse users. Cartoon animation, in contrast, is exceedingly successful at engaging its audience.
http://research.sun.com/research/techrep/1995/smli_tr-95-33.pdf

DataTiles

Tagged transparent tiles are used as modular construction units. These tiles are augmented by dynamic graphical information when they are placed on a sensor-enhanced flat panel display.
http://www.csl.sony.co.jp/person/rekimoto/datatile/

Jeremiah

Jeremiah is based around two subsystems, a graphics system which constitutes the head and a vision system which allows him to see. There is also a simple, built in emotion engine which allows him to respond to visual stimulus via expressions or emotions.
http://www.ee.surrey.ac.uk/Personal/R.Bowden/jeremiah/jeremiah.html

Leonardo

Leonardo is the Stradivarius of expressive robots. It is our challenge to give Leonardo a computational brain that is worthy of its body.
http://robotic.media.mit.edu/projects/Leonardo/Leo-intro.html

Public Anemone

It is a robotic creature with an organic appearance and natural quality of movement. It interacts with the audience by orienting to their movements using a stereo machine vision system. But if you get too close, it recoils like a rattlesnake.
http://robotic.media.mit.edu/projects/anemone/robot.html

Mixed Reality Pong

The players can play the game with their hands or other real-world objects. The game physics simulate the behavior of a real ball, except that the virtual ball doesn't slow down at all. The computer is completely hidden in Mixed Reality Pong. No specially marked objects are required, so any objects with enough contrast to the background can be used in the game.
http://www.mlab.uiah.fi/%7Ekkallio/mr-pong/

Museum on the Resistance

A table is divided into two halves by a series of vertical screens. By passing their hands over the surface of the table, the visitors can flick through a collection of stock footage as if it were a virtual book on the subject.
http://www.studioazzurro.com/opere/sarzana/index.htm

Reactrix

A visual display system that responds to people instantly and in real-time with dramatic visual effects and entertaining gameplay, to deliver an engaging experience people not only remember, but seek out.
http://www.reactrix.com/

reacTable*

The reacTable* is a novel electronic music instrument with a tangible user interface. The goal of the project is the creation of a state-of-the-art interactive music instrument. It is collaborative, intuitive, sonically challenging and interesting, totally controllable.
http://www.iua.upf.es/mtg/reacTable/

ToneTable

ToneTable is a sound and computer graphics installation which enables up to four people to collaborate on exploring varied dynamical relationships between media.
http://www.shape-dc.org/articles/pdf/DAFX-01.pdf

Open Window

An ambient virtual window for bolstering wellness and healing potential during a hospital stay.
http://www.medialabeurope.org/hc/projects/openwindow/

EventScope Table

The medium we chose to pursue is a table that is front projected from above. The interaction is done by an infrared stylus that is tracked by a camera.
http://www.etc.cmu.edu/projects/eventscope/goals.html

Stomping Ground

It is a permanent installation consisting of a musical carpet and a projection of live video with superimposed blobs.
http://acg.media.mit.edu/people/simong/stompingGround/index.html

Audiopad

It is a composition and performance instrument for electronic music which tracks the positions of objects on a tabletop surface and converts their motion into music.
http://web.media.mit.edu/%7Ejpatten/audiopad/

One2One

We are developing a technological infrastructure for creating personalized ambient communication links to enhance a sense of presence and togetherness between two distant individuals. We are exploring a variety of passive sensing and display devices to suit individual taste and the character of the relationship.
http://www.medialabeurope.org/hc/projects/one2one/

Wanderful Alcove

The magic wand presents an interesting design opportunity as a form for a tangible computer interface. In addition to exploring the technology needed to build a magic wand interface, this project focuses on role-immersion scenarios in which these interfaces can have a socially tranforming effect on their users.
http://www.medialabeurope.org/hc/projects/wanderfulalcove/

Virtual Space

The aim has been to create intuitive user interfaces that blend in with the surroundings of the user. The starting point of the project to research and develop bodily and spatial user interfaces has been the natural way in which people move and act.
http://www.vtt.fi/tte/projects/lumetila

Digital Seed

The Digital Seed is a virtual alter-ego of a real seed, he lives in a cube. The physical actions on the cube affect the inner virtual world where the seed lives and grows.
http://www.mle.ie/%7Emauroc/digitalseed/

UrineControl

The system uses computation to enhance the act of urination. Sensors in the back of a urinal detect the position of impact of a stream of urine, enabling the user to play interactive games on a screen mounted above the urinal.
http://www.monzy.org/urinecontrol/

Audience Interaction

A variety of techniques that enabled members of an audience to participate, either cooperatively or competitively, in shared entertainment experiences. These techniques allow audiences with hundreds of people to control onscreen activity.
http://www.monzy.org/audience/

Interactive Virtual Aerobics Trainer

This system creates a personalized aerobics session for the user and displays the resulting interactive virtual instruction on a TV screen. Here the user can choose which moves (and for how long), which music, and which instructor are desired for the workout.
http://www.cis.ohio-state.edu/CVL/Research/VirtualAerobics/aerobics.html

Calm Technology

Information technology is more often the enemy of calm. Technologies encalm as they empower our periphery.
http://sandbox.xerox.com/weiser/calmtech/calmtech.htm

EyePliances

Appliances and devices that detect and respond to human visual attention using eye contact sensors. EyePliances receive implicit input from users, in the form of eye gaze, and respond by opening communication channels.
http://www.hml.queensu.ca/papers/p550-Shell-CHI2003.pdf

HoloWall

The HoloWall is a wall-sized computer display that allows users to interact without special pointing devices. The display part consists of a glass wall with rear-projection sheet behind it. A video projector behind the wall displays images on the wall.
http://www.csl.sony.co.jp/person/rekimoto/holowall/

The Virtual FishTank

Twelve large projection screens form windows into a spectacular undersea world, populated by nearly 100 bold-colored, cartoon-like, mechanical fishes. Visitors use computers to simulate the movements of lots of fish and then to explore the kinds of patterns that emerge from the interactions of the fishes.
http://www.mos.org/exhibits/current_exhibits/virtualfishtank/vft_walkthrough.html

Installation

The system consists of a window through which the scene is viewed and a stylus with which objects are manipulated. The window is a flat panel display with a tiny camera mounted on the back showing a live video image of the room as seen through the window.
http://acg.media.mit.edu/people/simong/installationNew/intro.html

Zooming User Interfaces

ZUIs are based on the premise that navigation in an information space is best supported by tapping into our natural spatial and geographic way of thinking. The information space is represented by an infinite two-dimensional plane.
http://media.humboldt.edu/~continuum/rashmiweb/zui.html

Fisheye Menus

We are investigating techniques to support selection of an item from a long linear list. The primary technique we are looking at is the application of fisheye views to linear lists. Live demo.
http://www.cs.umd.edu/hcil/fisheyemenu/

Magic Lenses

Toolglass widgets are new user interface tools that can appear, as though on a transparent sheet of glass, between an application and a traditional cursor. These widgets may incorporate visual filters, that modify the presentation of application objects to reveal hidden information, to enhance data of interest, or to suppress distracting information.
http://www2.parc.com/istl/projects/MagicLenses/doc/TGMLSiggraph93.ps

VisiPhone

VisiPhone is a communication object that opens a graphical as well as an audio portal through space. It is designed to provide a continuous, ubiquitous connection between people in different places.
http://web.media.mit.edu/%7Ekkarahal/projects/visiphone/

iCom

It is a media installation that forms a bridge between different locations. It operates in a continuous and background mode, integrated with the surrounding space. The portal enables awareness of remote activity and promotes a sense of connection among those generating it.
http://www.medialabeurope.org/%7Estefan/hc/projects/icom/

Exertion Interfaces

An Exertion Interface is an interface that deliberately requires intense physical effort. We designed, developed, and evaluated an Exertion Interface that allows people who are miles apart to play a physically exhausting ball game together.
http://www.exertioninterfaces.com/

Gesture and Object Tracking for Augmented Desks

A spontaneous and unimpeded interface between the physical and virtual worlds. Objects are recognized and tracked when placed on the display surface.
http://www.gvu.gatech.edu/ccg/people/david/papers/starner-perceptive-mva02.pdf

Jam-O-World

This is a multi-user interactive musical device. Intuitive input devices with real-time computer graphics on a tabletop surface for collaborative gaming and music-making.
http://www.etc.cmu.edu/projects/jamoworld/index.htm

The Gesture Pendant

A wearable device for control of home automation systems via hand gestures. By combining other sources of context with the pendant we can reduce the number and complexity of gestures while maintaining functionality.
http://www.gvu.gatech.edu/ccg/publications/gesture_pendant.pdf

Roomservice, AI-style

Can a room be intelligent? Thoughts and work of four people who not only believe the answer is yes, but are working towards making this happen.
http://www.ai.mit.edu/people/mhcoen/ieee.pdf

Tangible Bits

Tangible Bits allows users to "grasp & manipulate" bits in the center of users’ attention by coupling the bits with everyday physical objects and architectural surfaces.
http://tangible.media.mit.edu/papers/Tangible_Bits_CHI97/Tangible_Bits_CHI97.pdf

SenseTable

A system which tracks the positions of intelligent objects on a tabletop surface, and projects information onto the objects themselves.
http://tangible.media.mit.edu/papers/SenseTable_CHI01/SenseTable_CHI01.pdf

Narrative Spaces

Interactive narrative spaces. These spaces are orchestrated such that people's presence and movement drives the presentation of digital media.
http://xenia.media.mit.edu/~flavia/Papers/NarrativeSpaces.pdf

Computer Vision for Interactive Computer Graphics

Technical report from MERL. Computers looking through a camera at people is a potentially powerful technique to facilitate human-computer interaction. The computer can interpret the user's movements, gestures, and glances.
http://www.merl.com/papers/docs/TR99-02.pdf

Attentive Toys

We describe an attentive system that pay attention to people so they can attend to people's needs using visual and audio sensors. We implemented it as a visually interactive toy robot.
http://www.umiacs.umd.edu/users/yaser/haritaoglu.pdf

Anthropos Project

The Anthropos project explores the fusion of AI research both as a tool for investigating human machine interaction and human cognition as well as a technological arena inspired by human cognition. The objective is to understand how to explore the role of anthropomorphism in HMI and balance issues of mechanistic vs. humanlike capabilities.
http://anthropos.mle.ie/

The Computer for the 21st Century

The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.
http://www.ubiq.com/hypertext/weiser/SciAmDraft3.html

Towards Digital Experience

By Ramesh Jain. Experience is fundamental to human existence. As a result, the desire to share experiences motivates the development of exciting technology. The time has come to define and design the digital experience.
http://jain.faculty.gatech.edu/media_vision/Towards_experience1.pdf

ChatCircles2

ChatCircles2 is a chatroom that uses graphics in a different way than most current graphical chatrooms. Instead of having pictorial avatars, participants in ChatCircles2 use color and geometric form to convey social presence and activity.
http://chatcircles.media.mit.edu/

The KidsRoom

The KidsRoom was a fully-automated, interactive narrative playspace for children. Using images, lighting, sound, and computer vision action recognition technology, a child's bedroom was transformed into an unusual world for fantasy play.
http://vismod.www.media.mit.edu/vismod/demos/kidsroom/kidsroom.html

Computer Vision Games Using A Cheap Webcam

This paper presents a game suite which has been developed as an example of vision interaction between a computer and a human.
http://www.ece.nus.edu.sg/stfpage/eleks/ICARCV%272000.pdf

SmartDesk

SmartDesk is a project of the Perceptual Computing group at the MIT Media Lab and encompasses experimentation on a range of computer-based perceptual input and output systems in a personal work environment.
http://www-white.media.mit.edu/vismod/demos/smartdesk/

DreamSpace

The DreamSpace allows users to collaborate in a shared space. The system "hears" users' voice commands and "sees" their gestures and body positions. Interactions are natural, more like human-to-human interactions.
http://www.research.ibm.com/natural/dreamspace/

When Things Start to Think

This book is by Neil Gershenfeld, from MIT Media Lab.An important story about why and how computers will disappear, when and where your things will think.
http://www.media.mit.edu/physics/publications/books/ba/

Body Mnemonics

Body mnemonics is a meta tool for portable devices that enhances their usability and makes them more responsive to our cultural background on the basis of three principles: proprioseptic sense, our outlook on our own bodies, and the "method of loci" mnemonic device.
http://www.mle.ie/~jussi/projects/body_mnemonics/index.php

Pfinder

The Person Finder (Pfinder) combined a traditional segmentation framework with sound classification theory. The result was a system that provided solid segmentation, plus a model of the body that yielded more information about the human than classical figure-ground segmentation alone.
http://www.cs.ucsb.edu/%7Ecs281b/papers/Wren.pdf

Facilitator Room

The facilitator room project is an attempt to observe, model, and affect the interaction patterns of its users. This involves sensing the user's motions and sounds using computer vision and audition technologies, and then interacting with them using active components in the room.
http://whitechapel.media.mit.edu/facilitator/introduction.html

MagicBoard

The MagicBoard project aims at augmenting a perfectly ordinary whiteboard-like surface with electronic capabilities, via a video projector and a pan/tilt/zoom camera.
http://iihm.imag.fr/demos/magicboard

Rubella

The idea with Rubella is to make people aware of their bodies and the way they lead their lives. We visualize the lack of exercise by reflecting it on an artefact, the dog Rubella.
http://www.rubella.tk/

KidStory

Tangible technologies were developed to support room sized collaboration amongst groups. Groups of children could interact using a magic carpet for navigation and bar coded and tagged objects to insert story elements (pictures and sounds).
http://www.virart.nott.ac.uk/Projects_Kidstory.htm

Nebula

Nebula is an interactive projection system designed to enrich the experience of going to bed, sleeping and waking up. It provides intuitive and natural ways of physically participating in a virtual experience, through simple body movements and gestures.
http://www.design.philips.com/smartconnections/nebula/

Interactive Tapestry

Most of today's interior decorations in home environments are static. It is not possible to interact with and change them in real-time. Our thought is that it should be possible to create a form of informal interactive art that is easily accessible in ones homes.
http://outrun.idc.cs.chalmers.se/%7Eit3chje/uc/Projekt/index.html