Università degli studi di Cagliari

Dottorato di Ricerca in Matematica e Calcolo Scientifico

Ciclo XXIV - Settore scientifico-disciplinare MIUR: INF/01

Gestures and Cooperation

Considering non-verbal communication in the design of interactive spaces

Presentata da: Alessandro Soro
Coordinatore Dottorato: Prof. Giuseppina D'Ambra
Tutor/Relatore: Prof. Riccardo Scateni

February 26, 2012

University of Cagliari

PhD School of Mathematics
and Scientific Computing

Gestures and Cooperation

**Considering non-verbal communication
in the design of interactive spaces**

Author:
Alessandro Soro

Supervisor:
Prof. Riccardo Scateni

February 26, 2012

Abstract

This dissertation explores the role of gestures in computer supported collaboration. People make extensive use of non-verbal forms of communication when they interact with each other in everyday life: of these, gestures are relatively easy to observe and quantify. However, the role of gestures in human computer interaction so far has been focused mainly on using conventional signs like *visible commands*, rather than on exploiting all nuances of such *natural* human skill.

We propose a perspective on natural interaction that builds on recent advances in tangible interaction, embodiment and computer supported collaborative work. We consider the social and cognitive aspects of gestures and manipulations to support our claim of a primacy of tangible and multi-touch interfaces, and describe our experiences focused on assessing the suitability of such interface paradigms to traditional application scenarios.

We describe our design and prototype of an interactive space for group-work, in which natural interfaces, such as tangible user interfaces and multi-touch screens, are deployed so as to foster and encourage collaboration. We show that these interfaces can lead to an improvement in performances and that such improvements appear related to an increase of the gestures performed by the users. We also describe the progress on the state of the art that have been necessary to implement such tools on commodity hardware and deploy them in a relatively uncontrolled environment.

Finally, we discuss our findings and frame them in the broader context of embodied interaction, drawing useful implications for interactions design, with emphasis on how to enhance the activity of people in their workplace, home, school, etc. supported in their individual and collaborative tasks by natural interfaces.

Foreword

During the preparation of this manuscript I happened to attend a conference on HCI with a friend whose specialization was computational geometry. He was astonished because none of the presentations contained either a triangle or a mathematical formula. The research for mathematical abstractions as an aid for thinking begun quite early in the history of mankind, but it wasn't until the 16th century (see for example [19]) that the abstract (algebraic) representation of geometrical concepts was recognized a superiority that today is still largely accepted.

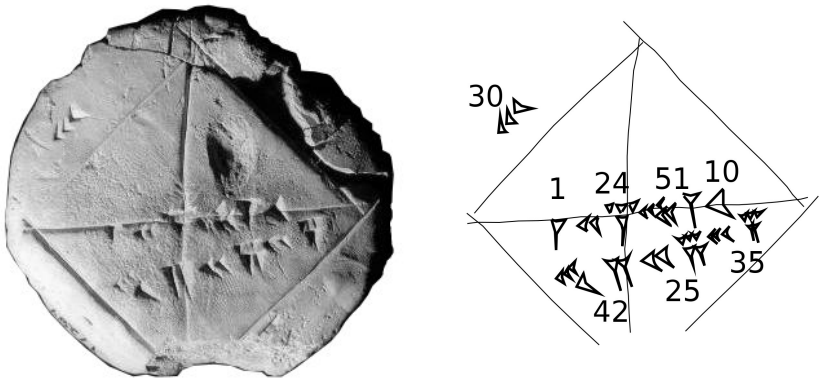


Figure 1: The YBC-7289 tablet (left) and a transcription of the mathematical cuneiform text (right)

However, at the same time, it is generally known that those same abstractions often require quick and dirty concretizations in order to be understood. For example, one of the first known written evidences of the pythagorean theorem is the YBC-7289 tablet[1] (see figure 1). It is a Babylonian school tablet, dated approximately 1800-1600 BCE (that is, 12-10 centuries *before* Pythagoras' birth), supposedly carved by a trainee scribe with the drawing of a square, complete of its diagonals, and numbers; the size and shape of the tablet, more or less round and 8cm in diameter, makes it perfect to fit into one's hand [58].

What caught the attention of mathematicians towards tablet YBC-7289 was the numbers carved along the diagonal. For those who don't read ancient Babylonian[2] the sexagesimal number $1;24,51,10$ is equivalent to $1+\frac{24}{60}+\frac{51}{60^2}+\frac{10}{60^3}$. That is to say 1.41421296, a surprisingly good approximation of the number $\sqrt{2}$,

[1] This tablet shows a Babylonian approximation to the square root of 2 in the context of Pythagoras' Theorem for an isosceles triangle. The original tablet is part of the Yale Babylonian Collection. Photo: Bill Casselman [31] (image source: Wikipedia, the image has been edited to remove text overlays).

[2] ...such as myself, the translation is from [31].

exact to the sixth decimal place. Note that the interpretation of old babylonian numbers is not unique and not at all easy if taken out of context. The other numbers on the tablet represent the length of the side (3 left-headed cuneiform symbols meaning 30 that in this case should be read $\frac{30}{60} = \frac{1}{2}$) and the length of the diagonal: $42, 25, 35 = 0 + \frac{42}{60} + \frac{25}{60^2} + \frac{35}{60^3} = 0.70710648 \approx \frac{\sqrt{2}}{2}$.

According to [58] YBC-7289 was the exercise notebook of a novice, as it is revealed by the unusually large handwriting. In the 35 centuries that have passed since then, lots of practices have changed, or disappeared at all. But I can't help imagining that young scribe, holding his (clay) tablet in one hand and a stylus in the other, racking his brains over a math problem. Does it remind of anything familiar? One other thing is outstanding here: 35 centuries ago a young trainee scribe bothered drawing a triangle (a concrete representation) to put numbers (the abstraction) into context, and this was useful not only to himself, helping to tame an otherwise difficult task, but also to us, that today are able to infer so much about his motivations and goals.

Thus, the rest of this work is about concrete representations, i.e. models, artifacts, behaviors and activities that can be viewed (I mean properly, not in one's mind's eye), manipulated, smelled; the next sections describe the journey that, starting from multiuser touch interfaces, leads to the design of computer assisted collaborative activities, informed by the observation and the analysis of human gesturing, as a visible, and only apparently unstructured, form of human communication.

Brisbane. 4 October 2011.

Contents

List of Figures

> History teaches the continuity of the development of science. We know that every age has its own problems, which the following age either solves or casts aside as profitless and replaces by new ones.
>
> David Hilbert

Part I

Background and Motivations

Chapter 1

Introduction

The dream of being able to operate computers just like a team leader can instruct the group is as old as the computer itself. In 1943, when he was a guest at AT&T lab, Alan Turing was heard saying:

> "No, I'm not interested in developing a powerful brain. All I'm after is just a mediocre brain, something like the President of the American Telephone and Telegraph Company." [77]

We can't say what Turing would think of modern interactive computers. Did expert systems or neural networks get even a bit close to his ideas? What about machine learning, computer vision, pattern recognition, would they look *intelligent* if seen through the spectacles of the mid' twentieth century scientist?

Anyway, in general, we have the sense that computers are far from being even comparable to a mediocre brain, and are not really getting any closer, although this is not necessarily a problem.

One of the goals when trying to create *intelligent* systems was to cope with human failures. Back in 1973, in a paper quite representative of such school of thought Wasserman suggests some design guidelines for creating *idiot-proof* interactive programs [182]; today we may smile at the triviality of recommendations such as these:

> Provide a program action for every possible type of user input;
> Allow the user to express the same message in more than one way;

if not for the fact that the underlying principle sounds quite sinister, and reminds of more than one masterpiece of science fiction: to create systems:

> designed to anticipate any possible action by its users and to respond in such a manner as to minimize the chances of program or system failure while shielding the user from the effects of such a failure. An idiot-proof

> program will continue to perform "intelligently" no matter what its users do. [182]

The nonsense of such principles is underlined by Bannon:

> a consequence of trying to make such a system is that an incredible amount of "intelligence" must go into its initial design and maintenance. Taken to the extreme, we have the prospect of artificially intelligent systems operated by morons, an absurd scenario. [8]

Fortunately, the evolution of interactive systems has taken a different path, one that leads to the enhancement and exploitation of human abilities, rather than to jailing users within a sandbox.

1.1 HCI reloaded

The history of interactive systems can thus be re-read in terms of what human abilities are exploited when operating the computer, what is the focus of the interface, from hardware to work setting, and what tools were used to evaluate the system, from chronometers to ethnography (see for example Grudin [68] and Dourish [46]).

The first computers were essentially *pure hardware*, meant to be operated by specialists, ideally those same ones that designed and assembled the machine. A step forward came with the introduction of some symbolisms, as with the first memory stored programs, and then some linguistic properties, as with the first programming languages.

In 1963 Sutherland presented the Sketchpad [174] opening the path to visual computing and graphical user interfaces, i.e an interaction paradigm that lets people exploit visual memory, the ability of spatial organization and visual metaphors. The principal 'users' of the system were no longer scientists or engineers, and certainly not the same ones that designed the application. Ergonomics and usability become central topics in the design of interactive applications.

The rest of the story, up to the present day, is about the exploitation of more and more natural abilities, such as manipulative and social ones, with focus on the inextricable relation that binds perception, cognition and motor action. Such relation is often referred to as *embodiment* (see for example [46, 51, 164]), and is mostly the focus of this work, and the tools of such new interaction paradigm are often referred as *natural interfaces*, and are the subject of sparkling research and sometimes heated debate.

A few points are worth remarking here; in the first place, every interaction paradigm that has been suggested has *wrapped* rather than *replaced* the preceding ones, and all the above interaction techniques are still available, though

differently accessible, in modern computers. As a consequence, the designer of new interfaces is more and more aware of how people feel, behave, work, learn, etc. in their everyday activities. That is, the design of new technologies starts from an observation of people expectations and goals, rather than from an analysis of problems and requirements.

Despite this, a thorough understanding of what characteristics define a *natural interface* is yet to come. If some authors point to *familiarity* [159], others pick out *predictability* [78]. These remarks may appear obvious: how could an interface be other than natural? What can we say of an interface *based on existing skills* [123], given that one of the skills of human beings is the ability to acquire new skills (and many consider it *fun*, even). And anyway they uncover the difficulties of designing the interaction with the complex environments in which we live, without the comfort of simple (maybe simplistic) guidelines and principles, such as the aforementioned *idiot-proof rules*.

Another interesting aspect of *naturalness* is that it is not necessarily a property of the *interface*, i.e. of the artefact, but is rather a feature of the *interaction*, that is, of the activity. For example, in human-human communication, a person that wants to attract the attention on a given subject has several options: she can name the subject verbally, nod with her head towards it, point with a finger, etc. Each one of these actions is, somehow, *natural*, and each one can be recognized by state of the art sensing devices. Yet we have the feeling that no interface based on *one* of the above would be more natural than, say, using a mouse or a stylus.

On the other hand, the possibility to choose the appropriate communication channel, together with the ability to mix several channels to reinforce the message and to move with continuity between different communication strategies, is much closer to the naive idea of a *natural* interaction.

In other words, in our view a natural interface is the one that allows people to stay (or at least *to feel* that they are) in control, by simulating, reproducing, or (even better) being part of the procedures and practices of everyday physical and social interaction. That is, the design of natural interaction with computers should start from the observation and understanding of how people interact with their natural environment and with each other.

In this sense, *natural interfaces* support collaboration, and throughout this work we will refer to *natural interaction* as the activity of people in their workplace, home, school, etc. supported in their individual and collaborative tasks by such specially designed artefacts.

1.2 Hands and Cognition

At this point a question could legitimately arise: an interface based on such natural behaviours would be suited to cope with abstract, for example mathematical,

thinking? After all, the symbolisms of algebra were initially conceived just to overcome the limits of verbal descriptions and geometrical representations [19].

This question is still open, and contributing to the advance of this field is one of the goals of this work. However there is strong evidence towards a positive answer. For example, we know that people make an extensive use of gesturing and non-verbal forms of communication when cooperating with others, both to express thoughts and emotions [50, 116, 121], which is hardly a surprise, though the mechanisms that regulate such behaviours can be very complex.

Perhaps less intuitive is the fact that gesturing helps people reasoning about specific problems, such as mathematical procedures [63], improves learning in children [36], improves spatial reasoning [49], etc. That is: not only people perform better when they have more chances for gesticulation (and thus for communication), but in otherwise identical conditions they perform better when encouraged to gesture. And subjects with more fluency in gestures score the better performances in tests. This should at least suggest to reconsider the design of office workstations to make them more *gesticulation-friendly*.

A similar role on cognition has been observed regarding manipulations. People use their hands to change the state of the world as well as to explore such state. In doing so it is possible to recognize either an *epistemic* or a *pragmatic* action. If the latter follow a predetermined plan or strategy, the former seems more related to the visual or tactile exploration of the possible alternatives (imagine reorganizing one's cards in a card game).

The reasons of such behaviour have been studied and explained by Kirsh and Maglio [113]: it helps moving the complexity of the task from one's head to the world, available strategies and possible solutions to a given problem appear at a glance. Additionally the (limited) resources of attention and memory are not wasted to concentrate on the strategy and can be used to explore alternative solutions. In practice such exploration, performed by means of manipulations on the world (or tools), is easier (less cognitive effort) and faster (less time) than it would be if performed mentally.

What is surprising here is that epistemic action (i.e. the explorative, unplanned, apparently *useless* manipulation) increases with skill [128] in computer programs. That is, the expert user doesn't head straight for the solution, but instead uses his/her abilities to explore multiple possibilities, and to do so makes a better use of epistemic action, when possible. Unfortunately not all computer interfaces are designed to exploit such attitude: tangible user interfaces [93] are so; WIMP interfaces, a bit less [150].

Manipulative interfaces have a long research background. Seminal work on multitouch screens dates back to the early eighties (e.g. [26,41,101,119,124,187]). and up to the recent explosion of interest on this topic (e.g. [71, 72, 76, 149, 190, 197]). Also tangible user interfaces (TUIs) build on the principle of direct manipulation of hybrid physical/digital entities (e.g. [6, 53–56, 93, 150, 179]). Manipulative interfaces have several peculiarities with respect to more traditional interaction techniques. The interaction takes place in the physical space: the

user is not projected onto the screen, but instead digital entities acquire a *touchable* or even *graspable* consistency.

Note, finally, that gestures and epistemic manipulation seem to share an overlapping area. Those *representational* gestures that people perform to indicate, mimic, or handle imaginary object appear analogous in their cognitive function to their physical counterparts [82].

1.3 Focus on Gestures

Studies on gestural interaction, on the other hand, have mainly gone towards the recognition and classification of arbitrary and definite sets of gestures (e.g. [17, 125, 126, 138, 151, 163, 173, 177, 188, 193, 195, 196]). Most applications presented in literature assume a certain vocabulary of gestures to be used as a command set in order to directly control a computer system. This is not a surprise since the way we interact with computers is mostly *narrative* rather than *conversational*, as Wexelblat points out in [188], but fail to consider a number of possible applications, and, most important, the cognitive role of gesturing discussed above is not given great emphasis:

> Finally in this category, there is the question of private versus public meanings. We all "know"that gesture has both private (for the self) and public (for others) functions.
> [...]
> If we could understand what the private uses of gesture were, we might be able to separate them out. Our systems might be programmed to ignore the private parts, or to make use of them, but either way it would be useful to know. [188]

This work is based on a different point of view. In the first place, *gestural interaction* doesn't assume necessarily *gesture recognition*. If the latter is about algorithms and error rates, the former encompasses the practices, usages, meaning and origins of gesturing. One example is to focus on gestures and postures to evaluate whether or not a certain technology is disrupting cooperation in a workplace scenario [139]. No matter how powerful our (desktop) computers become, it is clear that their shape and size is not evolving towards improving interpersonal relations. It is not rare to observe people use instant messaging to communicate within the same room.

Since gestures are relatively easy to observe and measure, they can provide an objective index of cooperation, useful for example to compare different design choices. One experiment described further on in this work builds on this concept to compare the *surface* and *desktop* versions of a collaborative application. It turns out that surface computing can outperform desktop in an *office automation* scenario. Moreover, a number of potential applications can stem from a deeper understanding of how gesturing relates to emotions and excitement, both at a

communicative (i.e. intentional), and at an informative level (i.e. what others perceive from our gesturing, independently of our intention to throw a message).

1.4 Dissertation Structure

The next sections will return in more detail on the topics that this introduction has only sketched out. Chapter 2 will declare precisely the scope and goal of this work, and enumerate more clearly the original contribution to the state of the art. Chapter 3 will describe several application scenarios.

Part 2 (chapters 4, 5 and 6) will describe the current state of the art, both from a technological and a psychological/anthropological viewpoint. Part 3 (chapters 7, 8, 9 and 10) describe the applications that have been prototyped and the experiments that have been performed to sustain the theses presented here.

Finally, Part 4 draws some conclusions and points to future research.

Chapter 2

Scope and Goals

This work aims at exploring the role of gestures in computer supported cooperative work. As such, it focusses in the first place on understanding how people use gestures when they cooperate (or compete) face to face using *at the same time* a digital tool. Common examples of this context may include sitting side by side browsing the Web, playing videogames with friends, looking for a flight on the *arrivals* panel together with a travel mate, etc.

Note that such scenarios are not easy to spot in daily life: our computers are, for the most part, only too *personal* to be used in cooperation; smart-phones, PDAs, etc., although they foster cooperation *at a distance* aren't of much help when people stand face to face[1]. This sounds particularly ironical: the computer industry has designed complex systems to cope with relatively specialized and not ubiquitous problems, such as spreadsheets and databases, but we still rely on paper for a task such as sketching, the base of communication and creative thinking.

Thus, the starting point of this work, that has been briefly addressed in chapter 1 is the observation of how people use gestures and manipulations when:

1. they interact with their environment to either make sense of it or to change it
2. they interact with each other
3. they cope with abstract (e.g. mathematical) thinking

what can we learn from the above? If such phenomena make a real difference, i.e. can be observed (if not measured) even in real world settings, what can we say of the present design of computer systems? Specifically:

[1] Of course cooperation happens: people sit together, browse the web in groups, share the use of devices every day. But in a sense they do it *in spite of* the design, rather than *thaks to* it.

1. what can we learn from the way people use to gesture that can improve the design of interactive systems?

2. what new application fields can we devise for *natural interfaces* in general, and to take advantage of the natural inclination of humans towards gesticulation in particular?

3. can we say something of how *natural interfaces* can impact traditional application fields?

2.1 Research Design

Despite (or because of) the above research questions being essentially interaction design issues, I've learnt from previous experience that a lot of hardware/software design would have been required in order to experiment with gestures and manipulations in a realistic (if not real) setting. Also, I considered carefully both the constraints and the opportunities offered by the workplace, in terms of physical space, hardware and software availability, opportunities for feedback, testing and cooperation. They are briefly described here.

The workplace is the Open Mediacenter Lab, situated withing the CRS4 building in the Science Park of Sardinia (Italy). Most of the work described has been carried on in cooperation with the lab's people, and the final prototypes were hosted in a space within the lab precinct, about $4 * 4$ mt. wide. This area is part of a larger open space, which partly limits the nature of the possible activities, but also fosters cooperation and informal feedback.

The hardware available for running software prototypes and building sensors and devices included several high end workstations, 2 short throw (wide-angle lens) video projectors, other high resolution video projectors, high resolution network cameras, high end webcams, Arduino prototyping boards and accessories, gigabit wired Ethernet and WiFi. Support and tools for some *craftsmanship*, such as soldering, cutting, assembling mockups, were also available.

External visits to the lab were planned on a regular basis, and included potential research and industry partners, representatives from the local administrations and government, students from primary and secondary schools. These gave great opportunities for informal feedback and passive or participant observation, that were exploited throughout the development of this work as will be discussed later. However, more structured testing in this context was not possibile due to privacy issues, ethical constraints and time availability. For example, attempts to organize research follow-ups to visits, administering a questionnaire to school children were unsuccessful.

Working on the field was possible and economical and logistic support was provided for participating to international exhibitions once a year. This was an excellent opportunity to test on the field both the prototypes and the research hypotheses. In practice, the participation to such an exhibition concentrates in one week all the visitors and all the technical, organizational and environmental issues that are typically spread in a whole year of lab life.

Formal testing was possible when colleagues or University students volunteered their participation. Such an opportunity is influenced by a combination of factors, such as matching deadlines, right *mood* in the lab (that, as said above, is a shared space), curiosity for the newly designed prototype, etc. As already mentioned, creating fully functional and engaging interactive prototypes ensure that the planning of testing is not left to serendipity alone, but at the same time demands a much greater effort.

The above reflect on the choice of approach and methodologies for this work, that balances between applied research, in the development of prototype, and lab experiments and field studies for evaluating and re-engineering the technology and for understanding and generalizing principles.

Note that the choice of what technology to implement was rather opportunistic: on the one hand there was the goal to foster participation of potential users. With respect to this, multitouch tables and walls are inviting and engaging, fun to use and to watch other people using. This choice of technology also encourages cooperation (several people can use the interactive wall or table at the same time) and exposes just that non-verbal communication that is the focus of this work.

On the other hand, the ultimate goal was, as said, to asses the suitability of natural interfaces to traditional application fields and to devise new applications for natural interaction. If passive observation and informal interviews may be suitable to inform the redesign of the technology, a quantitative result is desirable to give a positive answer to the former question. This has been obtained by means of a controlled laboratory experiment.

The latter goal, to devise new applications for natural interfaces, was accomplished (or better, a step in this rather unexplored scenario was moved) by exploiting the whole interactive space as a working demonstrator, in which to envision, enact and assemble new applications from existing interactive components, in cooperation with experts of confining disciplines. This approach is not too far from a common practice in HCI as described in [114], and is known to involve some issues[2]. In the first place the generalizability of lab experiments is problematic, even more in our case in which the choice of sample was opportunistic. Moreover, the observation that here are defined as *field*

[2]Though Kjeldskov and Graham examine practices in mobile HCI, their work fits well the case in exam.

work typically took place in the lab itself, though in an informal and otherwise unconstrained setting.

If in applied research the tool to develop is often the goal, and evaluation is targeted at improving the design, in this case the development of new technology (though in the end it contributed to the state of the art of multitouch devices) was not a goal per se. Instead the technology was meant to put in evidence a behaviour that I wanted to observe, acting in a sense like a catalyst.

Such behaviour (gesticulation in collaborative work) happens naturally in a workplace scenario such as the lab, and this partly mitigates the *lack of focus on real use contexts in relation to engineering* [114] that is common in HCI research. Most of the field work carried on involved observing, annotating and making sense of human gesturing. This can be done in several ways, with or without the aid of (semi)automatic tools. When possible, video footage has been collected for further analysis. Note that video analysis is typically the realm of very skilled and trained experts. However there are methods to extract useful insights through a not so formal, but yet sound approach, as described further below.

Several frameworks found in literature (for example [50, 111, 136, 141]) are a necessary starting point to define (and recognize in practice) the phenomenon to observe. This models however are far too broad; other researchers have suggested frameworks to explore the boundaries between natural gesturing, as observed in human-human communication, and gestural interfaces. These provide a contextualization to HCI that will nicely fit the above research goals.

Robertson [164] presented a taxonomy of embodied actions as observed in cooperative work, specifically in software design. It includes manipulation and gestures performed by individuals (both in relation to physical objects and to the workspace) and how these actions are used in group activities, for example to create shared representation, negotiating group attention, etc. The work of Brereton et al. [21] is particularly relevant here, given its approach and goals. It is based on the collaborative analysis of video footages showing activities in different workplaces. The video footages are analysed using the method of the *Video Card Game* [24], that allows non-experts to annotate and comment on relevant design concepts. The analysis is then used to define several *Gesture Themes* and discuss potential implications for design.

For the very nature of this research, most of the field work and lab experiments are aimed at exploring the blurred zone between the actions that people perform and the actions that computers can sense, with the idea that focusing on what happens within one meter from the screen could help to better understand the values, goals, practices and expectations of people, and, ultimately, to design better interactive spaces.

Design frameworks have been proposed to drop light into this dark area. Benford and co-workers describe this design space in terms of *expected, sensed, and desired* [13] actions and suggest to explicit consider those unexpected or unlikely patterns of interaction, treating them as design opportunities.

It is worth remarking here that the challenge of finding a balance between research and design, or doing *research through design* characterizes the field of HCI (see for example [199]). The interaction designer draws from several sources: technology from the engineers, field data from the anthropologists, behavioural models from social scientist. The creation and evaluation of prototypes provides feedback both to the engineers (in terms of technical opportunities) and to the the behavioural scientist (in terms of observed gaps between theory and practice).

Though my academic background is that of an engineer, during this work I had to wear, from time to time and at the best of my abilities, the hat of the anthropologist and that of the social scientist. Throughout the text, when such change of perspective happens, it is made clear.

2.2 Contributions

This work gives several contributions to the state of the art in the field of multitouch and natural interaction. Designing experiments to compare different interaction techniques is not always straightforward. Making such comparisons often require a working prototype of a complex (an potentially expensive) technology, such as interactive walls, multitouch tables, etc. Deploying such technologies on commodity hardware and at low cost is still a research problem, we describe in more depth our original contribution to this field in Part 3; in particular we have prototyped a fully functional interactive space with:

- a 3 mt wide interactive wall whose multi-touch sensor relies on an original technology described in section 9.1. The display consists in an array of projectors; obtaining a seamless image from the juxtaposition of projectors is not trivial using commodity hardware. Section 9.2 describes our original contribution to this aspect;

- a multi-touch table based on FTIR[3]. Our implementation adds several improvements to work in uncontrolled lighting exploiting the shadows that the hands of the user project on the surface in a noisy operational condition (section 9.3);

- a software framework for the development of cross-device multiuser applications [42] that provides abstractions over physical sensors, abstract and concrete multi-touch widgets (supporting manipulation of images, web pages, movies, etc.), fiducials recognition and tracking, and a distributed asynchronous event subscription/delivery architecture[4];

[3] Frustrated Total Internal Reflection [71]

[4] Although not discussed further in this work, it is the necessary infrastructure which all the applications described here rely upon.

The interactive space was used to conduct a number of experiments, both to test and asses the technology itself and to evaluate novel interaction techniques. Of the latter, one is described in great depth in chapter 10, and shows that:

- the adoption of a multitouch user interface fosters a significant, observable and measurable, increase of nonverbal communication in general and of gestures in particular, that in turn appears related to the overall performance of the users in the task of algorithm understanding and debugging.

Other experiments, that were helpful in shaping the interactive space, but are not reported in detail include:

- initial exploration on using the MS Kinect to detect hand shape and position and allow manipulative-like interaction with digital entities such as a 3D model, described in depth in [86, 88].
- a prototype application for the (semi)automatic recognition and annotation of human behaviour in a scientific video, described in [30];
- a prototype programmable emitter of scents for experimenting with olfactory interaction, described in [38];
- an experiment of multimodal touch/pen interaction described in [27];

2.3 Published as

Parts of this dissertation have been published before. Here we further develop those concepts and show them within the bigger picture.

A. Soro, S. A. Iacolina, R. Scateni, S. Uras. *Evaluation of User Gestures in Multi-touch Interaction: a Case Study in Pair-programming*, 13th International Conference on Multimodal Interaction - ICMI 2011. [171]

S. A. Iacolina, A. Soro, R. Scateni. *Improving FTIR Based Multi-touch Sensors with IR Shadow Tracking*. Proc. Of EICS 2011 ACM SIGCHI Symposium on Engineering Interactive Computing Systems. Pisa Italy 13-16 June 2011. [87]

A. Lai, A. Soro, and R. Scateni. *Interactive calibration of a multi-projector system in a video-wall multi-touch environment*. In Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST '10). ACM, New York, NY, USA, 437-438. [122]

A. Soro, M. Deriu, and G. Paddeu. 2009. *Natural exploration of multimedia contents*. In Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia (MoMM '09). ACM, New York, NY, USA, 382-385. [170]

Other papers have been published as part of the activities involved in this work, but have not been included in the final discussion: they are summarized in the Appendices where relevant.

S. A. Iacolina, A. Soro and R. Scateni. *Natural exploration of 3d models.* In Proc. Of the 9th Conference of the ACM SIGCHI Italian Chapter. Sept 13-16, 2011. Alghero (Italy). ACM, New York, NY. [88]

D. Cabiddu, G. Marcias, A. Soro, and R. Scateni. *Multi-touch and tangible interface: Two different interaction modes in the same system.* CHItaly 2011 Adjunct Proceedings, 2011. [27]

M. Careddu, L. Carrus, A. Soro, S. A. Iacolina, and R. Scateni. *Moravia: A video-annotation system supporting gesture recognition.* SIGCHI - CHItaly 2011 Adjunct Proceedings, 2011. [30]

V. Cozza, G. Fenu, R. Scateni, and A. Soro. *Walk, look and smell through.* SIGCHI - CHItaly 2011 Adjunct Proceedings, 2011. [38]

M. Deriu, G. Paddeu, and A. Soro. *Xplaces: An open framework to support the digital living at home.* In Proceedings of the 2010 IEEE/ACM International Conference on Green Computing and Communications & International Conference on Cyber, Physical and Social Computing, GREENCOM-CPSCOM '10, pages 484-487, Washington, DC, USA, 2010. IEEE Computer Society. [42]

S. A. Iacolina, A. Lai, A. Soro, R. Scateni. *Natural Interaction and Computer Graphics Applications*, EuroGraphics Italian Chapter 2010, pp. 141-146. Genova, Italy, November 2010. [86]

Chapter 3

Applications

As mentioned earlier, there is a number of potential applications that press towards including gestures in interactive systems in a way that differs from the main track of gesture recognition. In the first place we must be aware of the cognitive role of gestures and manipulations that have been introduced in chapter 1 and will be discussed in greater depth in 4. A second important track is the evaluation of human behaviour, such as emotion recognition. Intuitively we know we can say so much of a scene only observing gestures, and a great part of the memory that we carry of events (or of movies) is connected to nonverbal expressions. Finally, we mention an example application in which the human ability to mimic shapes and movements with the hands (ideographic gestures, or pantomimes) is used to animate, shape or interact with 3D digital characters.

It needs to be mentioned at this point that the following don't want to represent an exhaustive list of possible applications. Instead this chapter is meant to trace the context in which the present research has been carried on. I've used these three *application fields* to narrow the scope of my work and decide what natural actions were to be included in my literature review, and what was to be left to further work.

3.1 Gestures, Manipulation and Effectiveness

Anyone who has prototyped and demonstrated a *natural* interface, such as for example a multitouch screen, has probably at least once been addressed a question such as: *. . . nice but . . . how do you expect people to use a spreadsheet on this stuff?* There are several good answers to such a question, including replying very patiently that spreadsheets and office automation is not exactly the sort of applications which multitouch is best for. But it sounds like sort of a defeat. It would be much better to have grounds to affirm that, although user interfaces may have to be redesigned in various ways, natural interfaces can

perfectly cope with traditional application domains such as office automation, perhaps with advantages in terms of efficacy, in addition to pleasure of use.

Pursuing effectiveness may sound a bit old-fashioned in interaction design. I completely agree with those who suggest to reorient HCI research towards human goals, expectations and values (see for example Bannon [9]). I also quote Overbeeke and co workers when they say that:

> Interfaces should be surprising, seductive, smart, rewarding, tempting, even moody, and thereby exhilarating to use. [148]

Buxton provides and enlightening analogy for such philosophy in a recent interview:

> I think there's no profession that has a longer history of expressing powerful ideas to a technological intermediary than music [...] you look at somebody who picks up a beautiful guitar or instrument and before they play a note you can tell from the smile on their face the quality and the respect built into the instrument for that person's skills. [123]

However, nobody seems to really believe that the new instrument will be good for playing the old music. If we assume that natural interfaces are tailored on human abilities, isn't it reasonable to expect them to just *work better* than the old ones? Chapter 10 describes an experiment in this direction: in a scenario typically office-oriented, such as software development, we show that people perform better using a multi-touch table that they do using a desktop computer.

3.2 Evaluation of Human Behaviour

How much of our understanding of a conversation depends on gestures? Intuitively we may say that the *contents* depend on the verbal part, while the gestural behaviour conveys less precise, say *context*, information. However, this is not always the case: in the first place gesticulation can provide cues about the structure of the sentence, and the relation between gestures and semantic fragments appears to be relatively predictable and automatically recognized by a software [153]. Moreover, there are a number of real life examples of speech whose meaning can be caught *only* with the help of accompanying gestures. The following examples better explain this point.

> [...] a boy, excited about going out to the shops to buy himself a new camera, said. "Tonight I'm going to get my /GESTURE/" - the gesture here being an enactment of holding a camera to the eye and pushing the shutter. A colleague at a professional meeting sitting two or three seats from his friend caught his friend's eye as the meeting closed and said in a loud whisper ... "Do you want to go/GESTURE/" - here pointing his thumb at his mouth and tilting his head back slightly as he did so. A

> film director calls to a member of his lighting crew "Number five balcony /GESTURE/" - in this case gesturing his instruction to turn the light, but specifying in words which light is to be turned.
>
> Commonly and widely, then, gestures are employed as alternatives to spoken elements in utterances. [105]

Actually, there are mechanisms of which we are barely conscious, and that are just emerging as scientific evidence; for example it is relatively well known that people can discriminate visually (i.e. without any auditory clue) between a familiar (mother tongue) speech and a foreign one. More surprising, small children (4-8 months old) seem to have the same ability [183]. Furthermore, people rely on visual information when discriminating emotions [3], and many efforts exist that are targeted at emulating such human ability (e.g. [16, 32, 37, 70]).

Despite this, for example, the way we query the largest multimedia database ever created is still largely based on verbal descriptions, and ultimately on text keywords; considering the meaning carried by gestures and facial expression would allow a step further in this an in many other fields.

3.3 Natural Interfaces in Computer Graphics

Human Computer interaction and Computer Graphics share some of their milestones. The birth of computer graphics is traditionally connected to Sutherland's Sketchpad [174], the first computer application to provide direct manipulation, visual feedback, graphic abstractions and some of the powerful metaphors that today are so familiar; to borrow his own words:

> The Sketchpad system makes it possible for a man and a computer to converse rapidly through the medium of line drawings. Heretofore, most interaction between man and computers has been slowed down by the need to reduce all communication to written statements that can be typed; in the past, we have been writing letters to rather than conferring with our computers. For many types of communication, such as describing the shape of a mechanical part or the connections of an electrical circuit, typed statements can prove cumbersome. The Sketchpad system, by eliminating typed statements (except for legends) in favor of line drawings, opens up a new area of man-machine communication. [174]

It should not be surprising if computer graphics has been the playground of natural interfaces in the following decades, in order to explore the power and limits of direct manipulation on geometrical models, maps, mathematical representations, etc. [6, 85, 147, 147, 197, 198] However, as it is the case for other application fields, the role of gestures has been somewhat underestimated. Gestures have been explored as possible *commands* for navigating virtual worlds and manipulating graphical entities.

Again, from the observation and description of how people gesture in the *real world* we could draw many insights that should inform the design of *natural* computer graphics applications. Consider how people use their hands to describe the shape and the movements of objects in storytelling; observe how people *enact* imaginary characters with their hands and body; such skills are part of the natural *toolbox* of any person, but have found little space, so far, in interactive computer systems.

To see a world in a grain of sand,
And a heaven in a wild flower,
Hold infinity in the palm of your hand,
And eternity in an hour.

William Blake

Part II

Related Research

Chapter 4

Gestures

4.1 Gestures in HCI

The first experiments in gestural interaction date back to more than 30 years ago, fed, among other things, by several decades of studies on pen computing, shape and handwriting recognition. Bolt's *Put That There* [17] is generally recognized as the first human computer interface to exploit proper hand gestures. The system was meant to demonstrate and experiment with the *Media Room*, a technology augmented office space designed to be an *embodiment of the user terminal as an "informational surround"*. In such space Bolt managed to apply the concept of *Spatial Data Management System*, i.e. a system that allows its users to exploit the innate sense of spatiality to access information.

In his experiments Bolt coupled continuous speech recognition to hand pointing gesture sensing by means of a small hand-held device. Commands such as `create a blue square there` could then be issued to the system combining speech and gesture to specify what had to be done, where and how. A key aspect of Bolt's work is the possibility, when needed, to replace verbal descriptions with gestures and vice versa. For example the command `move the blue triangle to the right of the green square`, can equally well be formulated as `move that+GESTURE to the right of the green square`; as Bolt observes, here the object to be moved is *obstensively defined*, which has a sound theoretical ground on cognitive science [145].

Only a few years later, another landmark on gesture interaction is represented by Krueger' s VIDEOPLACE [119]. VIDEOPLACE was based on Krueger's own previous work on a *Responsive Environment*: a room equipped with sensors capable of tracing the position, movements and postures of people, that could then be used to control computer programs, such as navigating a maze displayed on a screen by moving one's own body in the room [118].

VIDEOPLACE added gesture recognition to explore the possibilities of

interaction with digital objects. The image of the user is captured and separated from the background exploiting pixel subtraction, and then projected in the virtual world displayed on a wall-sized display. The user becomes then part of this virtual world, and can *touch* and *manipulate* its virtual constituents. Krueger demonstrated a number of interaction techniques that would have been 'reinvented' (and patented) again and again in the following years, for example dual handed manipulation and gestures such as the *pinch to zoom*. One characteristic of these early works that got somewhat lost afterwards is their visionary nature. In the words of Krueger:

> VIDEOPLACE is not so much a solution to existing problems, as an effort to stretch our thinking about the human-machine interface. We have already entered an era where most of the people using computers are no longer programmers in the traditional sense. We can look to a day when most of the people interacting with computers will not be users in the current sense. [119]

The rest of this section provides an overview of the theoretical ground behind gestural interfaces. Further details on devices and algorithms are given, where relevant, in section 6.1. Of greater relevance are the following sections and chapters, that relate the state of research about gestures, manipulative action and how these impact the design of interactive spaces.

4.1.1 Gesture Recognition

The field of gesture recognition in human-computer interaction is very broad and encompasses the design and application of algorithms and devices capable of classifying the widest variety of body movements and postures; not surprisingly, the need for a unifying framework to gather and describe such diversity of viewpoints is generally accepted.

Pavlovic and co-workers present a thorough review of the methods and applications of computer vision based gesture interpretation [151]. They gloss over the problem of defining gestures outside the field of HCI, but within it provide a useful taxonomy that discriminates between intentional and unintentional movements. The former can be either manipulative or communicational gestures. Communicational gestures are then classified according to Nespoulous and Lecours [141] (although erroneously attributed to Quek [155]) into *acts* and *symbols*, depending on their level of abstraction:

Acts are characterized by a close and concrete relationship between their shape/trajectory and the meaning, they are either *mimetic* or *deictic* gestures;

Mimetic gestures are often performed as imitations of their referent, such as in pantomimes. The shape of the hand, or the movement of the hands is meant to recall the actual meaning;

Deictic gestures are gestures performed to point at the referent. They can be further distinguished in *specific*, when the subject indicates a particular object, *generic* when the subject points to an instance of a class of object meaning the class as a whole, *metonymic*, when the subject points to an object but refers to a concept related to the object, such as pointing at one's wrist to ask the time.

Symbols are abstract representations and are largely arbitrary: can be distinguished in *referential* and *modalizing*:

Referential gestures are literally words to be seen. They can be interpreted independently of other constituents of a sentence; examples are the *ring* or the *thumb up* symbol to indicate OK.

Modalizing gestures acquire meaning in relation to another channel of communication, typically speech; for example angler's tales will be often accompanied by a very eloquent modalizing gesture specifying the supposed size of a missed prey;

Karam [100] proposes a framework for classifying gesture based computer interactions according to *gesture styles*, *system response*, *enabling technologies* and *application domain*. Such categorization was informed by an extensive review of previous research and gives the sense of how rich this field is. At the same time it exposes clearly that the research in the field has been focused mostly on specific *styles*, mostly deictics (pointing), semaphoric (conventional hand symbols), manipulative (handling objects or mimicking the same action), and a combination of the above.

Mitra and Acharya provide a review of the models and algorithms involved in gesture recognition [138]. They provide a distinction between *hand and arm* gestures, such as hand poses or sign languages; *head and face* gestures, which include e.g. nodding, shaking of head, raising the eyebrows, but also eye gaze; and *body* gestures, i.e. tracking movements of people, analysing movements of a dancer, recognizing human gait or posture. When addressing specifically the recognition of hand and arm gestures Mitra and Acharya rely on the well known Kendon's classification into *gesticulation, languagelike gestures, pantomimes, emblems, sign languages*. However, they only present applications for sign language, handwriting recognition and emblems recognition.

The starting point of most (if not all) of these studies is the seminal work of Kendon [105–109] and Efron [48], and more recent research in speech and language by McNeill [134–136] and Goldin-Meadow and colleagues [36,49,63,64]. However, the attempt to tailor the theories and definitions of social science to the more practical problems of human computer interaction is leading to several, sometimes confusing, frameworks, that quite ironically, are at the same time too loose in scope and too narrow in perspective.

As we have already mentioned the term 'gesture' in HCI has been used in relation to almost any body movement, from eye gaze to foot steps. This may be

regarded as a matter of property of language, and any author could reasonably argue that in absence of a universally accepted definition of *gesture*, their own use of the word is just as valid as any other one. If the largest majority of authors don't even bother giving a definition of what they regard as a *gesture* in their research, those who do seem to focus opportunistically on rather different aspects:

> A gesture is a motion of the body that contains information. Waving goodbye is a gesture. Pressing a key on a keyboard is not a gesture because the motion of a finger on its way to hitting a key is neither observed nor significant. All that matters is which key was pressed. (Kurtenbach and Hulteen, 1990 [120])
>
> In the context of visual interpretation of gestures [...] A hand gesture is a stochastic process in the gesture model parameter space $\mathcal{M}_T$ over a suitably defined time interval $\mathcal{I}$(Pavlovic et al., 1997 [151])
>
> Gestures are expressive, meaningful body motions involving physical movements of the fingers, hands, arms, head, face, or body with the intent of: 1) conveying meaningful information or 2) interacting with the environment. (Mitra and Acharya, 2007 [138])

But at the same time, as we pointed out above, those researcher that focus on those *intentional hand movements whose explicit and primary goal is to convey a message*, which is largely the most accepted definition of gesture, are apparently focusing on sign languages or emblems, despite their being in general less *natural* (they must be learnt) and less *universal* (their meaning vary over time and across cultures).

On the other hand, works the focus on multi-modal speech/gesture interaction are generally less vague on specifying the nature of gestures they focus on; for example:

> we prototype a set of what Nespoulous and Lecours [141] term "coverbal gestures", in particular, gestures which serve to illustrate ongoing verbal behaviour. Within Nespoulous and Lecours's classification of illustrative, coverbal gestures, we have selected examples that are *kinemimic* (outlining the action referred to by one of the lexical items) or *spatiographic* (outlining the spatial configuration of the referent of one of the lexical items). (Bolt and Herranz, 1992 [18])
>
> The most general definition from the 1977 Lexis dictionary says that gestures are 'movements of body parts, particularly the arms, the hands or the head conveying, or not conveying, meaning'. Nespoulous and Lecours [141] divide gestures into centrifugal gestures ('having obvious communicational intent') directed toward the interlocutor, and centripetal gestures which may be observed and interpreted as mood and desire indicators although they are not directed at anyone. (Quek, 1994/95 [154, 155])

The following sections present the perspective offered on gesture by social science, specifically anthropological and psychological research. As Quek remarks:

> Turning to the work of semiotics who have investigated the subject, we derive a sense of how gesture is used and perceived among human beings. The purpose of this exercise is not to build an interface system that is able to interpret all the nuances of human gesture. The limitations of current technology will invariably limit our ability to do so. Ultimately the human user will have to meet technology 'halfway'. [154]

After nearly two decades it is interesting to repeat the *exercise* to evaluate what progress HCI has done towards that *halfway* meeting point, particularly in consideration of the fact that *computing* has moved from the office desk to the pockets, handbags, walls, living rooms, and in all these new settings it is pushing towards radically new forms of interaction.

4.2 Gestures in Social Science

The history of gesture studies is a fascinating topic that deserves to be mentioned here, if only because it is representative of how the availability of technology impacts not only the research agenda, but also the long term vision of any era. This is a particular pain for technologists, that must shape the tool of tomorrow using those ones available today.

In the case of gesture studies, the scientific evidence of the inextricable connection that bounds language, thought and gestures become available after affordable audio/video recording made its appearance in the toolbox of social scientists, that is to say, right after WWII. However, the study of gestures remained somewhat in the background until the mid 1970. The spark of such new interest was the study of sign languages, especially a successful attempt to teach sign language to a chimpanzee [61]. Further details on this topic are largely beyond the scope of this work and can be found in [111].

My goal here is to review the many possible definitions of what a *gesture* is, and to describe how such example of human action relates to the way humans think, learn, react to the environment, etc. At a certain extent, the adoption of a naive definition (or no definition at all) of gesture correspond to our shared understanding of the meaning and role of gestures. However, the observation of people, focusing on the relation between gestures and language, social behaviour, thought, memory, etc. throws more light on this subject.

Ekman and Friesen identified 5 *categories* of non-verbal behaviour [50]:

Emblems are non-verbal acts that can be given a verbal definition or correspondence, usually a word or short sentence.

Illustrators are movement directly linked to speech, and serve to better illustrate the verbal content.

Affect Displays are mostly coincident with facial expressions.

Regulators are gestures used during discourse to negotiate the flow of the conversation, requesting attention, askin the speaker to pause, continue, wait, etc.

Adaptors are actions learnt in childhood that supersede their original manipulative purpose and become communicative gestures. One example is the gesture of bringing the hands to the corners of one's eyes, as if wiping them from tears, to indicate sadness.

Part of this terminology remains in later literature, and the overall discussion of the five categories against *coding*, *usage*, and *origin* of gestures is by itself a useful framework.

Kendon reported of experiments in which people, when asked to naively identify *gestures* in a video footage, consistently pointed at intentional and deliberately communicational movements of the hands [104]. Additionally, Kendon points out several features of proper *gestures* that distinguish such movements from any other that people perform at different level of consciousness [110] for example:

- gestures move to and from a rest position;
- gestures have a peak structure (a stroke);
- gestures are well bounded (we can recognize when a gesture begins and terminates);
- gestures are mostly symmetrical.

Such characteristics can be exploited to segment hand movements in *gestural phrases* to analyse in conjunction with body posture, eye gaze, speech, etc. Kendon also puts emphasis on how humans move and combine with great effectiveness their gestural toolbox, using gestures, from time to time, and according to necessity and opportunity, as a substitute for speech, as a visually evocative system, as a substitution for single words, or plainly to accompany speech [109]. This concept was later re-elaborated by McNeill that dubbed it *Kendon's continuum* and suggests a classification for gestures that encompasses:

Gesticulation is the (almost) continuous movement of the hands that people perform while speaking. It is probably the more common and less structured form of gesturing, and has very deep roots in how people think, how spoken language is produced and learnt. This attitude seems to be innate (for example people who are blind from birth are known to gesticulate no less that sighted people [95]);

Speech-linked gestures are closely related to speech structure and occupy precise grammatical slots in the accompanying sentence, often they underline a concept, or can be regarded as a sort of *visual accent*;

Emblems are conventional signs, linked to a well defined concept, such as the *ring* symbol for OK, etc. such gestures are more dependent on cultural aspects, and can have different meaning across different communities;

Pantomimes are gestures that visually represent a concept, movement or shape;

Sign languages are complete languages equivalent to spoken ones, with complex grammatical rules and abstract symbols. They stay at the far end of the continuum with respect to gesticulation.

As a final note, Kendon doesn't give a special role to deictics (i.e. pointing gestures) in this schema, but rather underlines that *Many gestures have a pointing component, as well as many that seem to be 'pure' points.* [110]. Pointing gestures can differ in what body part is used to point, and when the hand is used, the shape can vary depending on the context. The *target* can also vary, and can be either a real object concrete and visible in the environment (actual pointing), objects that are not present in the environment, although they could or they use to (removed object pointing), concrete objects that are given an imaginary location to point at only for the purpose of discourse (virtual object pointing), objects that are not concrete, and thus cannot have any real location (metaphorical object pointing).

The Kendon's continuum has been thoroughly discussed by McNeill [136] that puts in great evidence the relation between gesture and language. In particular McNeill observes that *gesticulation* (i.e. free and unstructured hand movements) *always* occur in coordination with speech. The presence of speech decreases while we move along the continuum through *speech-linked gestures*, *emblems*, *pantomimes* and up to sign languages, that almost *never* coexist with speech[1].

A similar consideration can be done regarding the linguistic properties of the system of gestures: gesticulation doesn't show any linguistic property (in terms of abstraction, conventionality, syntax, etc.), sign languages have the characteristics of a complete language, while emblems and pantomimes only show weak linguistic properties in the way they can be combined. Finally, McNeill underlines that gesticulation is inherently *global* (i.e. the meaning of the parts are inferred from the meaning of the whole) and *synthetic* (i.e. a gesture covers a set of meanings that in a sentence would be typically bounded to different words, as it happens for example in any gesture that has a pointing component, that in a sense encompasses verb and predicate). On the contrary sign languages are clearly *segmented* and *analytic*.

McNeill contributed to the classification of gestures the well known quartet: *iconic*, *metaphoric*, *deictic* and *beat* gestures:

[1] This is due, according to McNeill to the way sentences are formed and articulated: while the orchestration of the sentence follows temporal rules in spoken language, it is subject to spatial rules in sign languages. The two things cannot coexist easily.

Iconic gesture are visible representations of entities or actions: the form of the gesture, the shape of the hand, the trajectory, speed or anyway the manner in which it is performed are meant to recall some semantic aspect of the entity referred to. These gestures are closely linked to the accompanying speech, but in a non redundant way.

Metaphoric gestures are visible representations of abstract objects or concepts. They may involve the manipulation of an imaginary tool, using either one or two hands. They are frequently observed when subject discuss abstract concepts, and resemble the act of holding or offering an object, while in fact the object offered is the *idea* or the mental representation that the subject has formed of the object itself.

Deictic gestures have their own place as distinct, autonomous gestures. Pointing is one of the first forms of gestural communication in children, and one that humans continue to evolve for a long period before mastering; it can be referred to concrete and abstract objects, and McNeill observers how the latter emerges surprisingly late, around the twelfth year of age.

Beats encompass Kendon's gesticulation and speech-linked gestures: they accompany spoken language and serve a *beating* and *highlighting* purpose in the underlying narrative. Beats can be used to mark and underline a word, to give emphasis to a pause, to highlight a concept, etc.

So, why people gesture? Though answering this question may be out of reach, the observation of the *consequences* of gesturing can be enlightening. Kendon's emblems, pantomimes and sign languages, or McNeill iconic and deictic gestures, have a clear communicative intent: they are words (or full sentences) to be seen. On the other hand, the consequences of gesticulation (or beat gestures) are less evident, and somewhat surprising.

People gesture while speaking, and do so not only with communicative purposes, but also for themselves. As a result people tend to gesture when no-one is viewing them [64], or when talking to someone that cannot view the gestures, such as a blind person [95]. The effects of such gesturing on a simultaneous task can be observed and measured.

Rauscher and colleagues [160] conducted an experiment in which 41 students where asked to describe 6 short video-footages, 3 of which freely gesturing, while 3 refraining from gesturing. Each one of these two conditions was then related to three additional variables: 1) in *normal speech* condition, subject were given no constraint on the words to use; 2) in *obscure speech* condition subjects were asked to use as many obscure words as possible; 3) in *constrained speech* condition subjects were asked to refrain from using a certain set of words in their descriptions. As a first result, Rauscher and co-workers observed that gesturing appeared three times more often in conjunction with sentences relative to spatial contents. Additionally they found that in sentences involving spatial contents, gesturing improved verbal fluency.

In a more recent research [63] Goldin-Meadow and co-workers tested 40 children and 36 adults on math exercises (addition problems for the children and factorization problems for the adults): participants were asked to solve the problems at a blackboard. Right after solving the problems each participant was given a list of items to remember (words for children, letters for adults) and then were asked to explain how they computed the solution to the problem. After the explanation all participants were asked to repeat the list just memorized, and the recall rate was measured in two different conditions: gestures permitted during task explanation and gesture forbidden. As a result, both children and adults remember a significantly higher number of items when allowed to gesture that when forbidden. Goldin-Meadow and colleagues suggest different possible explanations, all of which implies that gesturing while explaining the task frees cognitive resources, that a person can exploit to improve the recall rate.

Going further in describing more and more experiments that explore the impact of gesturing on the widest variety of educational or working tasks would be pointless here. Similar phenomena have been observed however in children's learning to count [2], in learning to solve simple equations [36], in spatial transformation tasks [49], etc. What is relevant here is the subtle nature of such phenomena, and at the same time the fact that the vast majority of people are completely unaware of their own relying on gesture for partially supporting (among other things) memory, learning and spatial reasoning.

In the following chapter we present the natural companion of gesturing: manipulation. Although manipulative act has been sometimes regarded as a special kind of gesture (or alternatively gestures have been regarded like a special *free-hand* manipulation), there are practical and theoretical reasons for treating these two skills separately, before trying to identify a unifying framework.

Chapter 5

Manipulative Action

Manipulations are actions intentionally performed on objects in order to change the state of the world. Although in a broader sense manipulation can be performed by means of any part of the body or without any physical contact, such as in *media manipulation* or *genetic manipulation*, it is clear that in the largest majority of cases manipulation is the action of our *hands*.

As it often happens when human abilities are involved, the simple mechanics of grasping reveals, when explored, a whole universe of orchestrated movements and perceptions [33]. At the same time the tight relationship between hand and mind, confirmed by scientific evidence, has been intuitively known for ages, at the point that verbs such as *to grasp, to catch, to get* share the double meaning of *grabbing* and *understanding*.

Abstracting from the underlying *biomechanical and physiological complexity*, Guiard's *Kinematic Chain Model* [69] can be exploited to make sense of the division of labour between the dominant and non-dominant hands in bimanual skilled manipulations. The KC model is hinged on three principles (the following examples hold for right-handed people:

right-to-left spatial reference: motion of the right hand is spatially referred to the motion of the left hand, such reference role of the non-dominant hand is also responsible for keeping in place the object to manipulate;

different temporal-spatial scales of motion: compared to the dominant hand, the non-dominant hand moves at a relative low frequency (longer periods of rest) and relative low spatial frequency (larger movements);

Left hand precedence the contribution of the non-dominant hand to cooperative action stars earlier than that of the dominant hand.

Applications of Guiard's KC model to human computer interaction tend to uncover the difficulty of implementing virtual manipulation. For example

Hinckley and colleagues show that in the manipulation of physical objects people use to switch to asymmetric bi-manual action (i.e. hold steadily the object with one hand and operate precise manipulations with the other) depending on a combination of factors that include the nature of the task, the mass of the objects being manipulated and the possibility to keep the preferred viewpoint [75].

HCI research has pursued two complementary approaches to this issue. One is to work on the digital side, implementing dual handed input to graphical user interfaces, such as with multi-touch screens. Another path is to link digital information to physical objects, thus keeping manipulation in the real world, such as with tangible user interfaces. Both approaches are described in more depth later on in this chapter.

Before moving on to a review of the technologies that are based on direct manipulation it is important to underline that the tight link between hand and mind shown in the previous chapter regarding gestures, exists (with several distinctions) regarding manipulations.

Hand action is known to fulfil either an explorative or a proper manipulative function which Cadoz [28] dubbed *epistemic* and *ergotic*, respectively. However, as it happens for gestures, there are subtle effects on problem-solving performances related to a proper use of manipulative skills.

Kirsh and Maglio further refine such distinction, exploring the cognitive effects of manipulative action. They define *pragmatic actions* as actions whose primary intent is to bring an agent closer to a predetermined goal, while *epistemic action* are physical actions whose primary purpose is to improve cognition by: *(i)* requiring less memory, *(ii)* requiring fewer steps, *(iii)* reducing probability of errors, i.e. improving reliability and time/space complexity [113].

This behaviour is evident in tasks involving the manipulation of symbols, such as algebra or arithmetic. Kirsh and Maglio however managed to uncover it also in a task that, though not clearly symbolic, required a quick and efficient reaction to visual stimuli, such as the game *Tetris*. They conclude that the commonly adopted model of planning as a *state-space* search should be revised in order to account for both *physical* and *informational* states. In other words, when solving problems, people often trade-off between actions that bring them closer to the solution and actions that provide an *informational payoff*.

Superficially this may recall a plain trial-and-error learning strategy, but Kirsh and Maglio managed to prove that epistemic action *increases* with skill: expert players of the Tetris game perform more trial and backtracking actions, thus apparently translating from the mind to the world a certain computational burden [128]. In practice, the better performances of skilled users are partially due to their greater ability to use the world as an *interactive visuospatial sketchpad*.

The next sections explore technological frameworks developed to take advantage or people's natural manipulative skills: multi-touch interaction exploits special touch screens capable of sensing more than one touch point in order to

allow a richer, yet generally GUI based, interaction. Tangible user interfaces aim at bridging the gap between the digital and the physical world.

5.1 Multi-touch Interaction

Multitouch interaction makes use of special touch screen capable of sensing more than one point of touch in order to provide a richer user experience in terms of manipulative actions (as in the case of the *pinch-to-zoom* function on modern smart-phones), dual-handed interaction (of which the already mentioned VIDEOPLACE [119] o Han's FTIR [71] screen are excellent examples[1]), or multi-user interaction.

The specific technological solutions used to implement multi-touch screens vary from arrays of antennas to computer vision and will be reviewed where appropriate in Part 3. Here we trace an history of the development of multi-touch screens in order to highlight the new possible interaction styles that such technology has made available.

One of the first application scenarios for multi-touch interaction has been the manipulation of shapes and graphical elements on a screen. Buxton and Myers explore [26] the effect of bimanual continuous control on tasks of positioning/scaling and navigation/selection. They found a (then surprising) attitude on the users' part to adopt dual-handed parallel manipulation, which leaded to improved performances of the tasks.

In 1991 Paul Wellner's *Digital Desk* [186, 187] was a visionary tabletop application prototype used to demonstrate bare-hand interaction with digital documents. Wellner started from the *desktop* metaphor, whose goal is to make the computer similar to an office desk, and suggest instead to make the desk appear more similar to the computer, projecting digital documents on it. As it happened for the VIDEOPLACE, the underlying principles of this work are so powerful in their simplicity that several of Wellner's scenarios has been re-invented and re-implemented, with greater and greater effectiveness while technology became available at cheaper prices.

Balakrishnan and colleague explored the use of bi-manual techniques borrowed from automotive design and applied them the *Digital tape drawing* [5] through which the user could sketch smooth curves with great precision on an interactive whiteboard.

An important characteristics of multi-touch screens is that often they can be used by many people *at the same time*. Concurrent co-located use of technology, either in collaborative or competitive scenarios, is a topic of which the research community is yet barely scratching the surface. If the potentialities of shared interactive screens, e.g. in education [162] are reasonably foreseeable, other

[1] Video demontrations of these and other technologies are available online `http://www.youtube.com/watch?v=dmmxVA5xhuo` and `http://www.youtube.com/watch?v=QKh1Rv0PlOQ`

phenomena are emerging as unexpected characteristics, still worth exploring. The negotiation of the interactive space between strangers is one aspect that can impact the design of public ambient displays: it results that sometimes conflicts may arise when people compete to access the device or parts of it [152].

Furthermore, despite the fact that multi-touch is a main character in current technological development, its application to real world settings in public spaces is still in its early envisioning phase, and users are far from *familiar* with it, revealing that this technology, despite its high potentialities, is far from being *natural* in the naive sense of the term [79].

5.2 Tangible User Interfaces

Tangible User Interfaces (TUIs) exploit physical objects to control the state of digital data: the physical control is coupled to its digital model to grant easy access to the model's functionalities. TUIs represent a growing and increasingly popular research area that sits at the intersection of industrial design, cognitive science and computer science, and aim at overcoming the limitations of the digital data by allowing direct manipulative access to information.

In the early phases of the development of this research field tangible user interfaces almost invariably consisted in the use of so called *phycons* (physical icons) and tools as graspable interface elements of a generally more complex background interactive screen or environment [93, 178, 179].

There are a number of advantages from such approach.

- the interaction takes place in the physical space: instead of manipulating graphical entities that represent digital objects, the user manipulates objects themselves;
- the interaction and its effects on an artifact happen at the same time and in the same place;
- the interface encompasses the state of the model, i.e. the user interface is not meant to *represent* the state of the system, but rather the interface *is* the state of the system.

The analogy with the duality of epistemic/pragmatic action is not casual: a key aspect and fundamental design principle for TUIs is their suitability for explorative manipulation. The natural user interface described in [56] is also designed to explicitly support natural interleaving of epistemic and pragmatic action in an augmented reality environment. Attempts have been made to identify epistemic action in other specific context, such as interaction with 2D and 3D maps [147], bi-manual editing and manipulation of 3D objects [6] and 2D curves [149]. In general, User Interfaces that allow epistemic action are perceived as more powerful, even if this preference on the users' part doesn't always correspond to an increase in performance.

Despite the fact that epistemic action has been first analyzed in the context of a graphical video game, little work exists that explore the impact of epistemic action in Graphical User Interfaces (GUI). Patten and Ishii have compared how people use spatial organization of physical objects on a desk with respect to icons on the desktop and found that it is significantly easier to recall object-resource associations in the case of TUIs than it is with GUIs [150].

Current research on tangible user interfaces has differentiated across a wide range of application domain, some of which are perhaps more easily labelled under pervasive computing or robotics. Ishii reviews some of these domains [92], that share the characteristics of letting the digital information emerge in physical form and become concrete to the user's perspective, either at a perceptual or functional level. Of the example application domains for TUIs presented by Ishii (see [92] for a thorough discussion), two are relevant to this work:

Interactive Surfaces-Tabletop TUI explore the possible synergies and combinations of tangible interface elements and shared interactive screens, typically in a *tabletob* or *whiteboard* form factor. Two recent research works [197] and [76], explore such domain: the combined use of pen and touch on a tabletop multi-touch screen opens a number of possible new interaction techniques that borrow the manipulative experience and skills that people generally have from pen+paper usage, and projects them onto their digital counterpart.

Augmented Everyday Objects aims at augmenting with digital capabilities the tools of the real world, either to create new perspectives for use or to bridge the gap between physical and digital information. As an example, the Mediacup [11] is a coffee mug augmented with a number of sensors for temperature, movements, whether the cup is placed on a surface or held in one's hand, etc. This allows to implement several scenarios regarding the interaction with intelligent appliances, meeting schedulers, activity viewers, etc.

The next chapter will present a unifying theory that gathers under a unique perspective the common aspects of gestures and manipulations. The shared link is the inextricable relation that binds together action and perception, hand and mind, generally known as *embodiment*.

Chapter 6

Towards a Unifying Framework

More than 25 years ago, talking about *collaborative virtual environments*, Benford and colleagues commented

> [...] it appears that many collaborative systems still view users as people on the outside looking in. [12]

Apparently such view is still dominant, although challenged by advances in tangible interaction, computer supported collaborative work and pervasive computing. An alternative point of view considers the interaction situated *in time* and *in space*, and any dichotomy between action and context as purely artificial, i.e. the users participate in the interaction, either by means of their physical bodies in real space, or by means of suitable *embodiments* in virtual space.

We all experiment such embodiment in everyday life, and posses highly effective skills, some of them innate, some others painfully learnt, to cope with real world issues. Goodwin [65] discusses in great depth the dependency of social interaction on the mutual and systematic recognition of what is happening and what is likely about to happen. People participating in interaction articulate their contribution by simultaneously deploying several communication channels (or *semiotic fields*). Goodwin provides an example of children playing *hopscotch*: one of the children is challenging a move of the other. In doing so, she addresses multiple, redundant verbal messages and physically engages the other with her body.

In such a scenario: (*i*) the speech, specifically the choice of lexicon, characterizes the message thrown; (*ii*) the overall syntactic structure in which the message is embedded reinforces the message; (*iii*) prosody underlines salient aspects of the discourse; (*iv*) the spoken message is additionally embedded in

a particular activity, that constraints the possibilities for action; (v) the talk happens in a specific framework of interaction (a mutual position of the bodies), that the two children *actively* participate to create. It is this shared framework that makes the adoption of a certain sign system possible.

Goodwin shows how the *context* in which interaction occurs cannot be separated by the action itself, in that it is continuously and actively shaped and negotiated by the participants. The mutual position and orientation of the bodies determines the meaning of the signs and the possibilities of adopting a certain communication channel; for example, if one of the two participants would turn her back on the other, she would *at the same time* throw a very specific message through *body language* and close a potential communication channel (the gestures). The example of the two children playing hopscotch is representative of how sophisticated and effective human interaction is, specially when compared to the humble projection that we manage to implement in manipulative and gestural interactive applications.

Starting from a different perspective, Hostetter and Alibali [82] explore the mechanisms that give raise to gestures and suggest that *representational* gestures[1] arise spontaneously from the involvement of visual and motor mental imagery in the cycle of action/perception: they call this phenomenon of *Visible Embodiment*: Gestures as Simulated Action. Hostetter and Alibali provide a thorough literature review to sustain their hypotheses:

embodiment of language: the understanding of sentences appear to trigger the sensorimotor affordances of the objects or events spoken about;

embodied cognition: there are both behavioural and neural evidences that viewing an object appears to trigger the actions necessary to grasping that object;

embodiment of mental imagery: a mental image is the representation of an object, in the absence of the stimulus that is connected to that object. For example *visual mental imagery* is the experience of viewing a form (let's say, in one's mind) in absence of the relative visual input. *Motor mental imagery* is the representation of our body in action, and is involved in planning motor action.

There is strong evidence that the actions of thinking about, talking about, using, grasping, manipulating an object rely (partly) on the same mental processes (and activate the same brain regions). This is true both for actual manipulation of real objects and for mental representations (mental imagery). Hostetter and Alibali conclude that representational gestures stem from concurrent use of visual/motor mental imagery (simulated action) and oral/motor system.

[1] that is, *movements that represent the content of speech by pointing to a referent in the physical environment (deictic gestures), depicting a referent with the motion or shape of the hands (iconic gestures), or depicting a concrete referent or indicating a spatial location for an abstract idea (metaphoric gestures)* [82]

In human computer interaction, or more precisely, in computer supported and computer mediated collaboration, embodiment involves coping (at least in part) with such complexity by providing the user with appropriate tools and support for making sense, reshaping and *using* in a broad sense, the physical or virtual interactive space.

Robertson provides a comprehensive framework for understanding embodied actions in the domain of distributed (i.e. geographically dispersed) collaborative design [164]. Embodied actions are practices publicly (i.e. visibly and explicitly) deployed in a shared (work)space, ant thus therein available to the perception of all participants to a given activity. Robertson observed in a field study how the team used to deploy such actions during co-located group-work, and how technology, *together with specifically evolved work practices*, was used to support communication in remote work.

The classes of embodied action observed in practice included *individual* as well as *group* activities. The former range from drawing and writing, moving and manipulating objects, emitting an monitoring signals, enacting various aspects of the use/design process (such as while pretending to be the user), moving in the space, pointing at a particular object or position, etc. Group activities include communicating or organizing communication resources, focusing group attention, looking at the same thing at the same time, negotiating changes of subject in the talk, etc. All these actions are crucial in face to face group-work, and effective computer-mediated collaboration should in principle support them by means of appropriate *embodiments*.

The next sections discuss two aspects in which the concept of embodiment is of particular relevance: *embodied interaction* and *interactive spaces*.

6.1 Interactive Spaces

Back in 1991 Mark Weiser envisioned *the computer for the 21stcentury* as an ubiquitous network of highly specialized devices, so pervasive as to disappear, not only from attention, i.e. in metaphor, but *in fact*, embedded in badges, clothes, appliances, etc [184]. *Ubiquitous computing* was later refined into several, partly overlapping, research fields. *Pervasive computing* is generally regarded as an equivalent term for ubiquitous computing. The two threads have leaded for more than a decade separate conferences, although the communities that attended them were *largely* overlapping. *Ambient intelligence* focuses on environments capable of sensing and reacting to the presence and actions of people; the *Internet of things* thread explores issues and potentialities of connecting real (typically everyday) objects to the Internet;

When the vision of ubiquitous computing was first shared by Weiser, the state of the art of hardware and networking barely allowed to create working proof-of-concepts of the many tools and gadgets. Satyanarayanan, 10 years later, pointed out a research agenda for pervasive computing [167], underpinning

its roots from distributed systems and mobile computing, and further stressing on 4 key aspects:

Smart Spaces created from a tight coupling of building and computing infrastructure, as to obtain responsive environments, capable of (for example) adapting to the profiles and preferences of its inhabitants;

Invisibility of machinery, although loosened with respect to Weiser's radical vision, to a more practical *minimal user distraction*;

Localized Scalability, i.e. the capability of the smart environments and networked computers to calibrate their interaction demands depending on the users distance or activities;

Masking Uneven Conditioning, i.e. the system takes care of covering as far as possible the changes in service availability, for example by caching relevant information, deferring synchronization, etc.

This rather *technology-centric* vision was perfectly suited for driving the technological progress that just in those years was about to put one smart-phone in almost every pocket on Earth. From our point of view it fails to address the richness of human skills and expressions. Further on, Satyanarayanan addresses the problem of *context awareness*:

> A user's context can be quite rich, consisting of attributes such as physical location, physiological state (e.g., body temperature and heart rate), emotional state (e.g., angry, distraught, or calm), personal history, daily behavioral patterns, and so on. [167]

Henricksen and colleagues [74] develop this model including location, device in use, communication channel, personal information and activity descriptions. As said above, such view of the context as a *boundary condition*, although useful in the design of adaptive and personalized services, doesn't have in our opinion the power to inform the design of *natural interaction*. Dourish explores in great detail such dual nature of context, underpinning the incompatibility of the *positivist* interpretation, largely adopted in the Ubicomp community, and the *phenomenological* interpretetion, more common in the CSCW community [45]. If the former considers the context essentially as a datasource available to the system (and thus *describable*, relatively *stable*, and *distinct from activity*), the latter considers context as an emergent property of interaction, and *contextuality* as a *relational property* (and then *dynamic*, *occasioned* and not clearly separated from the activity, but *actively produced, maintained and enacted in the course of the activity at hand*). Dourish concludes that:

> the major design opportunity concerns not use of predefined context within a ubiquitous computing system, but rather how can ubiquitous computing support the process by which context is continually manifest, defined, negotiated and shared? ([45], p. 26)

On the contrary, the aspects of interaction *between people* and the continuous and dynamic negotiation of resources are often neglected or taken for granted in pervasive computing scenarios. Vogel and Balakrishnan [181] in their work on shared public displays, propose a framework to identify 4 interaction phases, which the user fluidly and continuously passes through: *ambient display*, *implicit interaction*, *subtle interaction*, and *personal interaction*. However, the possibility of multiple users interacting with the system at the same time is regarded as a special case, and mutual interferences as problems to cope with, rather than a normal aspect of cooperation.

Analogously, in their work on *proxemic interaction* Ballendat and colleagues describe the potentialities of devices made aware of people's *position, identity, movement* and *orientation* [7]. They describe the characteristics of a system capable of reacting meaningfully to changes that happen within the above interaction *dimensions*: for example a video is paused when the user distract his attention. Ballendat and colleagues consider

> the complete ecology present in a small space Ubicomp environment[...]: the relationships of people to devices, of devices to devices, and of non-digital objects to people and devices.
> [...]
> Proxemic interactions also extend beyond pairwise interaction and consider one person or multiple people in relation to an ecology of multiple devices and objects in their nearby environment.

What is left out of the picture is the relationship of people to other people. On the other hand, when the interaction between people is given appropriate emphasis, it happens to play a key role in the use of the system too. Peltonen and colleagues [152] describe their experience when deploying a large shared display in the city center of Helsinki. They report how the presence of other users is crucial for encouraging the approach.

> Visibility of the screen is not merely a sum of its physical properties. As the urban landscape is already full of visual clutter, people appear to be more attentive to other people's behaviour there

Moreover, people negotiate the use of the system making their own presence gradually visible to the others, in a stepwise manner. Peltonen and co-workers report a number of examples of cooperative use of the display, including a case of *playful conflict* in an attempt to gain simultaneous access to a same interface element (hence the title of the paper: "*It's mine, don't touch!*").

A description of the technicalities involved in the implementation of interactive spaces is beyond the scope of this literature review, but will be included in the chapters that report on the technical achievements in Part 3. The next section presents the framework of *embodied interaction*, that gathers under a coherent perspective many of the concepts presented so far.

6.2 Embodied Interaction

Dourish presents the concept of *embodied interaction* in his highly influential book [46] as the common ground upon which tangible and social computing are based. The key aspect of tangible computing is that interaction takes place *in the real space* and by means of natural (manipulative and sensory) skills. Dourish underlines that even if many computing paradigms use the real world as a *metaphor* (for example the Desktop metaphor), a characteristic of tangible computing is to use the real world as a *medium* for interaction. The emphasis is on the *active participation* of the users:

> Even in an immersive virtual-reality environment, users are disconnected observers of a world they do not inhabit directly. They peer out at it, figure out what's going on, decide on some course of action, and enact it through the narrow interface of the keyboard or the dataglove, carefully monitoring the result to see if it turns out the way they expected. Our experience in the everyday world is not of that sort. [...] We inhabit our bodies and they in turn inhabit the world, with seamless connections back and forth. ([46], p. 102)

Similarly Dourish stresses on embodiment as a foundation for social computing: users participate in the interaction by means of their physical bodies in the real world, and doing so they participate in the complex network of social interactions of the everyday life. Social computing is then HCI seen through the lens of the social scientist, with a key role generally recognized to field work and ethnography, i.e. to the study and description of *practices*.
Thus, Dourish indicates in phenomenology the common theoretical ground, and in embodiment the practical *trait d'union* between interaction design informed by social studies and tangible computing.

Hornecker and Buur [80] also suggest a viewpoint on tangible interaction strongly focused on embodiment, and encompassing: *tangible* manipulations and *physical* representations; *embeddedness*, either in real space or in augmented reality; use of body movements as an *essential* aspect of interaction. They emphasize strongly how such approach overcomes the perspective of *interface* design, focusing instead on the design of the *interaction*, by exploiting the richness of bodily movements, and propose a framework for understanding tangible interaction, articulated in 4 themes:

Tangible manipulation covers the material and tactile qualities of the artefacts that are manipulated in tangible interaction;

Spatial interaction insists on the fact that tangible interaction happens in the real space, and thus implies movement in the space;

Embodied facilitation copes with the organization of space and artefacts, and how the possible configuration affect the possibilities for interaction;

Expressive representation examines how the tangible representations are created, shared and linked to their meaning;

What is particular relevant of this approach to our work is the emphasis on the activities, rather than on the artefacts, and specially on the collaborative aspects of such activities. Hornecker and Buur present several case studies that expose how apparently trivial or marginal design decision can heavily affect the *group activity*.

Taken further, a key issue in natural interaction design is that the effectiveness of the interaction between people (i.e. *as a group*) is not a direct consequence of the effectiveness (or usability) of the single artefacts. Too often the evaluation of collaborative technologies fragments the overall interaction in single-user interaction bursts, taking the negotiation, conflict management, interpersonal communication, etc. out of the picture.

The remain of this work will describe our experiments in the design of a space for work-group, the lessons learnt from the design, evaluation and redesign of every piece of technology, and a final evaluation of a collaborative activity conducted in the interactive space, and aimed at assessing the quality of cooperation, and understanding its practical impact in a typical office scenario.

> Many of the most exciting new research and development in computing will not be in traditional areas of hardware and software but will be aimed at enhancing our ability to understand, analyze, and create interaction spaces.
>
> Terry Winograd

Part III

A Space for Groupwork

Chapter 7

Interactive Space

In principle it is possible to experiment with interactive environments without building any working prototype. Using methods such as the *Wizard of Oz* [39] it is possible to involve potential users in testing very early interface designs.

In a nutshell, the Wizard of Oz technique consists in letting the user believe that he or she is using a fully functional system, while in fact some (or all) features are simulated by a human operator; not surprisingly one of the first, and still more popular, application field is with natural language based interfaces. Nevertheless, this approach has been exploited in a number of studies regarding gestural input (e.g. [29, 52, 73, 83, 166]).

However, the Wizard of Oz technique has severe limitations as well as advantages. First, there are ethical issues that arise from deceiving the users that participate in the test. Basic ethical considerations in HCI research prescribe to thoroughly brief the users about the goals and methods of the study, *before* the test begins.

Other concerns are of more technical nature; according to [59]

1. it must be possible to simulate the future system, given human limitations;
2. it must be possible to specify the future system's behaviour;
3. it must be possible to make the simulation convincing.

Actually, a Wizard of Oz simulation of a complex manipulative interface could be just as challenging as the realization of a working prototype. Additionally, as some authors point out, it is sometimes preferable to focus on the opportunities and limitations of an existing system, rather than on the desired features of an imaginary one.

In a recent article on the *Interactions* journal, Bill Gaver reports:

> We call the things we make "prototypes", but they're actually highly finished one-off products. We spend time crafting the details of our designs - the exact form and color of a casing, the timing of a graphical transition - often making several iterations to ensure that our prototypes are highly robust and easy to use. We do this for two reasons: First, so our prototypes can be deployed to people and treated as products rather than lab demos, and second, because we love it. [62]

The second reason doesn't claim for further clarification, the first one deserves some thought. Experience shows that to get engaged in interaction with a new *device* users need a degree of comfort that rarely is attained in lab demos.

Figure 7.1: The interactive space: notice the multitouch table and the wall sized interactive displays; camera are arranged in the space for capturing the activities for research purposes

If flickering, jerkily devices mounted on unsteady frames can be suitable *proof of concepts* to show to fellow researchers, testing with users can require solid structures, fluid animation, robust software and a overall *polished* aspect; in most cases fewer but solid features are preferable, both for hardware and software, if we want to observe a truly *natural* behaviour.

My first attempt at building a such device was inspired by the famous presentation by Jeff Han at TED[1]. What caught my enthusiasm (as well as that of many others, I guess) while watching at Han playing as a technology *juggler*, was that it looked powerful, easy, natural, but most of all, it looked *fun*.

In the following months I had to learn, to my own cost, how challenging it is to generate a similar excitement in the potential users of a new technology. My

[1] available online at
`http://www.ted.com/talks/jeff_han_demos_his_breakthrough_touchscreen.html`

goal was to create tools capable of engaging users in playful interaction in order to observe and make sense of their behaviour, motivations and expectations; a first prototype of a multi-touch interactive wall [172] was presented at the international exhibition CeBIT (in Hanover, Germany) in 2008.

After that, several other devices were added to form a fully functional interactive space (see figure 7.1):

- a larger interactive wall, multitouch sensing is based on bezel cameras and the display is composed of a tile of commodity projectors [122];
- a FTIR multitouch table whit several improvements to allow pre-contact feedback and robustness to changing lighting conditions [87];
- surveillance cameras to monitor the activity in the experimental area and to sense motions/gestures and use of fiducials for tangible interaction;
- video-cameras, multitouch table and the interactive wall are handled by means of a custom software framework, designed almost from scratch [42] that provides software abstractions over physical sensors, abstract and concrete multitouch widgets (supporting manipulation of images, web pages, movies, etc.), fiducials recognition and tracking, and a distributed asynchronous event subscription/delivery architecture;

The following chapters describe the technology, with particular emphasis on original contributions to the state of the art, as well as the experiments that have been carried on during and after the development, and how these experiments changed my understanding of HCI.

Chapter 8

Manipulative Interfaces

In the following sections we present a research experience of manipulation and exploration of multimedia contents. The *Troll*, as we dubbed it[1], was made from the combination of a paper brochure and a LCD display. Browsing the brochure the user can access multimedia contents that are displayed on the screen; the animations and movies are related to each specific page of the brochure, that in turn reports a description of our projects. We show the rationale behind the exploitation of TUIs in the specific design space of the demo corner of our research lab, discuss the implementation and evaluation with users of a working prototype, and suggest a broader viewpoint in the context of natural interaction.

8.1 Exploring Multimedia Contents

Though information technologies have faced an impressive evolution over the last decades, the way we access and explore multimedia contents (and digital information in general) has not changed much after the widespread adoption of graphical user interfaces.

Ironically, the typical metaphors of desktop computing, such as windows, icons etc. are making their way also to more personal and informal scenarios, such as home life, TV sets, mobile phones. However, extensive research is being carried out in order to overtake such paradigms in favour of more natural and unobtrusive models of interactions, especially in the vast and growing community of tangible interaction [93].

TUIs have already been thoroughly presented in part 2, when applied to the exploration and creation of multimedia contents, TUIs are about binding the digital world of audiovisuals to the physical world or to graspable representations of virtual entities.

[1] With no particular meaning or reference to the mythological creature.

8.1.1 Related Research

The practice of binding multimedia and TUIs has a long history and has been explored in many forms. Initial research may be traced back to the early ninety's to the work of Wellner [186,187] on the DigitalDesk: a system meant to augment paper documents with computer generated data. This is one of the first attempt of supporting a tangible interaction with digital media, although a formalization of the TUI paradigm was still to come; so Wellner's work concentrates on letting users easily alternate the work on digital and physical documents.

The Listen-Reader [4] is a book which pages act as controls for the playback of sound effects related to the narration. An interesting aspect of the Listen-Reader is that it emphasizes continuous gestures (like moving one's hand towards the book) versus discrete ones (such as pushing a button). However the particular choice of gestures didn't exploit the specific affordance of the paper book and so (in the authors' accounts) the interaction wasn't as natural as expected. Early work of Fitzmaurice [53, 54] on Graspable User Interface and then the work of Ishii [93] on TUIs informed subsequent research.

A key aspect of tangible interaction is that it allows people to actively explore and make sense of the world (either physical or digital). Examples of systems that allow tangible interaction with multimedia contents include [115], [133] and [66].

Differences with these and progress respect the state of the art are presented below.

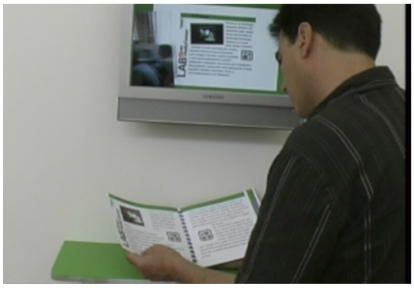
8.1.2 Implementation

The goal of this activity was to design a multimedia kiosk to, at the same time, inform visitors about our research activities and let them experience a simple but representative example of tangible interaction.

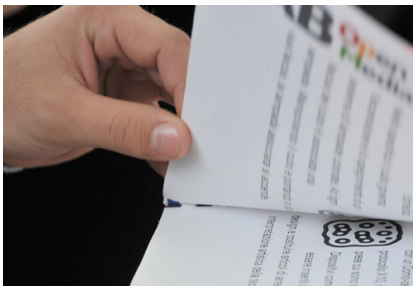
In our view, engaging people in active exploration represents at once the best way to let others know about our activities and a valuable chance for us to get users' feedback: a key point is that the exploration of the multimedia kiosk is only a part of a broader interactive experience that is the visit to the lab.

The multimedia kiosk was made coupling a paper booklet and an LCD screen (see figure 8.1(a)); a digital camera hidden behind the screen monitors the booklet and reacts to page changes. The multimedia contents displayed on the screen are relative to the page of the booklet that the used is reading.

As shown in Fig. 8.1(b), each page of the booklet bears a visual tag that encodes the page number. The recognition of such tags is performed by means of the reacTIVision framework [99], which is capable of distinguishing several hundred different tags, their position in the frame and rotation on the plane. To avoid accidental occlusions the same tag is replicated on both the left and right page of the booklet.



(a) the *Troll*



(b) booklet

Figure 8.1: The *Troll* in its final installation site (a) and a detail of the booklet that acts as a *tangible* controller (b)

A system was then built to coordinate the playback of multimedia contents while users browse the booklet. Whenever a new tag appears (the user turns a page) an animation starts that gently slides away the current movie (from right to left if the new tag, i.e. the page number, encodes a number that is higher of the current one, from left to right otherwise, which is consistent with the manipulation of the booklet).

This basic manipulation actually addresses most of the practical cases, however, from usage evaluation, we found that some further features could have been provided:

- if two different tags appear at the same time for a given interval, the user is hesitating in the middle of turning a page, there are good reasons for reacting to this situation, since in most case he/she is seeking feedback on the system's part (see later);
- when tags appear upside-down the user has rotated the booklet, in general as a mean of understanding how the system works; again appropriate feedback should be provided. Of course many actions may prove appropriate in this case depending on the application; if there's a map on the booklet, rotating it has a different meaning that if there's only text on it.

Clearly, a constraint for the system to function is that the booklet must stay under the camera. This turned out to be not an issue, though. Actually, people are willing to accepts some constraints in the use of a tool, as soon as they understand why these constraints exist: a remote control will not work through a bricks wall or from behind a blanket (although it may, adding some, not so expensive, technology) and this is in general regarded to as a feature of the tool, rather than a problem.

Our first hypothesis (confirmed by later observation) was that the occasional visitor would, more likely, be engaged by a familiar tool such as a book than by a keyboard or by a touch screen.

The following step of understanding how to operate the system was, however less straightforward. Ideally, while browsing the booklet, the attention of the user should have been captured by the animations playing on the screen but in our first prototype this didn't always happen. We realized that this was due to the fact that while reading the booklet the attention of the user is focused on this, and little attention is paid to what happens on the screen especially if unexpected such as in this case. To address this problem we had to play both with the ergonomics aspects (placing the screen next to the booklet so that it falls within the peripheral view of the user) and cognitive aspects (stressing on the coherence of interaction and system feedback).

The system has been deployed at the demo corner of our lab, and, in the following months, has been used by many visitors. We conducted usage evaluation passively observing people engaged in interaction and, occasionally, interviewing users that showed particular interest or difficulties (further discussion is presented below).

8.1.3 Evaluation

We discuss here how people managed to use our system, what insights we did get about the benefits of Tangible Interaction in the context of multimedia fruition and trace some further research on the basis of these results.

First of all, as mentioned before, it is important to notice that the exploration of the multimedia database is only part of a wider interactive experience: a phase of the overall visit of the demo corner and the lab. As such it was designed to be as unobtrusive as possible, that is, visitors are not meant to be absorbed by the presentations, but rather to engage in them and, at the same time, be aware of the mood of the place, engage in conversation, ask clarifications, etc.

This goal was met thanks to the specific nature of the booklet medium that allows a superficial browsing as well as deeper reading. On this basis our observations and interviews showed peculiar behaviours; the feeling of natural interaction was due to the coherence between the animations on the screen and the action of turning the pages. This was not surprising and adheres to common understanding of action and feedback.

However the choice of a continuous gesture (in contrast to a discrete one such as pushing a button) reinforces the feeling of tangibility, and most important, adds to the accountability of the overall system.

This idea of accountability indicates, as in [44] that the interface is designed in order to *present, as part of its action, an account of what is happening* ([46], p. 84). Since the understanding of such account is, in this case, shared among all the visitors and the researchers of the lab, which, in different roles, participate in the visit, the interaction is not a private matter that involves the system and the user, but is (consciously) one of the actions that take place in the environment, and can easily coexist with others.

A behaviour that we observed several times was the attempt on the users' part to freeze the animation of the multimedia contents by half-turning a page on the booklet: users begin to turn a page, then pause before completing the gesture, as if they had changed their mind.

The expected response was that the system paused (and eventually rolled back) the animation. Apart from arguing on the usefulness of such functionality, in our view such behaviour appears related to the accountability of the application: from the users' point of view if the system is aware of one's hesitations in the interaction (and reacts accordingly) there's more chance to learn its use by trial and error (i.e. actively exploring its functionalities) without reaching an unrecoverable dead-end.

Our initial mistake was to consider the action of turning a page as an atomic one, whereas users consider it as a continuous transition that can be paused and withdrawn at any time.

8.2 Lessons Learnt

The experience described in this chapter was quite significant for two distinct aspects. In the first place it originated the idea of *playful interaction* that was to guide most of the experiments that come later. We put great efforts in the graphical design of the booklet, as well as in the animations that appeared on screen. Our visitors rewarded us with plenty of useful feedback, insights and suggestions for improvement, that ultimately led to an improved installations.

A second important remark is that many people used to spend a few minutes exploring the technology to make sense of how it worked. As said above, the digital camera that monitors the booklet and triggers the animations is *hidden* behind the LCD display. This was initially done as a deliberate choice, aimed at fostering the curiosity of the users, but was also subject to debate during the evaluation: should the technology be hidden to the user?

Of course there's no general rule and each case is different. However it seems that sometimes the concept of *invisible computer* is taken too literally in the Ubicomp related communities. In this case (as in others, see for example [10] and [79]) users manifested the desire to understand the system *before* exploring the information in it.

This phase of familiarization is often regarded by natural interfaces researchers as a necessity to overcome (possibly) or to minimize: the idea underlined is that of an interface that users operate without even being aware of it.

A second possibility, more fruitful in our view, is that natural interfaces should account for their behaviour and support such phase of sense-making, and the visibility of the technology is often a facilitating factor on this path. In the following experiments we didn't try to hide the artefacts, but instead

managed to give them more visibility, in order to facilitate appropriation on the users' part.

Acknowledgement: this chapter is a revised version of the papers:
A. Soro, M. Deriu, and G. Paddeu. 2009. Natural exploration of multimedia contents. In Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia (MoMM '09). ACM, New York, NY, USA, 382-385. [170]

Chapter 9

Tabletop and Wall-sized displays

Shared displays are often considered a *must-have* of interactive spaces. In a recent video concept, *Microsoft at 2020* [137] almost every surface available is turned into a multitouch capable device. Boarding passes give navigation hints through the airport, interactive walls and desks open a view on virtual or distant worlds, etc.
An analogous vision is expressed by Corning Incorporated (world leader in specialty glass and ceramics): in the video concept *Future Technology: Watch your day in 2020* [89] people wake up into a technological house that has every piece of glass or ceramic turned into an interactive display.
Nokia envisions a future in which wearable devices, such near-to-eye displays, coupled with eye-gaze and gesture recognition enable interaction in augmented reality [142].

These (and other analogous) videos have drawn severe criticisms. Bret Victor, formerly designer at Apple and now *purveior of impossible dreams*, published a *Brief Rant on the Future of Interaction Design* on his Website [180], attacking the multi-touch interface that he calls *Pictures Under Glass*. In his own words:

> We live in a three-dimensional world. Our hands are designed for moving and rotating objects in three dimensions, for picking up objects and placing them over, under, beside, and inside each other. No creature on earth has a dexterity that compares to ours.
> The next time you make a sandwich, pay attention to your hands. Seriously! Notice the myriad little tricks your fingers have for manipulating the ingredients and the utensils and all the other objects involved in this enterprise. Then compare your experience to sliding around Pictures Under Glass.
> Are we really going to accept an Interface Of The Future that is less expressive than a *sandwich*?

A second, but maybe more important, aspect is again collaboration: most of the activities featured in the the videos are only fancy and animated flavours of web browsing and email: people are apparently more interested in checking the weather forecast than in talking to each other. These and other considerations, such as the top-model-looking characters and the luxury penthouses in which most of the action takes place, suggest that one of the goals of such videos (if not the only one) is to impress the executives rather than depict a visionary scenario.

In out view, as remarked before, a key aspect of multi-touch displays is the possibility for many people at once to use them *in cooperation*. We have focused then on the implementation of a videowall and a multitouch table. Both are wide enough to lodge at least two people, and potentially more, that can comfortably collaborate at specific tasks. In this chapter we explain the technical challenges that we had to face, and report on our solutions.

t-Frame is a hardware/software architecture that allows the implementation of multi-user interactive walls. t-Frame brings multi-touch sensing to a generic display by means of low cost digital video cameras. The design of t-Frame is illustrated in detail, together with a prototype installation.

Systems suitable for multi-user interaction (either collaborative or competitive), require almost invariably the adoption of projectors tiles. However projectors always displays a deformed image, due to lens distortion and/or imperfection, and because they are never perfectly aligned to the projection surface.

Multi-projector video-walls are typically bounded to the video architecture and to the specific application to be displayed. This makes it difficult the development of interactive applications, in which a fine grained control of the coordinate transformations (to and from user space and model space) is required. We have designed a solution to such issues: implementing the blending functionalities at an application level allows seamless development of multi-display interactive applications with multi-touch capabilities.

Finally we report on our improvements on the FTIR technique, that allows the implementation of a multi-touch robust to adverse lighting conditions, as are often found out of the lab. We show how our approach differs from others and discuss our findings, how it influenced our subsequent research, and lessons learnt.

9.1 t-Frame Interactive Wall

Interactive walls are a special kind of computer applications that deliver a highly impressive, shared view of information, and are suited to many exciting applications, ranging from work-group collaboration to pervasive computing and entertainment.

t-Frame is a low-cost hardware/software architecture that enables multi-touch interaction on a generic display. Specifically, t-Frame is intended to be used in large size, multi-user interactive walls. With respect to other approaches t-Frame can be installed as a touch sensor device on any flat surface regardless of size, shape and display technology. t-Frame allows researchers in the field of human-computer interaction to set up with minimal effort an environment for experimenting in the field of computer supported collaborative work and tangible user interfaces: the goals of the project can be summarized as follows:

1. provide multi-user and multi-touch interaction to any display, regardless of the specific technology of the display itself;
2. minimize both the cost of installation and maintenance, using standard hardware and simplify the installation and calibration procedures;
3. be applicable to very wide installations: modern hardware allows the creation of interactive walls several meters long using cluster systems or multi-head graphic adapters.

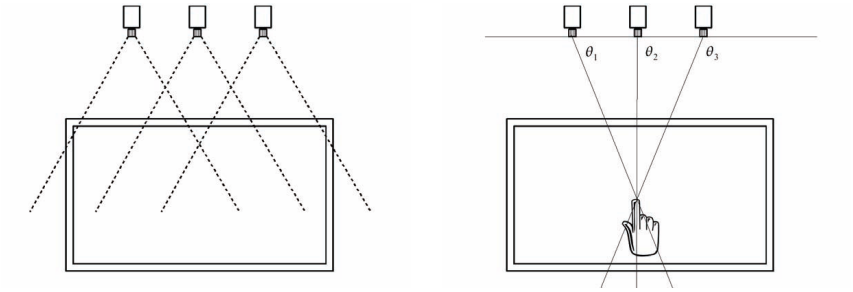
At the moment we designed the t-Frame system, the most adopted techniques to implement a multi-touch displays were based on Frustrated Total Internal Reflection (FTIR), Diffuse Illumination (DI) or Diffuse Surface Illumination(DSI) [71] [169], and required the adoption of expensive high resolution IR cameras, ambient IR screening and was in practice bounded to rear projected screens.

t-Frame requires less space than other technologies, can be easily transported because it does not require a single-piece touch surface (the first prototype implementation described in this paper consists of a LCD display tile, although any other display solution can be exploited) and, most important, can adapt to the size of the display seamlessly, without sensibly affecting the performances.

9.1.1 Related Research

As already mentioned, pioneer work on multi-touch sensing devices can be tracked back to the mid eighties, see for example [101] [124] [119]. An overview of the evolution of multi-touch technologies is maintained in [25]. Given that multi-user interaction is a straightforward extension of multi-touch sensing, the obvious playground in this field consists in displays capable of accommodating a number of users, such as tabletop and wall-size displays. [43] and [191] are examples of the former, [190] and [40] of the latter.

Several techniques have been exploited to implement multi-touch sensing devices: [43] consists of an array of antennas whose signals get transmitted, through the body of the user, to a receiver that elaborates touch events. Among optical techniques [190] exploits stereo cameras to compute hands position, but the cameras are located behind the semi-transparent screen, thus the system is bounded to front/rear projected display.



(a) video cameras are arranged along one edge of the screen

(b) the position of a point of touch is triangulated taking into account the input from each camera

Figure 9.1: The overall setup of the *t-Frame* system, and a schema of the triangulation algorithm

The same holds in [71], which relies on an infrared camera that captures the light that escapes the display surface when finger contact occurs. In [144] the optical sensor is located above the display surface, and thus the hands of the user(s) stay between the camera and the screen. Although feasible for the interactive desk described, this approach would be not practical for interactive walls, since the body of the user would in general cover the hands from the viewpoint of the camera.

The system described in [41] exploits an approach similar to t-Frame: the cameras are arranged around the screen and the position of the fingers is determined through triangulation, but the cameras are located on the corners of the screen: in t-Frame the particular arrangement of the cameras limit the complexity of the algorithms involved in finger triangulation and allows the system to scale in size almost indefinitely.

9.1.2 Technical design

A t-Frame installation consists of a set of cameras arranged on the plane of the display. In a typical installation of an interactive wall, cameras are aligned on the top or bottom edge of the screen, facing down or upwards, as shown in Figure 9.1(a). However, cameras are not bounded to a fixed position or orientation, and can be arranged anyhow, as long as they lie on the same plane of the screen and the exact position and orientation of each camera is known with respect to screen coordinates. In order to simplify the setup of the installation t-Frame provides a calibration facility that computes the exact position of each camera, this operation involves three steps:

1. every camera takes a snapshot of its field of view, no touch must occur during this process, and saves it as a known background;

2. the user is requested to point an horizon in the background image of each camera: the horizon must be specified as close as possible to the surface of the screen;

3. the user is requested to point with her finger three given points on the screen: the position of the cameras is triangulated exploiting the images captured at each touch.

Periodically the images are compared to gather touch events that are then pushed in an event queue that applications can poll and consume. In the following the two most critical steps of t-frame are presented.

Finger Triangulation The frames captured by each camera are elaborated to spot touch events. A frame is compared against the known background: when a significant difference is found, the algorithm assumes that the background is covered by a finger touching the screen (see Figure 9.2) and measures its position. The position of the finger is reported as the angle under which the finger is seen with respect to the center of the field of view. To do this the exact aperture of the field of view must be known, since we only can take measures in terms of relative positions, i.e. counting pixels in the image. Figure 9.1(b)

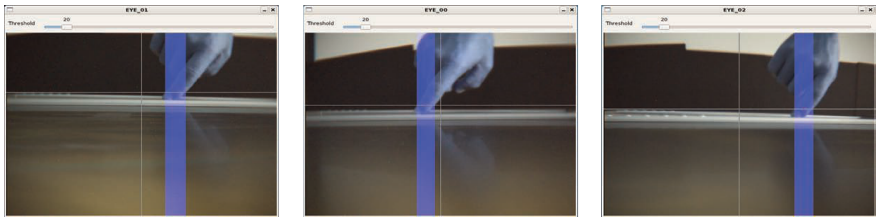


Figure 9.2: Images captured from different video cameras provide differentiate views that allow finger triangulation

shows how the exact position of a finger is computed in screen coordinates: when a finger touches the surface of the screen it is seen by every camera whose field of view covers that position. Additionally every camera sees the finger under a given angle.

Then, computing the position of the finger is as easy as calculating the intersection of two lines passing through the position of any given camera (that is known from calibration), and intersects the axes of the screen under the given angle, which is a matter of bare trigonometry. With a single touch and only two cameras this approach doesn't differ from stereo vision, but by exploiting several cameras t-Frame can easily recognize an arbitrary number of touch points.

Multitouch disambiguation Consider the situation represented in Figure 9.3(a). The user touches the screen with two fingers: each camera sees its

background covered by two distinct obstacles and computes two angles. We can therefore trace two lines for each camera; the intersections of such lines determine the exact position of the fingers, but we need a strategy to exclude false spots. To this end the algorithm of multi-touch disambiguation compares every intersection (that is a candidate finger) against any camera that has that specific point within its field of view, and checks if the candidate finger is compatible with the image seen. In the case of Figure 9.3 the candidate finger

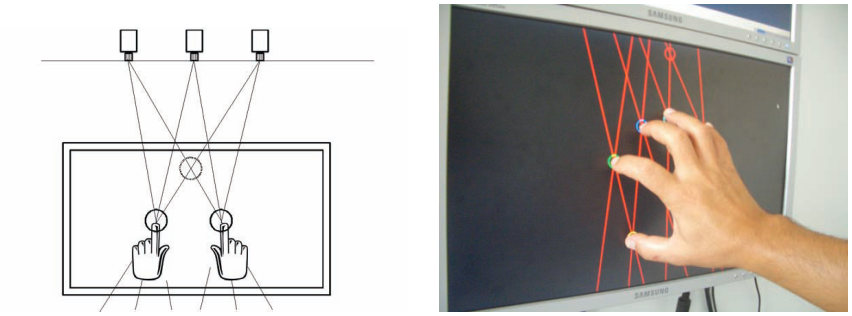


Figure 9.3: The algorithm of multi-touch disambiguation exploits redundant camera to cope with possible occlusions

highlighted with a grey circle is the result of the analysis of the frames taken by the leftmost and rightmost cameras. In order for this to be accepted as a finger-touch, it must be confirmed by the frame taken by the central camera. The comparison against the central camera shows that this is a fake finger, as expected, and that only the two spots highlighted in black circles represent true finger-touches.

The number of cameras needed to disambiguate a given configuration depends on many factors. In general with N cameras we can always disambiguate N-1 finger-touches, but in practice fewer cameras are enough to monitor many finger-touches. In fact, in many cases, false candidates fall (largely) outside of the display area and can be excluded.

Additionally finger-touches tend to follow predictable paths, and heuristic techniques can be exploited to distinguish false candidates. It can be shown that the complexity of the algorithms is linear with respect to the number of cameras installed, and depends primarily on the number of fingers that are seen at once by a given set of adjacent cameras.

9.1.3 Prototype implementation

The t-Frame design and algorithms have been tested initially through a prototype installation, consisting of a 60" wide display tile, driven by a high-end graphical workstation. The optical apparatus consisted of 9 VGA cameras (25 frames/sec).

The system performs very well in terms of efficiency, since the algorithms involved in the image analysis don't comport a significant overload. In addition, the exact computation of fingers' position only involves some basic operations.

Our experiments show that multi-camera triangulation based on background subtraction is a suitable technique for implementing a multi-user & multi-touch interactive wall. The event models that are common in computer systems, based on clicks, drags and so on, have been designed for single pointer operations. Although such a model can be extended to support multi pointer interaction (such as in MPX `http://wearables.unisa.edu.au/mpx/`) this cannot be other than a first step towards a new, commonly agreed, event model specifically designed for multi-touch widgets. However, in order to deploy t-Frame as

Figure 9.4: The first working prototype of *t-Frame*

collaboration platform for complex information retrieval and manipulation, we had to face many other technical issues, related to the setup of a low cost video-wall wide enough to accommodate several users. As shown in the next sections this proved a challenging issue.

9.2 Multi-projectors Screen

Interactive video walls require almost invariably the adoption of display tiles, made either of LCD screens or projectors. Choosing the first solution, although possible in principle, one has to face high costs, logistic hassles, high power consumption and heat emission, all factors discouraging to adopt it. It's true that, apart from this, arranging an array of multi-touch displays is barely a problem of setting up a scaffold holding it, and connecting the array to an appropriate hardware that supports a display of that size.

In the second case, when using projectors, the cost per surface unit is reduced, and the final result can be absolutely seamless, due to the absence of any type of frame inside or around the display. Our choice fell on this second solution, both for costs and logistics considerations. However, as we will discuss later on, low cost and higher flexibility is obtained at the cost of facing and solving the problem of blending, in term of geometry, colors and lightness, the images coming from each different projector. Figure 9.5 shows the enhanced setup of the t-Frame system based on two wide-angle Nec WT510 projectors mounted for overhead-front projection on an aluminium Trabes scaffold, of about 3 meters width.

Figure 9.5: The *t-Frame* interactive video-wall in its final, 3 mt wide, setup

9.2.1 State of the art

The blending problem is already theoretically solved by previous works as it is well summarized in [129] and practically implemented in many ways; these solutions mainly relies on hardware, using expensive projectors with in-hardware blending capabilities, or in software, typically bounded to the video architecture and to the specific application to be displayed, thus restraining the portability of the system.

A typical example of this is the Chromium framework, targeted to OpenGL

based applications [84]. This issue appears even more important in the development of multi-touch video-wall applications. As an example, coordinate transformation (from sensor space to GUI space) is affected by the blending functionality, and is better addressed if the blending is realized at the application level, rather than at the device level.

Here we describe our proposal to address such issues: implementing the blending functionalities at an application level allows seamless development of multi-display interactive applications with multi-touch capabilities. We chose to start from a well known and largely used user interface development toolkit as Qt [143] and to extend it adding geometric calibration and blending to the, Qt based, t-Frame multi-touch application framework already described.

9.2.2 Technical Design

A projector always displays a deformed image, due to spherical lens distortion and/or imperfection, and because it is almost never perfectly aligned to the projection surface. So, when multiple projectors are joined to realize a video-wall, this distortions are easily noticed, and just aligning projectors side by side is not a feasible solution.

This problem is well known as are several solutions [129], typically using a partial overlap between projectors, achieving first a geometrical calibration, and then applying a darkening mask to achieve the blending and hide the double lighting in the overlapping zones.

Interactive video walls have the additional constraint that users must be able to get close to the display (for interacting). Since front projection would cause the shadow of the user to be projected in the interaction area, such systems typically require rear projection or wide-angle front projectors.

As said, our setup adopts the latter solution (see figure 9.6), projectors are attached upside-down over the head of the users leading to a compact setup; however, the lens distortion is further enhanced by the wide-angle lenses; note however how the shadows projected by the hands of the users don't interfere with the interaction, thanks to the great asymmetry of the Nec wide angle projectors.

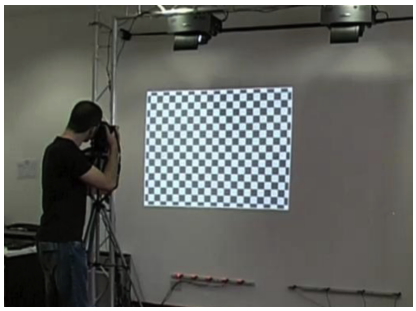
Geometric Calibration Projectors get calibrated one at a time: a black and white checker-board sample is captured by a camera positioned just in front of the projection (see figure 9.7(a). The camera itself need to be calibrated to avoid lens distortion (this task is easily done using OpenCV [175]). Tilt and orientation with respect to the display surface must be known. In absence of any distortion from camera and projector lenses the image projected on the screen and the one captured by the camera would have identical proportions. OpenCV allows to precisely determine the position of the internal corners of a

Figure 9.6: Users collaborating at the *t-Frame* video wall

chessboard pattern, and we use it to compute the deformation matrix for the projector.

The resulting (inverted) transformation is then applied just before the rendering phase: the application is not directly shown on the screen, but instead is captured as a texture (rendering on a Frame Buffer Object). Such texture is then divided in tiles that match the deformation matrix just captured. This compensates the projection distortion, and possible rotations, achieving geometric calibration inside each projector (intra-projector geometric calibration). Different (partially overlapping) areas of the model are then rendered separately, and each one is deformed according to the appropriate matrix before being rendered to the screen. In this way we achieve geometrical consistency between projectors, using this alignment to compensate the space wasted by overlapping projection regions. (extra-projector geometric calibration).

Blending Once we have obtained a geometrically consistent projection, we have to apply a darkening mask to obtain a luminance consistence. With the geometric calibration, we can assume that the projections are aligned, so we can apply a graduated shading that follows a double cosine function to the two edges of the windows that cover the overlapping zones. In this way, every point in the resulting projection has a constant sum of lighting (see figure 9.7(b)). Finally, the system is responsible for the transformation of user input (such as mouse events and multi-touch gestures) according to the transformed model.



(a) the operator captures the checkerboard for calibration

(b) comparison of the overlapping projections before and after luminance blending

Figure 9.7: Geometric calibration and blending of the multi-projector setup

9.3 Multitouch Table

Multi-touch tables are the natural companion for the interactive wall, just as desks are for blackboards. Frustrated Total Internal Reflection (FTIR) is a key technology for the design of multi-touch systems. With respect to other solutions, such as Diffused Illumination (DI) and Diffused Surface Illumination (DSI), FTIR based sensors suffer less from ambient IR noise, and is, thus, more robust to variable lighting conditions. However, FTIR does not provide (or is weak on) some desirable features, such as finger proximity and tracking quick gestures.

This section presents an improvement for FTIR based multi-touch sensing that partly addresses the above issues exploiting the shadows projected on the surface by the hands to improve the quality of the tracking system. The proposed solution exploits natural uncontrolled light to improve the tracking algorithm: it takes advantage of the natural IR noise to aid tracking, thus turning one of the main issues of MT sensors into a useful quality, making it possible to implement pre-contact feedback and enhance tracking precision.

9.3.1 Building a Reliable Sensor

Multi-touch displays offer a suitable working environment for computer supported cooperative work and fosters the exploration of new forms of social computing. A key technology for the design of multi-touch systems is Frustrated Total Internal Reflection (FTIR). Common FTIR setups [71, 72] have a transparent acrylic pane with a frame of LEDs around the side injecting infrared light. When the user touches the acrylic, the light escapes and is reflected at the finger's point of contact. The infrared sensitive camera at the back of the pane can clearly see these reflections. As the acrylic is transparent a projector can

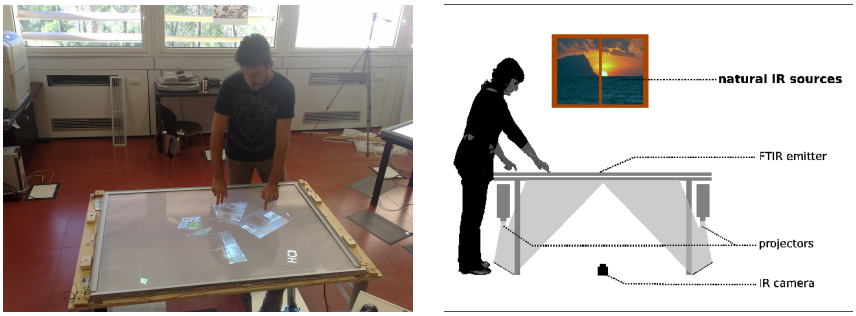


Figure 9.8: The multi-touch interactive tabletop: the picture was captured while using the table (left) and schema representing the overall setup of the table (right). Notice the operational conditions: strong direct lights and sharp variations of the luminosity.

be located behind the surface (near to the camera) yielding a back-projected touch sensitive display. The software framework relies on a set of computer vision algorithms applied to the camera image to determine the location of the contact point. An advantage of FTIR based sensors over competing solutions (such as DI, DSI [169]) is that this technology suffers less from ambient IR noise, and is thus more robust to changing lighting conditions. On the other hand, it is well known that FTIR has some disadvantages:

- it does not sense finger proximity, the user must touch the surface;
- it is difficult to track the fingers during movements;
- though more robust to changes in ambient light, it still relies on a control over lighting conditions.

To partly address such issues we propose to take advantage of the shadows that the hands of the user project on the interaction surface. Our experiments show that such solution allows to effectively sense user interaction in an uncontrolled environment, and without the need of screening the sides of the multi-touch table (see Figure 9.8).

9.3.2 Tracking IR Shadows

Tracking infrared shadows to improve the quality of multi-touch interaction has been studied before. Echtler and co-workers [47] describe a system to sense hovering on the surface, and thus provide pre-contact feedback in order to improve the precision of touch on the user's part. However the system they describe is based on a controlled IR lighting source above the table. In this

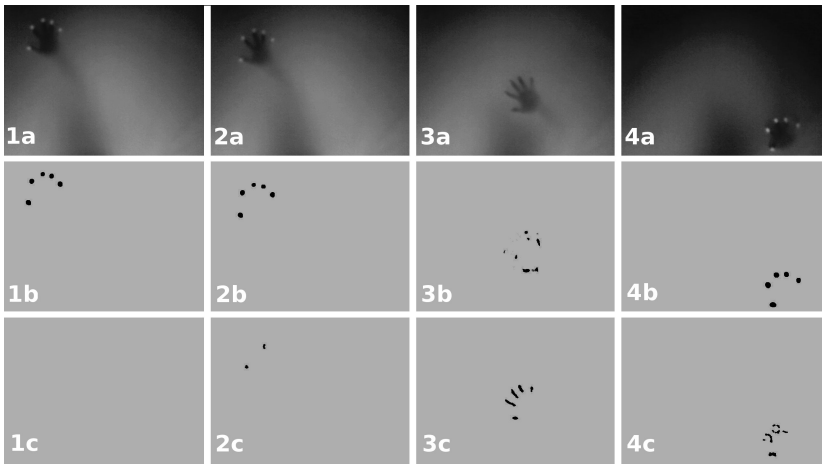
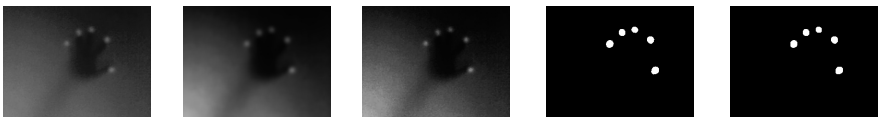


Figure 9.9: The result of tracking on IR light and IR shadow.

sense their system exploits an additional artificial lighting source, increasing the dependence on the lighting conditions.

Our solution, as further described below, exploits natural uncontrolled light to improve the tracking algorithm. We take advantage of the natural IR noise to aid tracking, thus turning one of the main issues of MT sensors into a useful quality, making it possible to enhance tracking precision and implement pre-contact feedback. The proposed technology exploits the shadows projected on the surface by the hands of the users to improve the quality of the tracking system.

As said above, ambient light has a negative impact on the IR based sensors when the light coming from the IR LEDs is not bright enough to prevail on the background noise. However, the hands of the user project a shadow on the surface (that will appear as a dark area in the noisy background). Such dark area is easily tracked because it is almost completely free of noise.



(a) raw image (b) blur (c) multiply (d) foreground (e) opening

Figure 9.10: Smoothing, enhancement and foreground segmentation on IR light blobs.

Furthermore, fingertips correspond to the darker parts of the shadow, and can be recognized with good accuracy. Note that tracking the shadow is more and more effective as the ambient light increases (as opposite from IR blobs

(a) raw image (b) blur (c) multiply (d) foreground (e) opening

Figure 9.11: Smoothing, enhancement and foreground segmentation on IR shadows.

tracking), thus IR tracking and shadow tracking tend to complement each other, the former working better in full darkness, the latter in full daylight. A second useful feature, consists in the ability of the shadow tracking system to sense objects that are only close (i.e., don't actually touch) the surface, thus allowing the sensor to recognize a richer collection of gestures.

Finally, a well known problem of FTIR based systems is that blob brightness decreases as the user moves her hands fast. This problem is typically addressed covering the screen with compliant surface and silicon rubber. Shadow tracking does not suffer from this issue, and can thus be exploited to improve finger tracking during sharp movements. Such complementarity is a key aspects of our work: it allows the system to work in less controlled environments, and to be more robust to changing lighting condition, as may easily happen in real world, off-lab installations. This latter is, as known, one of the major issues for computer vision based interactive systems.

Our implementation, based on OpenCV [175] for computer vision algorithms, shows significant improvements in the effectiveness of the sensor and, as a consequence, on the quality of interaction.

Figure 9.9 shows some frames from the image processing pipeline. Frames (1a-4a) are raw images as captured from the IR camera. The hand of the user is moving from top left to bottom right. Frames (1b-4b) are the output of the IR light tracking. Frames (1c-4c) are the output of IR shadow tracking. At (1a) the user has just touched the surface in an area relatively free of noise. The fingertips adhere well to the surface and the FTIR effect works perfectly as the result of IR tracking displayed in (1b) shows.

At (2a) the user is beginning to move her hand. As known, the IR light blobs tend to dim, but are still clear and trackable (2b). This is due to the fact that (i) the finger adhere less effectively to the surface while moving, and (ii) the hand is entering a noisy area. However the latter is partially counterbalanced by the IR shadow tracking (2c).

At (3a) the hand of the user is moving very fast and is within an area of high IR noise. The IR light blobs are invisible (3b), but the IR shadow appears clear and is easily tracked (3c).

Finally, at (4a) the user has completed the interaction phase and holds her hand still. Again the IR light blobs prevail on the noisy background and can be

tracked with great precision (4b).

At this point, combining the two input sources (light blobs and infrared shadows) is a straightforward task; details are given in the next section (Tracking).

9.3.3 Image Processing Pipeline

As known, the process of finger tracking for CV based multi-touch sensors is typically modeled as a pipeline consisting of several stages: from image acquisition to preprocessing, finger detection and tracking. All transformations are implemented by means of convolution matrices. The steps through which our implementation passes are as following.

Smoothing A blur filter is applied to smooth the image removing the Gaussian noise, thus getting rid of pixel size spots (see Equation 9.1 and Figures 9.10b and 9.11b).

$$G(x, y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{9.1}$$

Enhancement A rectification filter enhances the luminosity of each pixel (see Equation 9.2 and Figures 9.10c and 9.11c).

$$img(x, y) = \frac{(img(x, y))^2}{(max(img(x, y)))^2} \tag{9.2}$$

Background Removal Filter The picture is filtered in order to find the areas of the screen on which an interaction is happening. To this purpose a 7×7 matrix with Gaussian distribution was empirically determined. The result is matched against a threshold in order to select relevant areas. This operation in practice finds local maxima in the captured image. However the resulting image still presents some noise and must be further processed. Note that this same filter, applied to the negative image, is used in shadow tracking (see Figures 9.10d and 9.11d).

Opening An opening filter erodes spots whose size is smaller than a given value, often referred to as *salt and pepper* noise (see Equation 9.3 and Figures 9.10e and 9.11e).

$$img \circ m = (img \ominus m) \oplus m \tag{9.3}$$

Lens Distortion Removal The image is processed in order to compensate radial and tangential distortion due to the lens of the camera. Radial (Equation 9.4) and tangential (Equation 9.5) distortion correction require parameters p and k that can be computed by identifying distortions of images containing

known regular patterns [22] (see Figure 9.12). Note that OpenCV provides black-box functions to this purpose.

$$\begin{aligned} x_{\text{corrected}} &= x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ y_{\text{corrected}} &= y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \end{aligned} \tag{9.4}$$

$$\begin{aligned} x_{\text{corrected}} &= x + [2p_1 y + p_2(r^2 + 2x^2)] \\ y_{\text{corrected}} &= y + [p_1(r^2 + 2y^2) + 2p_2 x] \end{aligned} \tag{9.5}$$

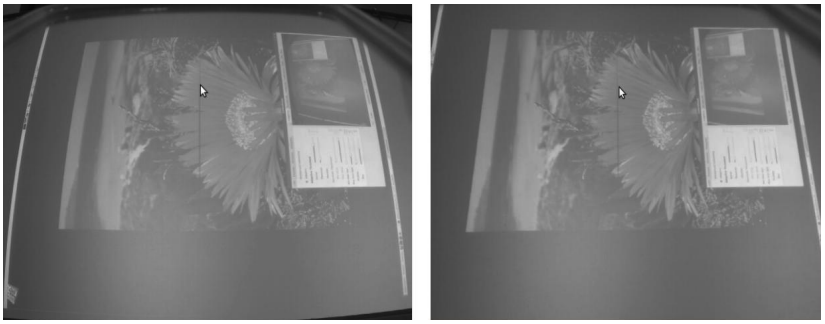


Figure 9.12: Correction of lens distortion (pincushion and barrel).

Perspective Distortion Correction This last stage aims at transforming between capture coordinates and display coordinates and getting rid of perspective when (as often happens) the camera is not placed perfectly perpendicular against the plane of interaction. This operation requires four points on the screen to be matched against 4 points in the capture. Usually this is performed manually (during an initial *calibration* phase). Such transformation is efficiently computed as an inverse mapping between triangular meshes [20].

To do so, the position of a point to be mapped from camera space to display space can be expressed in baricentric coordinates: if A, B and C are the vertices of a triangle, a point P inside the triangle is uniquely identified by $P = \lambda_1 A + \lambda_2 B + \lambda_3 C$, where $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Any deformation applied to the triangle does not change the baricentric coordinates of the point P, then since the coordinates of points A, B and C in the display are known from the calibration phase it's easy to compute the coordinates of point P on the display.

The complete pipeline, both for IR blob light tracking and IR shadow tracking is depicted in Figures 9.10 and 9.11. See from left to right how the image is filtered to enhance meaningful features.

Tracking Finally, tracking fingers that touch the screen is done as follows:

1. an improved Continuosly Adaptive Mean-shift algorithm (camshift) [175] is applied to determine a region of interest (ROI) surrounding the finger

in each successive frame, in order to track the finger and reduce the region of calculation, the camshift algorithm constantly adjusts the size of the search window;

2. for each video frame, a matrix that represents the probability distribuition of the foreground image is analyzed to determine the centre of the ROI;

3. the current size and location of the tracked object are reported and used to set the size location of the search window in the next video image.

4. based on the previous items, the system searches for fingers both in the shadow and light foreground images so that the tracking will continue even in variable lighting conditions.

All these allowed us to deploy a robust and reliable multi-touch table, easy to move and to calibrate

9.4 Lessons Learnt

From the very first proof of concept to the latest development of the interactive wall and table described above, there have been countless phases or redesign, during which the software implementation has been rewritten from scratch several times, and the choice of hardware has evolved, also according to market availability.

Both the interactive wall and the table have been shown to dozens of colleagues and visitors, that have tried the system, given feedback, sometimes criticized, even bitterly in some cases. The lack of a *killer application* was sometimes felt as a limitation for this type of technology, despite the obvious analogy with blackboards, bulletin boards, shop windows, etc.

This often drew criticism regarding the possibility to implement traditional applications for this new device. This became one central topic of our research, and is one key research question of this work, addressed later. Over time we were able to identify some key aspects that affected the way people experience a technology such as the one presented in this chapter:

Magic: in the very first demonstrations, when multi-touch was still a unknown to most non-specialists, the word *magic* was heard quite often. It was meant to summarizes the effect of surprise, not only for the novelty of the multi-touch interaction paradigm, but also for its apparent simplicity.

Participation: not only visitors asked to *try* the new technology, but were often proactive in suggesting new applications related to their own domain of interest, thus (in a sense) *volunteering* a creative effort, that turned out to be invaluable for our work.

Collaboration and competition: whenever possible users shared the workspace and managed to use the system together. Again the possibility to engage in playful conflicts, such as interfering with what others were doing, was readily exploited as a mean to negotiate the space or gain attention. The *demos* that involved multisensory stimuli, such as those ones that coupled animation and sounds, appeared to be the more engaging, and fostered the more cooperation;

Robustness: manipulative interaction appears to be demanding in terms of expectations just as much as engaging. People expect the system to give prompt reactions, just as the physical world does. This is not surprising given the above discussion on embodiment, but at the same time it is technically very challenging. Any malfunctioning is exceedingly frustrating, and spoils completely the sense of *magic*;

We managed to address the issues that emerged and take advantage of our own errors in the redesign of the interactive space. Ultimately the system was used to conduct a formal experiment aimed at evaluating the impact of tabletop systems on traditional, office related activities. The results are the subject of the next chapter.

Acknowledgement: This chapter is based on revised contents from the papers:

Alessandro Soro, Gavino Paddeu, and Mirko Lobina. *Multitouch sensing for collaborative interactive walls.* In Peter Forbrig, Fabio Paternò, and Annelise Pejtersen, editors, Human-Computer Interaction Symposium, volume 272 of IFIP Intl. Federation for Information Processing, pages 207-212. Springer Boston, 2008.

Alessandro Lai, Alessandro Soro, and Riccardo Scateni. *Interactive calibration of a multi-projector system in a video-wall multi-touch environment.* In Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST '10). ACM, New York, NY, USA, 437-438.

Samuel A. Iacolina, Alessandro Soro, and Riccardo Scateni. *Improving FTIR based multi-touch sensors with IR shadow tracking.* In Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems (EICS '11). ACM, New York, NY, USA, 241-246.

Chapter 10

Gestures in Multi-touch Interaction

As already thoroughly discussed, natural user interfaces are often described as familiar, evocative and intuitive, predictable, based on common skills. Though unquestionable in principle, such definitions don't provide the designer with effective means for creating a natural interface or evaluate a design choice against another. Two main issues in particular are open:

(a) how do we evaluate a natural interface, is there a way to measure 'naturalness';

(b) do natural user interfaces provide a concrete advantage in terms of efficiency, with respect to more traditional interface paradigms?

In this chapter we discuss and compare observations of user behaviour in the task of pair programming, performed at a traditional desktop versus a multi-touch table.

A generally accepted, while hard to quantify, advantage of multi-touch tables and walls over desktops is their being inherently multi-user: people cooperate to the task at hand, sharing or negotiating the use of the device in a natural manner. Depending on the specific task it is easy to observe an increase of non-verbal communication (gestures, body postures, facial expressions, etc.) that are proven to have a positive impact on many cognitive processes.

Gestures in particular represent an easy to measure virtuous practice that in desktop computing appear limited almost exclusively to pointing with hand or finger, while observing users of multi-touch tables it often happens to see fluent, dual-handed metaphorical gestures. This raises the questions we try to answer. Is there any practical advantage (e.g., in terms of efficient problem solving) when using a natural interface? More precisely: is multi-touch better

than the desktop for some traditional application? Moreover: can gesticulation be used as a suitable signal of natural interaction justifying the chain that more gestures provoke a more natural, and, thus, better interaction?

We opted to experiment with *pair-programming* [189]. It is a practice of software engineering strongly recommended by *agile methodologies* and, thus, represents a realistic and not artificial test-bed both for desktop and multi-touch setting. Additionally, gesticulation, which we aim to observe, is more easily, though not exclusively, triggered during group-work.

We will be able to show how interacting in a multi-touch environment determines a significative increase of non-verbal communication in general and especially, of gestures, that in turn appears related to the overall performance of the users in the task of algorithm understanding and debugging.

10.1 Related Research

As mentioned before, multi-touch interaction has been a topic of research since the mid-eighties (e.g. [26, 101, 119, 124]), but it's with the recent work of Han [71, 72] than this interaction paradigm has become popular and multi-touch interaction is now so often taken as an example of natural interface.

However, applications based on this interaction paradigm are still in a phase of creative envisioning (e.g. [76, 192, 197]) and little, if any, study exists on the real advantages of direct manipulation in traditional application fields.

Owen and colleagues [149] explore the advantage of bi-manual input on a curve matching task; Patten and Ishii [150] present a study that compares the strategies (and effectiveness) of spatial organization with tangible and traditional user interfaces.

These studies let foresee an advantage of direct manipulation, and by extension of multi-touch tables, over traditional desktop for very specific tasks that have in common a certain physicality, but don't settle the point on whether or not surface computing can replace the desktop in traditional work or learning scenarios.

Rogers and Lindley have compared how people interact and share the space when cooperating at vertical (such as interactive walls) versus horizontal (such as tabletops) displays. They found that when working at the multi-touch table users tend to switch more between roles, share more ideas, and are generally more aware of the contribution of each member of the group [165].

Buisine and colleagues have explored how people perform in a task of creative problem solving, supported by an interactive tabletop with respect to pen and paper. According to their results the form factor (the around-the-table setting) and the attractiveness of the multi-touch device improved collaboration by reducing the *inequity index*, that accounts for the balance of individual participation [23].

However, the effectiveness of a multi-touch tabletop setting in office related activities is still an open issue.

10.2 Experimental Setting

Figure 10.1: People participating in the experiment: at the multi-touch (left) and at a traditional desktop (right)

A convenience sample of 44 people participated to this study, age 20-35, all students of computer science or ICT professionals, thus quite literate in computer programming. Working in pairs (see Figure 10.1), the testers were asked to review 7 snippets of C code (1 demo, and 6 exercises), each one containing a bug, and to point out the bug to an assistant. The review of the code snippets was performed through a very simple interface implemented with the identical look and feel both for the desktop and for the multi-touch environment.

The appearance of the graphical interface is shown in Figure 10.2; it consists of

- a square text-area that shows one snippet of code at a time. The snippets of code are short enough to fit the visible area, so no scrolling is ever needed and no scrollbars are thus provided;
- a small control panel with a timer, and buttons to jump to forward and backwards between the exercises; multi-touch functionalities where enabled on the MT table and simulated with keyboard/mouse combinations on the desktop.

However, in practice, testers seldom manipulated the interface, except for hitting the *Next* button.

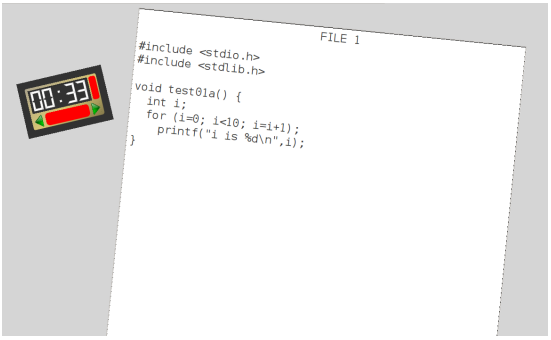


Figure 10.2: The appearance of the user interface

10.2.1 Pre-test briefing

Before the beginning of a test session the testers where briefed on the purpose and method of the research. We had great care to specify that the goal of the work was to evaluate the quality of the tool (Desktop vs Multi-touch) and not the ability of the users. The need of a video recording was justified by explaining our need to monitor "collaboration and non-verbal communication" but without explicitly mentioning gestures, or their supposed connection with efficient problem solving. The testers where then encouraged to cooperate to the solution of the problems. The testers were also informed that:

1. every snippet contains one (and only one) bug;
2. the bug is not in the syntax, but in the logic of the code;
3. comments (where provided) are not misleading;
4. bugs, although trivial to explain, where sometimes well concealed, and intended to be difficult to spot;
5. finally, although no time constraints were given, the testers were informed that the whole test required between 15 and 25 minutes on average. This was not intended to fix a goal for the performance, but to prepare the testers to the effort needed to complete the test.

Of course all participants were given written warranty of privacy and non-disclosure of videos and disaggregated data.

10.2.2 Test Session

The 44 testers (spontaneously organized in 22 couples) were then asked to complete the experiment. 11 tests were run at the Desktop and 11 were run at

the Multi-touch table, the assignment to one or the other setting was performed randomly. The F-test was used to verify if a significant difference exists between the two methods, multi-touch and desktop. Note that the same 7 exercises were administered at the 2 settings.

Of the 7 snippets of C code, the first one was intended as a demonstration to get into confidence with the interface and clarify latest doubts; results are not taken into account in the following discussion. For each one of the remaining 6 snippets, the testers had to perform the following:

1. examine the snippet for as much time as needed, discussing, if necessary, to decide what the bug was;
2. as soon as an agreement was reached on the exercise, press a pushbutton (that turns green) on the control panel;
3. testers could then point out the bug to an assistant, who annotated it in a block notes, without either confirming or refusing the answer;
4. by pressing a pushbutton on the control panel the testers could then proceed to the following exercise.

Note that in both settings:

- the interface didn't allow any editing of the C code; so the users were not able to correct the error;
- since the assistant did not comment on the proposed solution, the test actually measures the time spent before reaching an agreement, we did not measure the accuracy (i.e. if the testers positively solve the exercise or not) of the exercise; thus, wherever in the rest of the paper we talk of solving an exercise it should be clear that we mean reaching an agreement on the solution;
- the cases in which the testers were not able to reach an agreement (either on the correct or on a wrong answer), were also included; in a sense this results indicate the time spent before deciding that additional tools/information was needed to positively solve the exercise; of course such cases should better be taken into account in a deeper investigation;
- the testers hit a button after reaching an agreement and another one to switch to next exercise, thus the time spent in reporting the bug to the assistant is known and has been expunged in the following discussion.

10.2.3 The 6 Code Snippets

The various exercises have been designed to be of increasing complexity and length (and in general took increasing time to solve). The exercise can be

divided in 4 categories, and were administered in the same order in which they are described below:

Type 1: controversial exercises such as the one below are likely to cause debate between the testers.

```
1 void test2() {
2   int i;
3   for (i=0; i<10; i=i+1)
4     if (i=2)
5       printf("i is 2\n");
6     else
7       printf("i is not 2\n");
8 }
```

In the specific case the use of an assignment as argument of a truth evaluation, though not syntactically wrong, is typically deprecated. There are exceptions however, and the testers spent time discussing whether or not the use of such construct was acceptable in the context of the exercise. **Type 2**: slips or careless errors are very common in everyday programming and are easily spotted since often result in meaningless or inconsistent code.

```
1   void test3() {
2   int i;
3   i = 0;
4   while (i < 10);
5     i = i + 1;
6   printf("Finished. i = %d\\n",i);
7  }
```

In this case the body of the while construct is actually an empty statement (because of the semicolon), resulting in an infinite loop. **Type 3**: pattern matching error are those ones that require visual memory or recognition, and represent a class of errors almost unknown to programmers today, thanks to the use of visual editors that provide syntax highlighting. Examples include misplaced parentheses due to a wrong indentation, and comments opened and not closed, such as in the example below.

```
1 void test4() {
2   int i;
3   for (i=0; i<10; i=i+1)
4     /* check the value of i */
5     switch(i){
6       /* is i 0? */
7       case 0: printf("i is 0\n");
8               break;
9       /* is i 1?
```

```
10       case 1: printf("i is 1\n");
11               break;
12       /* now the default case */
13       default: printf("i is more than 1\n");
14    }
15 }
```

Most modern editors would help the programmer to find the error here: the comment at line 9 is not closed at the end of the line, and runs through to line 12, voiding in practice the body of the function. Without the help of syntax highlighting, the testers were forced to check the syntax of comments, which is trivial in practice, but not intuitive.

Type 4: algorithm understanding exercises are those ones for which the most effort was required. The bugs consisted in the overrun of array indexes, such as in the example below.

```
 1 void bubble_sort(int array[], int n) {
 2   int i, j;
 3   // sort array of length n
 4   for (i = (n - 1); i > 1; i++) {
 5     for (j = 0; j < i; j++) {
 6       if (array[j] > array[j + 1]) {
 7         // swap values
 8         int tmp = array[j];
 9         array[j] = array[j + 1];
10         array[j + 1] = tmp;
11       }
12     }
13   }
14 }
```

Here the outer for cycle will never end and causes an array overrun on the subsequent instructions. Testers were able to solve the exercise only after understanding (or recollecting from previous study) the basic logic of the algorithm.

10.2.4 Data Collection

Data collected during or following the test are:

1. the time spent on each exercise;
2. the proposed solution, that may or may not be correct;
3. the video footage of the whole session.

These were used in the analysis described in the next section. Other data gathered, but not discussed in detail here are:

1. whether or not the testers were able to reach an agreement on the solution of the exercise;
2. subjective scores of the difficulty of each exercise.

This information, as we already noticed, will be subject to further investigation on the accuracy of the performance and on the subjective perceived difficulty of the exercises in the two settings.

Analysis of the Video Log To better understand the role of gestures in collaborative work we have analysed the video logs of the test sessions in order to count the gestural events. As pointed out in Part 2 there is strong evidence that a fluent gesticulation has a positive influence, among others, on short term memory [63] and learning [36].

We claim that a similar behaviour exists in solving complex tasks such as the one considered here, and that a system that allows (or encourages) a fluent gesticulation allows better performances. The video collected were annotated using Anvil [112], a platform for multi-layered annotation of video with gesture, posture, and discourse information.

10.2.5 Experimental Hypotheses

As mentioned earlier, among the scopes of this work, a main goal is to answer the following research questions:

1. Is there any practical advantage (e.g., in terms of efficient problem solving) when using a natural interface? More precisely: is multi-touch better than the desktop for some traditional application?
2. Can gesticulation be used as a suitable signal of natural interaction (i.e., the more gestures, the more natural, and the better interaction)?

Hence the null hypothesis related to question 1.

H1. Participants will be no faster in solving an exercise containing a controversial bug, when using the Multi-touch table or the Desktop.

H2. Participants will be no faster in solving an exercise containing a careless error, when using the Multi-touch table or the Desktop.

H3. Participants will be no faster in solving an exercise requiring a pattern matching, when using the Multi-touch table or the Desktop

H4. Participants will be no faster in solving an exercise that require algorithm understanding, when using the Multi-touch table or the Desktop.

In order to positively answer question 2 we should first prove that the observed difference in fluency of gestures couldn't be otherwise explained:

H5. Participants will gesture with no more of less fluency (measured as gestural units/time) at the Desktop or at the Multi-touch table.

Further hypotheses, showing if fluency of gestures has any direct impact on efficient problem solving (i.e., couples with more fluent gesture actually perform better), or a deeper exploration in the nature of gestures involved in this specific task (e.g., what pantomimes, icons, metaphors were used in addition to deictics that helped the participants who scored the better results) are outside the scope of this experiment.

10.3 Results and Discussion

The experiments show that people perform significantly better at the multi-touch table (for the task examined) than at the desktop for some of the exercise, namely those ones involving cooperation, discussion, and, more generally, exchange of communicational information. Do participants perform better when solving an

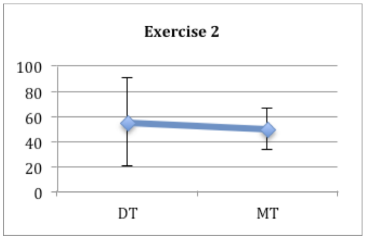


Figure 10.3: Results of exercise 2 (controversial bugs)

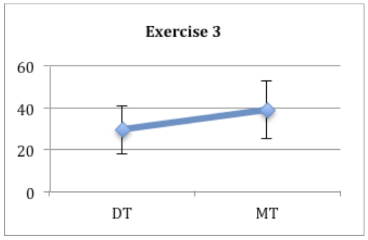


Figure 10.4: Results of exercise 3 (careless errors)

exercise containing a controversial bug, when using the Multi-touch table than the Desktop? As shown in figure 10.3 tester scored slightly better performances at the Multi-touch; the difference is significant, $F(10, 10) = 4.72$. Hypothesis H1 should then be rejected. The analysis of results of exercise 3 does not show any significant difference between the Desktop and the Multi-touch, $F(10, 10) = 1.46$, n.s. figure 10.4 shows means and standard errors for the results of the experiments. Similarly, no significant difference was observed in the execution of exercises 4 and 5, both containing errors requiring a pattern matching: precisely: $F(10, 10) = 1.29$, n.s. for exercise 4 and $F(10, 10) = 1.08$ for exercise 5. Figure 10.5 shows the results. Finally, exercise 6 and 7 required the most effort from the testers (as shown by the longer time to solve on average, figure 10.6), and the Multi-touch setting allowed a tighter cooperation resulting in a significant better performance: $F(10, 10) = 5.56$ for exercise 6, $F(10, 10) = 13.50$ for exercise 7. The timing are summarized in table 10.1.

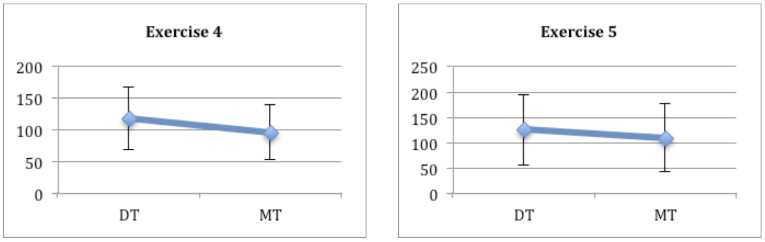


Figure 10.5: Results of exercise 4 and 5 (pattern matching)

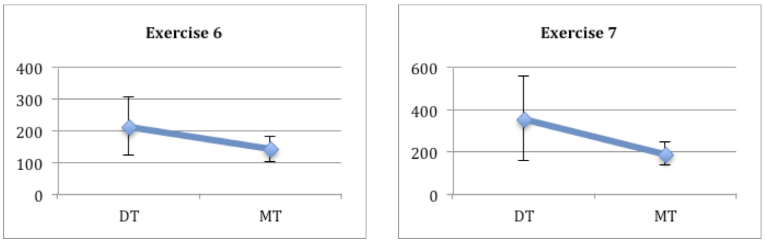


Figure 10.6: Results of exercise 6 and 7 (algorithm understanding)

Exercise #	Avg. Time (Desktop)	Avg. Time (Multi-touch)
2	55.18	49.91
3	29.64	39.18
4	119.27	95.82
5	125.64	109.36
6	213.55	143.09
7	356.55	190.80

Table 10.1: Time spent on average on each exercise. Desktop (left column) and Multi-touch (right column) performances are compared.

10.4 Gesture Fluency

Our last test was aimed at showing if users manifested a difference behaviour with respect to gesture fluency in the MT and DT settings. We observed proper gestures according to the related literature given in Part 2. In particular:

- Only movements of the hands were counted as gestures, thus excluding nodding and changes in body postures; specifically, pointing with the mouse was not counted as gesturing; in fact, mouse pointing is not a proper gesture and comparison to previous work is problematic. Additionally, we can't assume the visibility of the mouse gesture to the other user, i.e. there is no clear communicative intent (see later).

- Movements of the hands were counted as gestures when they had a clear communicative intent: folding the hands together is not a gesture; pointing, mimicking an action, and counting with fingers are all considered gestures;

- Gesture phrases were counted as their atomic components where possible; for instance, when a tester points a section of code, then another to show correlation, and finally makes sharp movements to show progress, even if these three movements are executed without any visible pause, were counted as 3 separate gestures.

We, thus, introduced for simplicity a measure of gesture fluency, as the number of gestural events per second of both testers, and, for each one of the 22 couples, we counted the gesture events of both testers. The gesture fluency of a couple is the total number of gestures performed divided by the total time spent solving the 7 exercises. Results are shown in figure 10.7. The experiment shows that

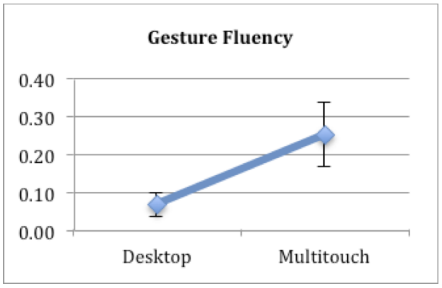


Figure 10.7: Gesture fluency compared for the two settings: desktop (left) and multi-touch (right)

participants use significantly more gestures when using the multi-touch than when interacting at the Desktop, $F(10, 10) = 7.70$; thus we reject hypothesis H5.

10.5 Discussion

The results shown above indicate that while working at the multi-touch table people perform better than at a traditional desktop, and such improvement is associated to an increased gesticulation. Some remarks are due here. Not all types of exercise seem to benefit from the adoption of a multi-touch system, in particular snippets containing careless errors (exercise 3) and pattern matching (exercises 4 and 5) were not significantly affected by the different setting. Controversial exercises (such as exercise 2) are better addressed at the multi-touch, where a tighter cooperation is possible. This is hardly a surprise, since this sort of problems requires discussion and sometimes negotiation between the users.

The results obtained for exercises 6 and 7 (algorithm understanding) are perhaps less intuitive and their implications in the design of interactive applications deserve some attention. On the one hand this work gives a further confirmation of the already observed connection between gesturing and problem solving. In this case an improvement in the graphical interactive systems did not involve improvements in the interface (as remarked above, people didn't do extensive use of the multi-touch features of the tabletop setting), but rather the design of a work-setting more suitable for cooperation, and fluent gesturing was taken as a metric for the cooperation itself.

On the other hand, one can notice that the exercises taking the most benefit from the multi-touch setting were the more difficult among the 7 administered, and still were trivial with respect to the typical problems that programmers face in daily work. Our experiments suggest that multi-touch tables, encouraging cooperation, help people express their potential, thus resulting in a better performance. The registered difference in performances for code understanding and debugging time could make a significant difference in many practical cases. These results may help reconsidering the design of our offices and programming labs towards a more widespread adoption of tabletops, that today are mostly regarded as research prototypes and curiosities.

Some further empirical observations are worth mentioning here. Our metric of gesture fluency was suitable for the work at hand, but hides the real complexity of gesture phrases. If the gesture largely more exploited by all participants was pointing with one finger, others where frequently observed:

- Gestures indicating progress or continuity, both single and dual-handed, are executed moving the hand(s) on a circle or sharply from left to right; such gestures are not easily performed when sitting, and not surprisingly they are less frequently seen at the Desktop;

- Some gestures are performed primarily for communicating, they are a sort of visible words; as such they have to be performed in a well defined and visible space; again, such space (close to the screen) is easier to reach at the multi-touch than it is at the Desktop;

- At the multi-touch pointing with the finger was sometimes used to negotiate the attention of the mate; testers often pointed at the same point on screen as to reinforce and confirm a gesture; this behaviour was not observed at the Desktop;
- In one case a tester asked if she could use paper and a pencil, which was not possible, actually; several participants at the multi-touch setting were observed while mimicking the use of paper and pencil on the palm of the open hand.

10.6 Lessons Learnt

In our experiment the adoption of a multi-touch user interface leads to a significant, observable and measurable, increase of non-verbal communication in general and of gestures in particular, that in turn appears related to the overall performance of the users in the task of code understanding and debugging.

Our results indicate that working at the multi-touch table people perform better than at a traditional desktop, and such improvement is associated to an increased gesticulation. Users at the multi-touch outperformed desktop user for specific classes of problem, and such gap corresponded to 4 as much gestural events, indicating an improved cooperation. As noted throughout the work however, several questions remain open.

- We didn't observe the accuracy of the solutions proposed. For the scope of this research a problem was considered solved when an agreement on the proposed solution was reached; though in principle cooperation and discussion lead to accurate results, a precise measure of such accuracy is likely to expose new insights;
- The choice of gestures to observe was arbitrary, though shared in literature; for example, pointing with the mouse is a common behaviour at the Desktop, whose impact should be evaluated;
- How strong an interrelation exists between gestures efficiency and accuracy? Do couples that show more gesture fluency perform better?
- What new insights may come from a more detailed analysis of gestures; gesture fluency doesn't capture the richness of expression that emerges at the multi-touch table, where dual-handed symbolic gestures are often used compared to bare single-handed deictic that form the majority of gestures at the desktop.

The implications of these results will be discussed in greater depth in the concluding remarks, in Part 4.

Acknowledgement: This chapter is a revised version of the paper:

Alessandro Soro, Samuel Aldo Iacolina, Riccardo Scateni, and Selene Uras. 2011. *Evaluation of user gestures in multi-touch interaction: a case study in pair-programming.* In Proceedings of the 13th international conference on multimodal interfaces (ICMI '11). ACM, New York, NY, USA, 161-168.

Some of the C code snippets were adapted from fragments available online under GNU GPL or analogous licenses. Snippets 1-4 were adapted from [127].

"Which road leads to the Wicked Witch of the West?" asked Dorothy.
"There is no road," answered the Guardian of the Gates. "No one ever wishes to go that way."
"How, then, are we to find her?" inquired the girl.
"That will be easy," replied the man, "for when she knows you are in the country of the Winkies she will find you, and make you all her slaves."

L. Frank Baum

Part IV

Final Remarks

Chapter 11

Discussion

While introducing this work, I suggested that the possibility to choose the appropriate communication channel and the ability to combine several channels, moving with continuity between different communication strategies, was one core feature of natural interaction. In human computer interaction, multi-modality is the exception to the rule, though.

Additionally, even multi-modal systems typically manage to exploit a narrow set of human abilities, say manipulation + gesture or gesture + speech, and often assume a simplification of the overall collaborative scenario in terms of individual interaction bursts. Let's recall, by contrast, the example of children playing hopscotch (from [65]): in a fragment of just a few seconds the two girls deployed messages through speech + prosody + gesture + facial expression + body movement + body posture.

It is clear that the implementation of such a rich communication is probably out of reach for state of the a art technologies (now and within a foreseeable future). At the same time, it is not at all clear wether people would be *willing* to use such communication strategies in interacting with computers. Quoting Winograd *People are primarily interested in other people, and are highly motivated to interact with them in whatever media are available* [194]. We can't simply assume a correspondent motivation when people interact with computers: not surprisingly, the *accuracy rate* that people require from speech or gesture recognition systems is far beyond the accuracy that humans expect from other humans.

A consequence of the thesis exposed in this work is that natural interfaces deployed in interactive spaces cannot be designed and evaluated *individually*, but should rather be framed and observed in the broader context of cooperation (or social interaction, or competition, depending on the purpose) in that space.

Manipulative skills, such as those ones that people exploit in multi-touch or tangible interaction, are a suitable paradigm for building such natural interfaces (although not the only one). *Gestures*, and specially gesticulation, that is the

unstructured and almost spontaneous movements of the hands that invariably accompany speech, are a suitable (although not the only) evaluation criteria.

The rationale behind such claim has been thoroughly exposed in Part 2: manipulations are not just our *actuators*, actions by which we change the state of the world. They are an integral part of the way people make sense of their environment, learn how to cope with its issues and exploit its opportunities.

Gestures on the other hand cannot be simply regarded as visible words (or visible computer commands): that's only one end of the *continuum*. Gestures serve a cognitive role in helping people think, reason and talk, improve our memory and serve as signals in our continuous negotiation of the interaction *context*.

In many related works gestures and manipulations that people perform for themselves or to the benefit of other people (i.e. not explicitly meant to operate the system) are regarded as somehow *extraneous* to the human-computer interaction problem, and thus cut out of the picture. As a consequence such behaviours are sometimes obstructed, instead of being fostered and encouraged, by the design of the interactive system.

Additionally, the subtle nature of the mental processes involved make it even more difficult to cope with such design issues. People will rarely list as a requirement: *to be left free to gesticulate*. We are barely conscious of possessing such skills, although at a certain extent we can learn to make better use of them.

In this work we have discussed a number of experiences of design of tangible and manipulative interfaces; we have shown how such interfaces help fostering a *playful interaction* that is an invaluable aid for encouraging interest and participation, both in the tools themselves and in the informational content that they carry.

In the past three years we have managed to deploy our prototypes, both in public spaces, during several exhibitions, and within our lab, as a complete interactive space in which to perform experiments and evaluate natural interaction.

In the final phase of the project described here we managed to asses the effectiveness of natural interaction in a scenario of office automation. We challenged the generally accepted claim that natural interfaces are not particularly well suited for office activities designing an experiment in the context of software engineering.

Our results indicate that the effectiveness of group-work is fostered by the adoption of a tabletop multi-touch setting.

This result may raise some perplexity: is the improvement in performance due to the multi-touch feature or rather to the tabletop setting? In other words, is it a matter of interface design or of better arranging the furniture in the office? In our view it is a matter of *embodied facilitation* (see [80]). The organization

of space and artefacts constraints the possibility of people to exploit their skills, and affects collaboration.

As an analogy we may consider how *eye gaze detection* has been used in the analysis of graphical user interfaces: We know that (western) people perform a visual scan of a Webpage focusing initially (and for a longer time) on the top left corner and then on the rest of the page. And we know that the fixation time decreases dramatically going towards the bottom right corner of the page. Publishers and advertisers take these aspects into account when selling and buying an advertising space.

But we also know how to interfere with such behaviour: images tend to attract eye gaze, and images in which the face of people is clearly visible attract the attention even more: a picture placed anywhere in the page will change radically the average pattern of visual scan and exploration.

We have exploited gestures in interactions design, just like eye tracking has been used in visual design: not only the single interactive artefacts (the natural interfaces), but also the way they are made available in the space, how they are interconnected, how suitable they are for concurrent cooperative use, determine the overall natural interaction. As shown above in many examples, this cannot be inferred from the use of the single device, or from an analysis of the interaction of a single person with an ecology of devices.

The fee to pay to the added complexity is that the results of any investigation become yet more difficult to generalize: any task becomes unique; any setting is potentially impossible to compare to another; any improvement obtained under one aspect of the overall collaborative activity may result in a corresponding issue arising elsewhere. This should not be considered as a limit, but rather as a sign of change.

The nature of human computer interaction studies has changed radically since when HCI affirmed itself as a discipline. Quoting Bannon: [...] *HCI has moved from evaluation of interfaces through design of systems and into general sense-making of our world* [9].

Our attempt to give a contribution to this process was then focused on the interaction *beyond the individual*, i.e. on the activities and behaviours of people, when they interact with their digital ecosystem *and with each other*.

The next step, will be to focus on people *beyond the work-group*, and is the subject of the next chapter.

Chapter 12

Conclusions

While I'm writing these final remarks the world is facing what seems to be the worst economic crisis of the last 100 years, if not ever. Europe, that in half century has built his unity on the hashes of the WWII, overcoming step by step political, cultural and economical differences is now struggling for its survival against what appears as a blind fury of the global financial market (or the blind greed of the global financial tycoons).

The crisis stroke dramatically on the African shore of the Mediterranean sea. In less than one year, a wave of social and cultural revolution has swept away long lasting tyrannies in Egypt, Libya and Tunisia, causing also hundreds of thousands of refugees to press towards the borders of southern European countries.

If the causes of the so called *Arab Spring* span from human rights violations to extreme poverty and governments corruption, the methods of the revolutions focused, at least in the beginning, on civil resistance, peaceful demonstration, and a *systematic use of social media to plan, coordinate and document the protests* in spite of any attempt of censorship. In a sense, while we were trying to answer questions such as *what can ICTs do for us as a work-group?* history was already moving two steps ahead.

What can ICTs do for us as a *people*? The change is happening already: user generated Web contents are a primary source of information for millions of people, specially among the youngest generation. Web based social media are blurring the boundaries between personal and professional life, even in fields in which privacy and confidentiality have always been considered of paramount importance (see for example [98]). Politicians have entered (willing or not) the social media arena with their Facebook pages.

In such arena, U.S.A. President Barack Obama is considered a first mover (see [57]): it is not just another stand from which to campaign, it allows

communicating with a network of followers, to engage and stimulate participation in supporters. But it is equally open to the critics of opponents: every single word gets scrutinized, registered and criticized, moderation is practically impossible, as many have learnt to their own expenses.

What happens when social media meet the mobile world of location sensitive services? A digital superstructure of social relationships that overlays the physical world becomes available. With rules and conventions analogous to the physical world, but loose spatio-temporal constraints, it enables totally new forms of interaction (see [15]). Will we be able to exploit the potentiality of social and mobile media to change the way we confront (or embrace) cultural and socio-economical diversity?

Shifting the focus from designing *for the individual* user to designing *with communities* of people, plenty of projects are currently in progress, that aim at involving, informing or rising awareness about, to name a few examples, sustainable food production [34], energy consumption [130], experience reporting and sharing [97].

From climate change to employment and education, from migration and integration to globalization, most of the challenges that we'll have to face in the 21stcentury will force us to rethink, negotiate, and actively reshape our mutual understanding and our networks of relationships, both as individuals and as communities. ICTs can provide the infrastructure to support a new culture of active participation, engagement, social awareness and reciprocal respect, and social computing will certainly have a leading role in this change.

Designing, implementing and exploiting such infrastructure, however, is ultimately entirely up to us.

Appendices

Appendix A
Web based Video Annotation

Given the importance of gestures in expositions and specially their utility in teaching, we could think to evaluate the quality of an exposition under the aspect of the gestural communication. Such evaluation will help to analyse speakers' teaching and communication skills, and help them to improve the overall quality, focusing strengths and flows of a performance.

In this chapter we present the design, development and initial evaluation of MORAVIA (*MOtion Recognition And VIdeo Annotation*): a collaborative web application for (semi)automatic gesture annotation. MORAVIA is able to detect the gestures using a particular video camera, a depth camera like the Microsoft Kinect^{TM}and extracting the body skeleton for the recognition of every single gesture. Then, our web application for video annotation allows collaborative review and analysis of the different video sequences, both for domain experts, as a research tool, as well as end users, for self-evaluation; in the end, the overall system is meant to be capable of giving a quality score to the entire performance.

We briefly recall the most relevant proposals in the different fields touched by our work.

Gesture recognition: gestures can be represented as multi-dimensional and time-dependent data, so a classic approach for their recognition is the use of Hidden Markov Model (HMM) [156, 157]: this is a valid method to recognize time varying data and consists of a network of nodes, transitions between nodes and transitions probabilities, starting from defined input symbols. These input symbols must be discrete and because gestures are multi-dimensional data it's mandatory to perform a discretization, viable in different manners and used in some works with HMM: manually specified conditions [176]; Self Organizing Map (SOM) [117] used by Iuchi et. al [94].

Other gesture recognition methods use state machines, as proposed by Matsunaga and colleagues [132] that used Support Vector Machines (SVM) [132] for transition conditions learning, while instead Oshita [132] used manually specified fuzzy-based rules. In any case the state machine is created manually and it's not an easy process: this can be an important limitation, but the work of Oshita and Matsunaga [146] obtained important results; they used SOM to divide gestures in phases, thus to create the state machine, and finally they used SVM to determine transition conditions between nodes. Beyond theoretical aspects, a real gesture recognition implementation requires the use of ad-hoc hardware instruments like sensors or tracking devices.

As analyzed by Mitra and Acharya [138], those devices can be classified in two categories:

- Wearable tracking devices, gloves, suits and similar.
- Computer vision-based devices and techniques, video cameras paired with algorithms that find movements from the video.

Wearable tracking devices are very accurate and can reveal sudden movements, like fingers movements while moving hands; on the other hand, methods based on Computer Vision are less invasive and are able to identify also colors and textures.

Depth Cameras: between different gesture recognition methods based on Computer Vision, depth-cameras are often used: an example of use of depth-cameras in gesture recognition is the work of Benko and Wilson [14], where they used the 3DV Systems *ZSense* camera (since June 2009, ZSense is part of Microsoft). The last Microsoft entertainment product, *Kinect*, incorporates a depth-camera; this device, commercialized as a game controller, allows users to interact with the console by moving the body, mainly hands and arms, in the real space.

To make this interaction possible, this device embeds several sensors: a regular VGA video-camera, an infrared projector and a sensor that reads the environment response of an infrared light pattern released by the projector. A freely available API decodes the raw signal and provides developers a digital description of the human body in 3D, recognizes different body parts (head, neck, shoulders, wrists, hands, hips, knees, ankles, and foots) and therefore can create a digital reconstruction of the human skeleton.

In particular, the infrared projector and the video camera that reads infrared response work well also in bad illumination conditions. Furthermore, *Kinect* embeds an array microphone that allows vocal control thanks to proprietary software. So *Kinect* has a high potential, not limited to its use in the game world. That is why we decided to use its features, combined with a video annotation system, to realize our work.

Of course, there are various problems on gesture recognition done with Computer Vision that are not taken into consideration here. For example occlusions, light conditions etc. One main issue in gesture recognition is tracking the person in his/her movements in the room, and this is the main advantage of coupling a depth camera like Kinect to the normal video recording.

Video annotation Videos have a high communicative potential, and therefore they are used as a tool for knowledge acquisition. Indeed it was the availability of cheap video-recording to foster new research on gestures and today we can expect a similar explosion thanks to the introduction of automatic gesture recognition on videos.

Several platforms exist to support researchers in the analysis and annotation of videos, among these we describe here those two that most influenced the design of MORAVIA. *VideoANT* [81] is a web based video annotation tool, characterized by a minimal user interface, it allows free text annotation, and is often adopted for collaborative annotation tasks. However it lacks several interesting functionalities like annotations downloading and a users system. *Anvil* [112] is a desktop annotation tool which offers multi-layered annotation based on a personalized coding scheme. It provides very useful features such as color highlighting for annotations and coding agreement analysis. However, being a desktop based application, co-annotation and project sharing is not always straightforward.

MORAVIA In our context a working group, that may consist of students, teachers or researches, collaborates to the annotation of a video marking significant moments. The video typically contains a teaching session that has to be evaluated. Sessions are recorded using a video camera, and then subsequently they are analysed in order to identify weak points and to suggest improvement. Through this technique, teachers operate an observation on themselves from the professional point of view, becoming aware of the manner in which their competencies are manifested, and manage to identify possible elements that interfere or hinder the training method. Extending this protocol to group evaluation allows to gather many different points of view, so leading to a more effective evaluation.

A further improvement, and our original contribution, is then to exploit, in addition to the plain video, the information on subject's body movements and postures captured by a depth camera (and thus suitable for automatic elaboration). Our proposal consists, as already anticipated, in a tool for quality evaluation of exposition in terms of gestures: this involves the creation of a classification model that, taken in input a video recording containing a speaker who performs an exposition, is able to detect different gestures performed by the same speaker and is able to give a score to the performance.

Now we describe the steps necessary to achieve this goal.

Constructing a Training set The classification model should be trained starting from a training set, that in this case, given the nature of the problem, is composed by several types of gestures and an expositive score associated with each of them. Because we did not have a training set of this type, it was necessary to build it by ourselves: to do gesture capturing we preferred to use techniques of video-recording combined with Computer Vision instead of techniques based on wearable sensors, because these last tend to be more intrusive in the exposition.

We decided to use *Kinect* that, as explained earlier, implement good quality sensors that facilitate the use of techniques of Computer Vision for the recognition of movements by identifying the human skeleton, providing a good enough performance. We expect that the affordable price and the good performance will make it the de-facto standard in a short time and will stimulate a renewed interest in gesture recognition research. Once we captured the gestures, we needed an evaluation about them; those evaluations, to be reliable, must come from experts on educational and psychological domain.

The best way to obtain evaluations is to collaborate with a group of experts. We contacted a group of experts in didactic valuations, that was already executing video recording of expositions.

MORAVIA: Collaborative Video Annotation on the Web After an analysis of existing video annotation tools and having discussed about it with the group, we decided to develop a video annotation software on our own. We identified the following features as essential:

- ease of use and minimal UI; since MORAVIA users may be very different in computer literacy;
- web interface for collaboration; the workgroup may be (and actually is in our case) spread in several departments/cities;
- possibility of downloading annotations to work offline;
- authentication of users; it is necessary to distinguish the attribution of any annotation;
- customizable annotation structure; from the simplest plain text note to a very detailed annotation convention;
- support for common video formats, including *Kinect* ONI format;
- extensibility to include automation filters (such as HMM gesture recognition).

None of the platforms available today have all these features. In figure (Figure A.1) there is a screenshot of MORAVIA with the various parts highlighted: part 1 contains the page header with site navigations commands; part 2 contains

Figure A.1: The main video page of MORAVIA.

RGB video and *Kinect* vision of the current video: this part is a Mockup, at the moment only RGB video can be viewed; part 3 contains the annotations markers and bars: the multi-colored upper bar is still a Mockup, it will reveal with red color video sections where gestures are frequent; part 4 contains additional video controls; part 5 contains annotation management buttons; part 6 contains the textbox to insert new annotation; part 7 contains currently available annotations for the current video. At the moment, the site provides a simple authentication and users system, with a permissions subsystem for videos and annotations; furthermore it provides multi-language support and currently it supports English and Italian languages.

The cooperation with the group included our presence during their classic video-recording sessions: we placed the *Kinect*, paired with a notebook PC, along with the classic video-camera owned by the group. Thus the design of the system was refined and adapted to support the field-work as described so far. Also we managed to collect recorded expositions paired with the skeletal tracking of the speaker, to use as a training testbed.

Once collected enough expositions (with related skeletons and video annotations), we proceeded with the definition of the evaluation model by using the techniques of gesture recognition and gesture classification already exposed in the state of art, and the association of scores to different gestures extrapolated from records provided us by the experts.

Main Issues Among the various difficulties encountered so far we can certainly mention the issues related to video-recording: in addition to classic problems of privacy and loss of naturalness in exposition due to the presence of a video camera, the group of experts often found difficulties to get teachers willing to be filmed and sometimes those which have given the availability gave up at the time of registration. Also, the registration with *Kinect* may cause further loss of naturalness since: it tends to be more cumbersome than standard video-cameras (has to be connected to a PC); it has to be placed closer to the speaker, and the speaker herself have to do a calibration pose of a few seconds necessary for the initial identification of the skeleton.

The group is also doing evaluations mainly on primary school teachers, with a series of additional problems. There are logistic problems, since classrooms sometimes are narrow, it is then difficult to obtain optimal positioning of video-recording; children are distracted by the presence of video-camera and *Kinect*. Issues of privacy are more delicate because of the presence of children, which are sometimes rowdy, and go on purpose in front of the video-camera to get filmed. Sometimes we also incurred in purely technical problems: the *Kinect* was a fairly new technology at the moment this study was carried on (2010) and Microsoft official development tools are not yet available (planned for mid-May 2011), at the moment of writing open source drivers were often causing malfunctions.

Acknowledgement: This section is a revised version of the paper:
Marco Careddu, Laura Carrus, Alessandro Soro, Samuel A. Iacolina, and Riccardo Scateni. *MORAVIA: A video annotation system supporting gesture recognition.* SIGCHI - CHItaly 2011 Adjunct Proceedings, 2011.

Appendix B
Combining Multi-touch and Tangible Interfaces

Many existing systems are bounded to a single modality of interaction, forcing the user to adapt to it. Such approach is targeted at well-defined application scenarios, in which an a-priori study of user behaviour is easier and more effective. However, some systems are meant to be used in different contexts, by different users, with less predictable goals and behaviour. In such cases, the availability of several interaction modalities lets the user choose the appropriate one.

Our project, *Didactic-Highlighter*, has been conceived as a tool for computer/supported education, art and entertainment, with the use of some most recent technologies in the Interaction Design area, like multi-touch and tangible interfaces. The targeted users are, thus, quite heterogeneous, and, most important, are not thought as being extremely familiar with multi-touch tables. The main goal was, hence, to keep the user interface as simple as possible, providing a restricted set of functions, including just the possibility to view and easily manipulate images and graphics, also by drawing and erasing lines, like in a teaching/learning environment.

Combining a drawing table with multi-touch and tangible technologies allows these features to be accessible in an intuitive way, keeping the training phase to almost zero and letting the final user to be immediately in control of the application at first sight. The tangible ones, furthermore, allow an increasing of the level of naturalness of the interaction: drawing on the screen holding an adapted pen is similar to do the same on a paper sheet with ink one.

Drawing Table The multi-touch table consists of a closed wooden box; the tabletop is an acrylic pane, topped by a semi-opaque retro-projection surface. IR LEDs surround the pane (see Figure B.1) and provide the source of infrared light that propagates within the pane.

Figure B.1: The LED frame of the multi-touch table

Figure B.2: The wooden-frame of the interactive table.

The most natural instruments used for drawing on a paper sheet are ink pens and pencils. The IR pen built is analogous to them and allows users to draw on the interactive surface. It has been built by replacing the tip of a thick pen with an IR LED, powered by a common battery. The user switches it on and off by pressing a little button placed on one side. Drawing with this tool is more intuitively than do the same with fingers, without deny to users the latter way. Choosing a new image or a different pen color, erasing of a drawn line, deleting and image, or the well known gestures used to move, scale and rotate an image seems to be, instead, more naturally when done with fingers.

The system was build on the Core Community Vision (CCV) [67] and Tangible User Inter-faces Objects (TUIO) [99] platforms. CCV/tBeta is an open source library for computer vision used for tracking interaction events (e.g.: finger down, moved and released) in building multi-touch applications. It is based on elaborating and analyzing a video stream to detect the boundary of blobs generated by the contact with objects on the surface.

TUIO is a widely used protocol designed to provide a general and versatile communication interface between multi-touch and tangible interactive systems. It works on UDP and defines which types of message and data formats can be exchanged between a tracking server and several client applications. The CCV/tBeta package provides tracking functionalities on the video stream, and the TUIO protocol is used to forward user events to the Didactic-Highlighter. Finally, the MT4J framework [140] provides the possibility of easily composing graphical objects and managing their reaction to input events, with particular emphasis on natural gestures.

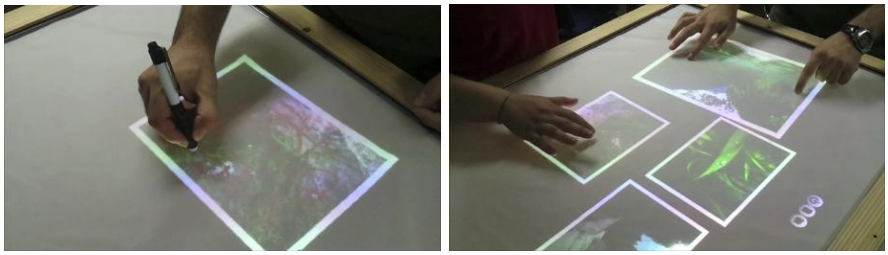


Figure B.3: Collaboration with the *Didactic-Highlighter*

Figure B.3 shows how different users work at once on different images without interfering with each other and anyone is allowed to choose his own preferred modality (touch or pen or both).

Acknowledgement: This section is a revised version of the paper:
Daniela Cabiddu, Giorgio Marcias, Alessandro Soro, and Riccardo Scateni. *Multi-touch and tangible interface: Two different interaction modes in the same system.* CHItaly 2011 Adjunct Proceedings, 2011.

Appendix C
Olfactory Interaction

In real life human beings interact with the world and with each other in a multi-sensory way. They perceive their surroundings through all five senses, leading to a deeply detailed perception. The history of HCI is mainly built over the visual communication channel. Additional auditory and haptic interaction came later and are still regarded as optional companions of the former. Only in rare cases attempts have been made to model multi-sensory environments. In these environments the authors tried to achieve systems where the user is encouraged to get involved in 360°.

There are several application fields where it could be an important limit to receive information only through the visual, auditory, and tactile channels. Examples about these applications are: simulation environments, e.g. for the fire-fighter training work [102]; the use of olfactory icons to show events (Microsoft, *smicons*); applications in wellness, such aromatherapy [91], multi-sensory rooms designed for the care of newborn babies or elderly people [131], and applications in cultural environments such as interactive multi-sensory museums [90].

However, many aspects are still open in terms of technological research, design and evaluation. This experiment was aimed at testing a prototype of interactive multi-sensory environment, built over computer vision techniques for motion detection, multimedia contents and actuators able of spraying aromas in the air. Our main goal is to experiment with this setup while collecting problems and exploring opportunities about unrolling a fully immersive *walk, look and smell* path, in the context of two specific scenarios:

- The implementation of an interactive wine museum;
- The representation of the whole year cycle catching the sensations of changing seasons.

Deploying these interaction scenarios without the olfactory component would be not enough satisfactory for the full sensorial experience. At the same time,

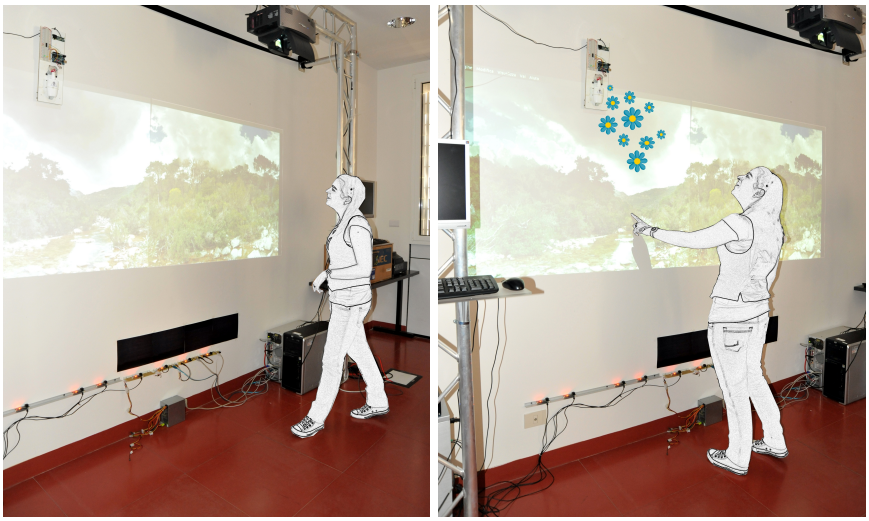


Figure C.1: Representation of the multisensory virtual path

in order to create such a full sensorial experience, it is necessary, somehow, to use a consistent percentage of stimuli present in real situations [158].

Moreover, the sense of smell has the important function of adding a profusion of information to the perception, and mainly to modify the human's emotional state [35, 96]. It is, thus, essential to support applications widely developed and used, with the sense of smell that is the most primitive among the human perceptions.

Related Projects. Making an application that includes the sense of smell within the channels of communication with the user requires one to venture into largely unexplored area; that is due, mainly, to the trouble to understand how the brain is able to elaborate information received by the sense of smell [168], and, afterwards, in which application area it's useful to add these kind of interactions.

The smell allows us to easily draw the attention and to focus on what it's considered important; Kaye suggests a sort of similarity between olfactory displays and "Calm Technologies" [185], suggesting that the perception of a scent changes quickly from the background to the foreground of our attention, and such changes in environmental scents could quickly attract one's attention [102]. Typically, there are two macro areas in HCI in which one can find smell included:

1. Using smell as input with the creation of sensors able to identify the constituents molecules of a complex scent by using of pattern recognition algorithm and that can return this representation of the scent in a digital format; this is the main task of electronic noses; they obtained a large

success in the past two decades and found a large usage in food and wine industries [60];

2. Reproduce scents by means of electronic devices; this is the main task of olfactory displays and its developments; in this way smell is an output channel and it is the focus of our interests here; we will see an overview of these only.

Olfactory displays do the inverse task as electronic noses do. These devices are designed with the goal of synthesizing scents from a digital description [103]. Olfactory displays are used in real life applications, such as, for example, "Incense Clock" of Japan and China life [102]. Olfactory displays are widely used as therapeutic devices in aroma therapy area, "aroma-chology" [91]. These projects are based on the idea that some scents stimulate good feelings in human beings like positivity, creativity, perspicacity.

It is this quality of scents that is mostly tried to be reproduced using electronic systems. It dates back to the 1950's the attempt to use smell in interactive system, with its introduction in cinemas world [102] as an advertising campaign to bring back people to the cinema, after the advent of television. Films were complemented with 3D goggles, vibrating seats and scents emitters. The aim was to provide people with a new approach to the plot.

Even if the "scented films" were not a success, the same technologies have been applied in environments of virtual reality. A seminal example is Sensorama, "an immersive virtual reality" [161], using 3D vision, vibrating seats, and scents, to reproduce real life scenes, such as motorbike ride, or a walk in a flower garden. Another example of virtual reality, as already cited, is the firefighter training developed by Carter and described by Zybura and colleagues in [200]. He focused his work in managing scents emission in quantity and quality; he says "olfactory output is completely proportional from hint of odour to a stench that makes you want to rip the mask off...".

The sense of smell was also introduced in closed spaces, as museum and exhibitions, with the goal to create scents that could help visitors to remember easily all the information acquired during the exhibition [1]. This idea was used in famous museum as Natural History Museum in London, Bow Street Old Whiskey Distillery in Dublin, Jorvik Viking Museum New York. Pletts Haque realized an artistic installation where she used colors and scent to mark off an exhibition area [103].

We can see examples in which smell completes the information received from other senses. The idea to provide information by the sense of smell was also realized in other applications such as "smicon" [102], where every scent is conventionally associated connected to the semantics of an information. For example Microsoft used this idea to create an extension of Outlook in which the sender of an email can be recognized by a smell that is emitted the a message arrives [102].

Our Prototype of an interactive multi-sensory system, supported by presence sensors, multimedia contents and actuators is capable of spreading specific aromas in the air [Figure C.1]. We aim at testing problems and opportunities concerning the realization of immersive paths. It is possible to split the project in the investigation of two different scenarios, interrelated, since they are using the same basic components.

Interactive wine museum. This scene describes an immersive multi-sensory experience inspired by the desire of incrementing the knowledge about wine making regional culture. The user visits a multimedia exhibition; when she comes in a predefined "interaction area", she is involved by several sensations including sounds, images and odours. The stimuli have an evocative and involving nature and enhance the message transmitted by the installation.

The representation of changing seasons This scene describes an experience that is inspired by the cultural heritage of Sardinia. The user visits a multimedia exhibition, when arriving at a transition point, for example an airport, she is caught by typical images of wildlife in Sardinia, with some typical scents of the underwood. The user gets involved in the cycle of seasons, and can look how earth and sea, colors and scents change in time and space.

In our prototype a video-camera captures the user movement; a mechanical activated device, correlated to the camera, is able to spray scent and, thus, reproduce olfactory experiences. Our main initial focus was on the realization of the device that can reproduce scents. The project was intended to model a kind of virtual path, an enabling technology to layout different applicative models for our studies.

The video device is realized using a low-cost webcam. The webcam is positioned above the interaction zone between the user and the virtual representation of the object. The webcam can detect a change in the interaction zone. To do this, the device, continually collects information from the video by means of image processing techniques. Such simple video capturing algorithms, one frame per second of the video, and comparing frames two by two, can be easily deployed to an embedded architecture at a later time. With the comparison, we can obtain differential connected regions from pairs of images in black and white. These connected regions allow understanding if any user is in the interaction zone. When connected regions are identified, and also the presence of the user is recognized in the area, the scents emitter is fired.

The scent emitter device is realized connecting a DC motor to an external micro-controller with a specific programming to provide the necessary pressure to a scent spray [Figure C.2]. We can imagine the motor as a sort of remotely controlled electronic finger, which spays scents, by pressing on a button.

As of today, it is only at a proof of concept stage and, in the future, the prototype will be improved along two directions:

1. We want to be able to setup the video acquisition system to allow to figure out how long a user stays in the interaction area. This information

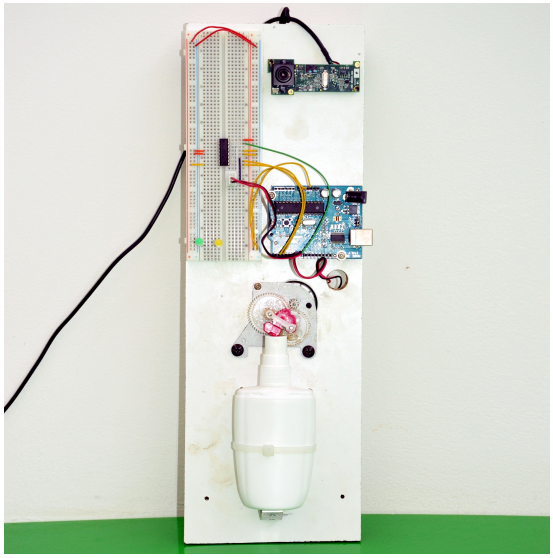


Figure C.2: Prototype scent emitter

will be used to determine how many times it will be necessary to emit scents and in which quantity;

2. A second main improvement will be to modify the hardware device, replacing the “electronic finger” with a piezoelectric device.

On the prototype we will do some measuring to choose optimal positioning of environment emitters. Moreover we will introduce methods of post ventilation in the environment, to leave unchanged the perception for users not affected by the current interaction.

At first, we recorded the paths of the people crossing the environment where we wanted to place the experience; we, then, recorded with a video-camera the movements and the reactions of the people passing by. On average, at the beginning, the system caused a little confusion in the participants; afterwards, the users were keen to play with the interactive system. We tracked the reactions and, overall, they considered the experience complete and enjoyable. This study allows us to claim that our proposal is reasonable, the sense of smell has an important role also in the HCI area and we can proceed toward the realization of more complex virtual experiences.

Acknowledgement: This section is a revised version of the paper:
Valentina Cozza, Gianni Fenu, Riccardo Scateni, and Alessandro Soro. *Walk, look and smell through.* SIGCHI - CHItaly 2011 Adjunct Proceedings, 2011.

Figure C.3: From top to bottom: people approaching the evaluation site, sneaking into the experience area, and smelling the scent

References

[1] John P. Aggleton and Louise Waskett. The ability of odours to serve as state-dependent cues for real-world memories: Can viking smells aid the recall of viking experiences? *British Journal of Psychology*, 90(1):1–7, 1999.

[2] Martha Wagner Alibali and Alyssa A. DiRusso. The function of gesture in learning to count: more than keeping track. *Cognitive Development*, 14(1):37–56, March 1999.

[3] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 3 1992.

[4] Maribeth Back, Jonathan Cohen, Rich Gold, Steve Harrison, and Scott Minneman. Listen reader: an electronically augmented paper-based book. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '01, pages 23–29, New York, NY, USA, 2001. ACM.

[5] Ravin Balakrishnan, George Fitzmaurice, Gordon Kurtenbach, and William Buxton. Digital tape drawing. In *Proceedings of the 12th annual ACM symposium on User interface software and technology*, UIST '99, pages 161–169, New York, NY, USA, 1999. ACM.

[6] Ravin Balakrishnan and Gordon Kurtenbach. Exploring bimanual camera control and object manipulation in 3d graphics interfaces. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 56–62, New York, NY, USA, 1999. ACM.

[7] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. Proxemic interaction: designing for a proximity and orientation-aware environment. In *ACM International Conference on Interactive Tabletops and Surfaces*, ITS '10, pages 121–130, New York, NY, USA, 2010. ACM.

[8] Liam J. Bannon. From human factors to human actors: the role of psychology and human-computer interaction studies in system design. In J. Greenbaum and M. Kyng, editors, *Design at work.: Cooperative Design of Computer Systems*, pages 25–44. Hillsdale: Lawrence Erlbaum Associates, 1991.

[9] Liam J. Bannon. Reimagining hci: toward a more human-centered perspective. *interactions*, 18:50–57, July 2011.

[10] Ditte Amund Basballe and Kim Halskov. Projections on museum exhibits: engaging visitors in the museum setting. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, OZCHI '10, pages 80–87, New York, NY, USA, 2010. ACM.

[11] Michael Beigl, Hans-W. Gellersen, and Albrecht Schmidt. Mediacups: experience with design and use of computer-augmented everyday artefacts. *Computer Networks*, 35(4):401–409, 3 2001.

[12] Steve Benford, John Bowers, Lennart E. Fahlén, Chris Greenhalgh, and Dave Snowdon. User embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '95, pages 242–249, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[13] Steve Benford, Holger Schnädelbach, Boriana Koleva, Rob Anastasi, Chris Greenhalgh, Tom Rodden, Jonathan Green, Ahmed Ghali, Tony Pridmore, Bill Gaver, Andy Boucher, Brendan Walker, Sarah Pennington, Albrecht Schmidt, Hans Gellersen, and Anthony Steed. Expected, sensed, and desired: A framework for designing sensing-based interaction. *ACM Trans. Comput.-Hum. Interact.*, 12:3–30, March 2005.

[14] Hrvoje Benko and Andrew D. Wilson. Depthtouch: Using depth-sensing camera to enable freehand interactions on and above the interactive surface. Technical Report MSR-TR-2009-23, Microsoft, March 2009.

[15] Mark Bilandzic and Marcus Foth. A review of locative media, mobile and embodied spatial interaction. *International Journal of Human-Computer Studies*, 70(1):66–71, 1 2012.

[16] Kirsten Boehner, Rogerio DePaula, Paul Dourish, and Phoebe Sengers. How emotion is made and measured. *Int. J. Hum.-Comput. Stud.*, 65:275–291, April 2007.

[17] Richard A. Bolt. "put-that-there": Voice and gesture at the graphics interface. In *SIGGRAPH '80: Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pages 262–270, New York, NY, USA, 1980. ACM.

[18] Richard A. Bolt and Edward Herranz. Two-handed gesture in multimodal natural dialog. In *UIST '92: Proceedings of the 5th annual ACM symposium on User interface software and technology*, pages 7–14, New York, NY, USA, 1992. ACM.

[19] Carl B. Boyer and Uta C. Merzbach. *A History of Mathematics*. John Wiley and Sons Inc New York, Hoboken edition, 2011.

[20] Christopher J. Bradley. *The Algebra of Geometry: Cartesian, Areal and Projective Co-ordinates*. Highperception, 2007.

[21] Margot Brereton, Nicola Bidwell, Jared Donovan, Brett Campbell, and Jacob Buur. Work at hand : an exploration of gesture in the context of work and everyday life to inform the design of gestural input devices. In Robert Biddle and Bruce Thomas, editors, *Australian User Interface Conference, Conferences in Research and Practice in Information Technology*, pages 1–10, Adelaide, South Australia, 2003. Australian Computer Society Inc.

[22] Duane C. Brown. Decentering distortion of lenses. *Photogrammetric Engineering*, 32(3):444–462, 1966.

[23] Stéphanie Buisine, Guillaume Besacier, Améziane Aoussat, and Frédéric Vernier. How do interactive tabletop systems influence collaboration? *Computers in Human Behavior*, 28(1):49–59, 1 2012.

[24] Jacob Buur and Astrid Soendergaard. Video card game: an augmented environment for user centred design discussions. In *Proceedings of DARE 2000 on Designing augmented reality environments*, DARE '00, pages 63–69, New York, NY, USA, 2000. ACM.

[25] Bill Buxton. Multi-touch systems that i have known and loved (overview). `http://www.billbuxton.com/multitouchOverview.html`, 2007.

[26] William A. S. Buxton and Brad A. Myers. A study in two-handed input. *SIGCHI Bull.*, 17(4):321–326, 1986.

[27] Daniela Cabiddu, Giorgio Marcias, Alessandro Soro, and Riccardo Scateni. Multi-touch and tangible interface: Two different interaction modes in the same system. CHItaly 2011 Adjunct Proceedings, 2011.

[28] Claude Cadoz. Le geste canal de communication homme/machine: la communication 'instrumentale'. *TSI. Technique et science informatiques*, 13(1):31–61, 1994.

[29] Sebastien Carbini, Lionel Delphin-Poulat, Laurence Perron, and Jean Emmanuel Viallet. From a wizard of oz experiment to a real time speech and gesture multimodal interface. *Signal Processing*, 86(12):3559–3577, 12 2006.

[30] Marco Careddu, Laura Carrus, Alessandro Soro, Samuel A. Iacolina, and Riccardo Scateni. Moravia: A video-annotation system supporting gesture recognition. SIGCHI - CHItaly 2011 Adjunct Proceedings, 2011.

[31] Bill Casselman. Ybc 7289. `http://www.math.ubc.ca/~cass/Euclid/ybc/ybc.html`, 2001.

[32] Ginevra Castellano, Loic Kessous, and George Caridakis. Emotion recognition through multiple modalities: Face, body gesture, speech. In Christian Peter and Russell Beale, editors, *Affect and Emotion in Human-Computer Interaction*, volume 4868 of *Lecture Notes in Computer Science*, pages 92–103. Springer Berlin / Heidelberg, 2008.

[33] Umberto Castiello. The neuroscience of grasping. *Nat Rev Neurosci*, 6(9):726–736, 09 2005.

[34] Jaz Hee-jeong Choi and Eli Blevis. Advancing design for sustainable food cultures. In Marcus Foth, Laura Forlano, Christine Satchell, and Martin Gibbs, editors, *From Social Butterfly to Engaged Citizen : Urban Informatics, Social Media, Ubiquitous Computing, and Mobile Technology to Support Citizen Engagement*. MIT Press, Cambridge, Mass, 2011.

[35] Simon Chu and John J. Downes. Odour-evoked autobiographical memories: Psychological investigations of proustian phenomena. *Chemical Senses*, 25(1):111–116, 2000.

[36] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. Gesturing makes learning last. *Cognition*, 106(2):1047 – 1058, 2008.

[37] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32 –80, jan 2001.

[38] Valentina Cozza, Gianni Fenu, Riccardo Scateni, and Alessandro Soro. Walk, look and smell through. SIGCHI - CHItaly 2011 Adjunct Proceedings, 2011.

[39] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of oz studies - why and how. *Knowledge-Based Systems*, 6(4):258–266, 12 1993.

[40] Kelly L. Dempski and Brandon Harvey. Supporting collaborative touch interaction with high resolution wall displays. In *Proceedings of the 2nd Workshop on Multi-User and Ubiquitous User Interfaces (MU3I)*, 2005.

[41] Michael B. Denlinger and Haworth NJ. Ambient-light-responsive touch screen data input method and system, 1986/10/02 1988.

[42] Massimo Deriu, Gavino Paddeu, and Alessandro Soro. Xplaces: An open framework to support the digital living at home. In *Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*, GREENCOM-CPSCOM '10, pages 484–487, Washington, DC, USA, 2010. IEEE Computer Society.

[43] Paul H. Dietz, Hopkinton MA, and Darren L. Leigh. Multi-user touch surface, 2001/05/24 2002.

[44] Paul Dourish. *Accounting for system behavior: representation, reflection, and resourceful action*, pages 145–170. MIT Press, Cambridge, MA, USA, 1997.

[45] Paul Dourish. What we talk about when we talk about context. *Personal Ubiquitous Comput.*, 8:19–30, February 2004.

[46] Paul Dourish. *Where the Action Is: The Foundations of Embodied Interaction*. The MIT Press, new edition edition, September 2004.

[47] Florian Echtler, Manuel Huber, and Gudrun Klinker. Shadow tracking on multi-touch tables. In *AVI '08: Proceedings of the working conference on Advanced visual interfaces*, pages 388–391, New York, NY, USA, 2008. ACM.

[48] David Efron. *Gesture and Environment*. King's Crown Press, Morningside Heights, New York, 1941.

[49] Stacy B. Ehrlich, Susan C. Levine, and Susan Goldin-Meadow. The importance of gesture in children's spatial reasoning. *Developmental Psychology*, 42(6):1259–1268, November 2006.

[50] Paul Ekman and Wallace V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica*, 1(1):49–98, 1969.

[51] Brenda Farnell. Moving bodies, acting selves. *Annual Review of Anthropology*, 28:pp. 341–373, 1999.

[52] Fredrik Willem Fikkert. *Gesture Interaction at a Distance*. PhD thesis, Universiteit Twente, Enschede, March 2010. SIKS Dissertation Series No. 2010-07.

[53] George W. Fitzmaurice. *Graspable user interfaces*. PhD thesis, University of Toronto, Toronto, Ont., Canada, Canada, 1996. Adviser-Buxton, William.

[54] George W. Fitzmaurice, Hiroshi Ishii, and William A. S. Buxton. Bricks: laying the foundations for graspable user interfaces. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 442–449, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[55] Morten Fjeld and Wolmet Barendregt. Epistemic action: A measure for cognitive support in tangible user interfaces? *Behavior Research Methods*, 41:876–881, 2009. 10.3758/BRM.41.3.876.

[56] Morten Fjeld, Martin Bichsel, and M. Rauterberg. Build-it: a brick-based tool for direct interaction. In *Engineering Psychology and Cognitive Ergonomics (EPCE*, pages 205–212. Hampshire: Ashgate, 1986.

[57] Obama for America campaign. Barack obama facebook page. `http://www.facebook.com/barackobama`, 2011.

[58] David Fowler and Eleanor Robson. Square root approximations in old babylonian mathematics: Ybc 7289 in context. *HISTORIA MATHEMATICA*, 25:366–378, 1998.

[59] Norman M. Fraser and G. Nigel Gilbert. Simulating speech systems. *Computer Speech & Language*, 5(1):81–99, 1 1991.

[60] Julian W. Gardner and Philip N. Bartlett. *Electronic noses. Principles and Applications*. Oxford University Press, 1999.

[61] R. Allen Gardner and Beatrice T. Gardner. Teaching sign language to a chimpanzee. *Science*, 165(3894):664–672, Aug 1969.

[62] Bill Gaver. Interaction research studio: : Goldsmiths, university of london. *interactions*, 18:84–87, January 2011.

[63] S. Goldin-Meadow, H. Nusbaum, S.D. Kelly, and S. Wagner. Explaining math: Gesturing lightens the load. *Psychological Science*, 12(6):516–522, November 2001.

[64] Susan Goldin-Meadow. The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11):419 – 429, 1999.

[65] Charles Goodwin. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522, 9 2000.

[66] Raphael Grasset, Andreas Duenser, Hartmut Seichter, and Mark Billinghurst. The mixed reality book: a new multimedia reading experience. In *CHI '07: CHI '07 extended abstracts on Human factors in computing systems*, pages 1953–1958, New York, NY, USA, 2007. ACM.

[67] The NUI Group. Community core vision platform. `http://ccv.nuigroup.com/`.

[68] Jonathan Grudin. The computer reaches out: the historical continuity of interface design. In *CHI '90: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 261–268, New York, NY, USA, 1990. ACM.

[69] Yves Guiard. Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model. *Journal of Motor Behavior*, 19:486–517, 1987.

[70] Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 11 2007.

[71] Jefferson Y. Han. Low-cost multi-touch sensing through frustrated total internal reflection. In *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 115–118, New York, NY, USA, 2005. ACM.

[72] Jefferson Y. Han. Multi-touch interaction wall. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Emerging technologies*, page 25, New York, NY, USA, 2006. ACM.

[73] Alexander Georg Hauptmann. Speech and gestures for graphic image manipulation. *SIGCHI Bull.*, 20:241–245, March 1989.

[74] Karen Henricksen, Jadwiga Indulska, Andry Rakotonirainy, Friedemann Mattern, and Mahmoud Naghshineh. *Modeling Context Information in Pervasive Computing Systems*, volume 2414, pages 79–117. Springer Berlin / Heidelberg, 2002.

[75] Ken Hinckley, Randy Pausch, Dennis Proffitt, James Patten, and Neal Kassell. Cooperative bimanual action. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '97, pages 27–34, New York, NY, USA, 1997. ACM.

[76] Ken Hinckley, Koji Yatani, Michel Pahud, Nicole Coddington, Jenny Rodenhouse, Andy Wilson, Hrvoje Benko, and Bill Buxton. Pen + touch = new tools. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 27–36, New York, NY, USA, 2010. ACM.

[77] Andrew Hodges. *Alan Turing : the enigma of intelligence*. HarperCollins Publishers Ltd, 1985.

[78] Kay Hofmeester and Dennis Wixon. Using metaphors to create a natural user interface for microsoft surface. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, pages 4629–4644, New York, NY, USA, 2010. ACM.

[79] E. Hornecker. I don't understand it either, but it is cool - visitor interactions with a multi-touch table in a museum. In *Horizontal Interactive Human Computer Systems, 2008. TABLETOP 2008. 3rd IEEE International Workshop on*, pages 113 –120, oct. 2008.

[80] Eva Hornecker and Jacob Buur. Getting a grip on tangible interaction: a framework on physical space and social interaction. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pages 437–446, New York, NY, USA, 2006. ACM.

[81] Bradford Hosack. VideoANT: Extending online video annotation beyond content delivery. *TechTrends*, 54(3):45–49, 2010.

[82] Autumn Hostetter and Martha Alibali. Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin and Review*, 15:495–514, 2008. 10.3758/PBR.15.3.495.

[83] Johanna Höysniemi, Perttu Hämäläinen, and Laura Turkki. Wizard of oz prototyping of computer vision based action games for children. In *Proceedings of the 2004 conference on Interaction design and children: building a community*, IDC '04, pages 27–34, New York, NY, USA, 2004. ACM.

[84] Greg Humphreys, Mike Houston, Ren Ng, Randall Frank, Sean Ahern, Peter D. Kirchner, and James T. Klosowski. Chromium: a stream-processing framework for interactive rendering on clusters. *ACM Trans. Graph.*, 21(3):693–702, 2002.

[85] Edwin L. Hutchins, James D. Hollan, and Donald A. Norman. Direct manipulation interfaces. *Hum.-Comput. Interact.*, 1:311–338, December 1985.

[86] Samuel A. Iacolina, Alessandro Lai, Alessandro Soro, and Riccardo Scateni. Natural interaction and computer graphics applications. In *Eurographics Italian Chapter Conference*, pages 141–146, 2010.

[87] Samuel A. Iacolina, Alessandro Soro, and Riccardo Scateni. Improving ftir based multi-touch sensors with ir shadow tracking. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems*, EICS '11, pages 241–246, New York, NY, USA, 2011. ACM.

[88] Samuel A. Iacolina, Alessandro Soro, and Riccardo Scateni. Natural exploration of 3d models. In *Proceedings of the 9th ACM SIGCHI Italian Chapter International Conference on Computer-Human Interaction: Facing Complexity*, CHItaly, pages 118–121, New York, NY, USA, 2011. ACM.

[89] Corning Incorporated. Future technology watch your day in 2020. `http://www.youtube.com/watch?v=bBjvqnKQsTI`.

[90] Naomi Inoue. Application of ultra-realistic communication research to digital museum. In *Proceedings of the 9th ACM SIGGRAPH Conference on Virtual-Reality Continuum and its Applications in Industry*, VRCAI '10, pages 29–32, New York, NY, USA, 2010. ACM.

[91] Alice M. Isen, F.Gregory Ashby, and Elliot Waldron. The sweet smell of success. *The Aromachology Review*, VI(3):1, 1997.

[92] Hiroshi Ishii. Tangible bits: beyond pixels. In *Proceedings of the 2nd international conference on Tangible and embedded interaction*, TEI '08, pages xv–xxv, New York, NY, USA, 2008. ACM.

[93] Hiroshi Ishii and Brygg Ullmer. Tangible bits: towards seamless interfaces between people, bits and atoms. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 234–241, New York, NY, USA, 1997. ACM.

[94] Hirotaka Iuchi, Sakashi Maeda, and Naoyuki Tsuruta. Gesture recognition using Self-Organizing Maps and Hidden Markov Model. *IPSJ SIG Notes, Computer Vision and Image Media*, 2001(36):127–134, 2001.

[95] Jana M. Iverson and Susan Goldin-Meadow. Why people gesture when they speak. *Nature*, 396(6708), 11 1998.

[96] Tim Jacob. A tutorial on the sense of smell. `http://www.cf.ac.uk/biosi/staffinfo/jacob/teaching/sensory/olfact1.html`, 2007.

[97] Giulio Jacucci, Antti Oulasvirta, and Antti Salovaara. Active construction of experience through mobile media: a field study with implications for recording and sharing. *Personal Ubiquitous Comput.*, 11:215–234, April 2007.

[98] Sachin H. Jain. Practicing medicine in the age of facebook. *New England Journal of Medicine*, 361(7):649–651, 2009.

[99] Martin Kaltenbrunner and Ross Bencina. reactivision: a computer-vision framework for table-based tangible interaction. In *TEI '07: Proceedings of the 1st international conference on Tangible and embedded interaction*, pages 69–74, New York, NY, USA, 2007. ACM.

[100] Maria Karam. *A framework for research and design of gesture-based human-computer interactions*. PhD thesis, University of Southampton, 2006.

[101] Leonard R. Kasday and Plainsboro NJ. Touch position sensitive surface, 1981/12/23 1984.

[102] Joseph "Jofish" Kaye. Making scents: aromatic output for hci. *interactions*, 11:48–61, January 2004.

[103] Joseph Nathaniel Kaye. Symbolic olfactory display. Master's thesis, Media Lab, Massachusetts Institute of Tech, 2001.

[104] Adam Kendon. Differential perception and attentional frame in face-to-face interaction: Two problems for investigation. *Semiotica*, 24:305–315, 1978.

[105] Adam Kendon. Gesture. *Journal of Visual Verbal Language*, 3(1):21–36, 1983.

[106] Adam Kendon. Current issues in the study of gesture. In J-L Nespoulous, P. Peron, and Lecours A.R., editors, *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, pages 23–47. Lawrence Erlbaum, Associates, Hillsdale, New Jersey, 1986.

[107] Adam Kendon. Some reasons for studying gesture. *Semiotica*, 62:1–28, 1986.

[108] Adam Kendon. On gesture: Its complementary relationship with speech. In A. Seigman and S. Feldstein, editors, *Nonverbal Communication*, pages 65–97. Lawrence Erlbaum, Associates, Hillsdale, New Jersey, 1987.

[109] Adam Kendon. How gestures can become like words. In F. Poyatos, editor, *Crosscultural Perspectives in Nonverbal Communication*, pages 131–141. C. J. Hogrefe, Publishers, 1988.

[110] Adam Kendon. An agenda for gesture studies. *The Semiotic Review of Books*, 7(3):7–12, 1996.

[111] Adam Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[112] Michael Kipp. Multimedia annotation, querying and analysis in anvil. In M. Maybury, editor, *Multimedia Information Extraction*, chapter 19. IEEE Computer Society Press, 2010.

[113] David Kirsh and Paul Maglio. On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18(4):513–549, December 1994.

[114] Jesper Kjeldskov and Connor Graham. A review of mobile hci research methods. In Luca Chittaro, editor, *Human-Computer Interaction with Mobile Devices and Services*, volume 2795 of *Lecture Notes in Computer Science*, pages 317–335. Springer Berlin / Heidelberg, 2003.

[115] Scott R. Klemmer, Jamey Graham, Gregory J. Wolff, and James A. Landay. Books with voices: paper transcripts as a physical interface to oral histories. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 89–96, New York, NY, USA, 2003. ACM.

[116] Mark L. Knapp and Judith A. Hall. *Nonverbal communication in human interaction*. Wadsworth/Thomson Learning, 2001.

[117] Teuvo Kohonen. The self-organizing map. In *Proceedings of the IEEE*, volume 78, pages 1464–1479, September 1990.

[118] Myron W. Krueger. Responsive environments. In *Proceedings of the June 13-16, 1977, national computer conference*, AFIPS '77, pages 423–433, New York, NY, USA, 1977. ACM.

[119] Myron W. Krueger, Thomas Gionfriddo, and Katrin Hinrichsen. Videoplace—an artificial reality. *SIGCHI Bull.*, 16(4):35–40, 1985.

[120] Gordon Kurtenbach and Eric A. Hulteen. Gestures in human-computer communications. In Brenda Laurel, editor, *The Art of Human Computer Interface Design*, pages 309–317. Addison-Wesley, 1990.

[121] Marianne LaFrance. *Nonverbal communication.*, pages 463 – 466. American Psychological Association, 2000.

[122] Alessandro Lai, Alessandro Soro, and Riccardo Scateni. Interactive calibration of a multi-projector system in a video-wall multi-touch environment. In *Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 437–438, New York, NY, USA, 2010. ACM.

[123] Larry Larsen. Interview to bill buxton on nuis. `http://channel9.msdn.com/Blogs/LarryLarsen/CES-2010-NUI-with-Bill-Buxton`, 2010.

[124] Seonkyoo Lee, William Buxton, and Kenneth C. Smith. A multi-touch three dimensional touch-sensitive tablet. *SIGCHI Bull.*, 16(4):21–25, 1985.

[125] Sören Lenman, Lars Bretzner, and Björn Thuresson. Computer vision based hand gesture interfaces for human-computer interaction. *CID Stockholm Sweden*, 2002.

[126] Nianjun Liu and Brian C. Lovell. Gesture classification using hidden markov models. In *Proceedings of the Seventh Biennial Australian Pattern Recognition Society Conference*, 2003.

[127] Tim Love. Ansi c for programmers on unix systems. `ftp://svr-www.eng.cam.ac.uk/misc/love_C.ps.Z`, 2010.

[128] Paul P. Maglio and David Kirsh. Epistemic action increases with skill. In *In Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pages 391–396. Erlbaum, 1996.

[129] Aditi Majumder and Michael S. Brown. *Practical Multi-projector Display Design*. A. K. Peters, Ltd., Na-tick, MA, USA, 2007.

[130] Jennifer Mankoff, Deanna Matthews, Susan R. Fussell, and Michael Johnson. Leveraging social networks to motivate individuals to reduce their ecological footprints. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, page 87, jan. 2007.

[131] Patrizia Marti, Henrik Hautop Lund, Margherita Bacigalupo, Leonardo Giusti, and Claudio Mennecozzi. Blending senses: A multi-sensory environment for the tratment of dementia affected subjects. *Journal of Gerontechnology*, (6(1)):33–41, January 2007.

[132] Takefumi Matsunaga and Oshita Masaki. Recognition of walking motion using support vector machine. In *Proceedings of ISIC 2007*, pages 337–342, 2007.

[133] Ali Mazalek, Glorianna Davenport, and Hiroshi Ishii. Tangible viewpoints: a physical approach to multimedia stories. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 153–160, New York, NY, USA, 2002. ACM.

[134] David McNeill. So you think gestures are nonverbal?. *Psychological Review*, 92(3):350 – 371, 1985.

[135] David Mcneill. *Hand and Mind: What Gestures Reveal about Thought*. University Of Chicago Press, August 1992.

[136] David McNeill. *Gesture and thought*. University of Chicago Press, 2005.

[137] Microsoft. Microsoft at 2020. `http://www.youtube.com/watch?v=cZanf5FkkkA`.

[138] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE transactions on systems, man and cybernetics, Part C, Applications and reviews*, 37(3):311–324, May 2007.

[139] Cecily Morrison, Matthew Jones, Alan Blackwell, and Alain Vuylsteke. Electronic patient record use during ward rounds: a qualitative study of interaction between medical staff. *Critical Care*, 12(6), 2008.

[140] MT4j. Multitouch for java. `http://www.mt4j.org/mediawiki/index.php/Main_Page`.

[141] Jean-Luc Nespoulous, Paul Peron, and Lecours André Roch. Gestures: Nature and function. In J-L Nespoulous, P. Peron, and Lecours A.R., editors, *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, pages 49–62. Lawrence Erlbaum, Hillsdale, NJ, 1986.

[142] Nokia Corporation. Nokia mixed reality - nokia world. `http://www.youtube.com/watch?v=CGwvZWyLiBU`.

[143] Nokia Corporation. Qt framework. `http://qt.nokia.com/`.

[144] Kenji Oka, Yoichi Sato, and Hideki Koike. Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems. In *FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 429, Washington, DC, USA, 2002. IEEE Computer Society.

[145] David R. Olson. Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77(4):257–273, 1970.

[146] Masaki Oshita and Takefumi Matsunaga. Automatic learning of gesture recognition model using SOM and SVM. In *6th International Symposium on Visual Computing 2010 (Lecture Notes in Computer Science 6453)*, pages 751–760, November 2010.

[147] Antti Oulasvirta, Sara Estlander, and Antti Nurminen. Embodied interaction with a 3d versus 2d mobile map. *Personal Ubiquitous Comput.*, 13(4):303–320, 2009.

[148] Kees Overbeeke, Tom Djajadiningrat, Caroline Hummels, and Stephan Wensveen. *Beauty In Usability: Forget About Ease Of Use!*, pages 9–18. Taylor & Francis, 2000.

[149] Russell Owen, Gordon Kurtenbach, George Fitzmaurice, Thomas Baudel, and Bill Buxton. When it gets more difficult, use both hands: exploring bimanual curve manipulation. In *GI '05: Proceedings of Graphics Interface 2005*, pages 17–24, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2005. Canadian Human-Computer Communications Society.

[150] James Patten and Hiroshi Ishii. A comparison of spatial organization strategies in graphical and tangible user interfaces. In *DARE '00: Proceedings of DARE 2000 on Designing augmented reality environments*, pages 41–50, New York, NY, USA, 2000. ACM.

[151] Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):677–695, 1997.

[152] Peter Peltonen, Esko Kurvinen, Antti Salovaara, Giulio Jacucci, Tommi Ilmonen, John Evans, Antti Oulasvirta, and Petri Saarikko. It's mine, don't touch!: interactions at a large multi-touch display in a city centre. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1285–1294, New York, NY, USA, 2008. ACM.

[153] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. Multimodal human discourse: gesture and speech. *ACM Trans. Comput.-Hum. Interact.*, 9(3):171–193, 2002.

[154] Francis K. H. Quek. Toward a vision-based hand gesture interface. In *Proceedings of the conference on Virtual reality software and technology*, pages 17–31, River Edge, NJ, USA, 1994. World Scientific Publishing Co., Inc.

[155] Francis K. H. Quek. Eyes in the interface. *Image and Vision Computing*, 13(6):511–525, 8 1995.

[156] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296, 1990.

[157] Lawrence R. Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, April 2003.

[158] Belma Ramic-Brkic and Alan Chalmers. Virtual smell: authentic smell diffusion in virtual environments. In *Proceedings of the 7th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, AFRIGRAPH '10, pages 45–52, New York, NY, USA, 2010. ACM.

[159] Jef Raskin. Viewpoint: Intuitive equals familiar. *Commun. ACM*, 37:17–18, September 1994.

[160] Frances H. Rauscher, Robert M. Krauss, and Yihsiu Chen. Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7(4):226–231, 07 1996.

[161] Howard Rheingold. *Virtual reality*. Simon & Schuster, Inc., New York, NY, USA, 1991.

[162] Jochen Rick, Amanda Harris, Paul Marshall, Rowanne Fleck, Nicola Yuill, and Yvonne Rogers. Children designing together on a multi-touch tabletop: an analysis of spatial orientation and user interactions. In *Proceedings of the 8th International Conference on Interaction Design and Children*, IDC '09, pages 106–114, New York, NY, USA, 2009. ACM.

[163] Gerhard Rigoll, Andreas Kosmala, and Stefan Eickeler. High performance real-time gesture recognition using hidden markov models. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 69–80, London, UK, 1998. Springer-Verlag.

[164] Toni Robertson. Cooperative work and lived cognition: a taxonomy of embodied actions. In *Proceedings of the fifth conference on European Conference on Computer-Supported Cooperative Work*, pages 205–220, Norwell, MA, USA, 1997. Kluwer Academic Publishers.

[165] Yvonne Rogers and Siân Lindley. Collaborating around vertical and horizontal large interactive displays: which way is best? *Interacting with Computers*, 16(6):1133–1152, 12 2004.

[166] Daniel Salber and Joëlle Coutaz. Applying the wizard of oz technique to the study of multimodal systems. In Leonard Bass, Juri Gornostaev, and Claus Unger, editors, *Human-Computer Interaction*, volume 753 of *Lecture Notes in Computer Science*, pages 219–230. Springer Berlin / Heidelberg, 1993.

[167] Mahadev Satyanarayanan. Pervasive computing: vision and challenges. *Personal Communications, IEEE*, 8(4):10 –17, aug 2001.

[168] Susan S. Schiffman and Tim C. Pearce. *Introduction to Olfaction: Perception, Anatomy, Physiology, and Molecular Biology*, pages 1–31. Wiley-VCH Verlag GmbH & Co. KGaA, 2004.

[169] Johannes Schöning, Peter Brandl, Florian Daiber, Florian Echtler, Otmar Hilliges, Jonathan Hook, Markus Löchtefeld, Nima Motamedi, Laurence Muller, Patrick Olivier, Tim Roth, and Ulrich von Zadow. Multi-touch surfaces: A technical guide. Technical Report TUM-I0833, University of Münster, 2008.

[170] Alessandro Soro, Massimo Deriu, and Gavino Paddeu. Natural exploration of multimedia contents. In *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*, MoMM '09, pages 382–385, New York, NY, USA, 2009. ACM.

[171] Alessandro Soro, Samuel Aldo Iacolina, Riccardo Scateni, and Selene Uras. Evaluation of user gestures in multi-touch interaction: a case study in pair-programming. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pages 161–168, New York, NY, USA, 2011. ACM.

[172] Alessandro Soro, Gavino Paddeu, and Mirko Lobina. Multitouch sensing for collaborative interactive walls. In Peter Forbrig, Fabio Paternò, and Annelise Pejtersen, editors, *Human-Computer Interaction Symposium*, volume 272 of *IFIP International Federation for Information Processing*, pages 207–212. Springer Boston, 2008.

[173] Thad Starner, Alex Pentland, and Joshua Weaver. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20:1371–1375, December 1998.

[174] Ivan E. Sutherland. Sketchpad: a man-machine graphical communication system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, AFIPS '63 (Spring), pages 329–346, New York, NY, USA, 1963. ACM.

[175] Opencv Dev Team. Open source computer vision library - reference manual. `http://opencv.itseez.com/`, 2011.

[176] Yukitaka Toyokura and Yoshihiko Nankaku et al. Approach to japanese sign language word recognition using basic motion HMM. In *Proceedings of the Society Conference of IEICE*, volume 2006, page 72, 2006.

[177] Matthew Turk. Gesture recognition. In K. Stanney, editor, *Handbook of Virtual Environment Technology*. Lawrence Erlbaum Associates, 2001.

[178] Brygg Ullmer and Hiroshi Ishii. mediablocks: tangible interfaces for online media. In *CHI '99 extended abstracts on Human factors in computing systems*, CHI EA '99, pages 31–32, New York, NY, USA, 1999. ACM.

[179] Brygg Ullmer and Hiroshi Ishii. Emerging frameworks for tangible user interfaces. *IBM Syst. J.*, 39(3-4):915–931, 2000.

[180] Bret Victor. A brief rant on the future of interaction design. `http://worrydream.com/ABriefRantOnTheFutureOfInteractionDesign`, November 2011.

[181] Daniel Vogel and Ravin Balakrishnan. Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In *Proceedings of the 17th annual ACM symposium on*

User interface software and technology, UIST '04, pages 137–146, New York, NY, USA, 2004. ACM.

[182] Anthony I. Wasserman. The design of 'idiot-proof' interactive programs. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*, AFIPS '73, pages m34–m38, New York, NY, USA, 1973. ACM.

[183] Whitney M. Weikum, Athena Vouloumanos, Jordi Navarra, Salvador Soto-Faraco, Núria Sebastián-Gallés, and Janet F. Werker. Visual language discrimination in infancy. *Science*, 316(5828):1159, 2007.

[184] Mark Weiser. The computer for the 21st century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3(3):3–11, 1999.

[185] Mark Weiser and John Seely Brown. *The coming age of calm technology*, pages 75–86. Copernicus - Springer-Verlag, New York, 1997.

[186] Pierre Wellner. The digitaldesk calculator: tangible manipulation on a desk top display. In *UIST '91: Proceedings of the 4th annual ACM symposium on User interface software and technology*, pages 27–33, New York, NY, USA, 1991. ACM.

[187] Pierre Wellner. Interacting with paper on the digitaldesk. *Commun. ACM*, 36:87–96, July 1993.

[188] Alan Wexelblat. Research challenges in gesture: Open issues and unsolved problems. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 1–11, London, UK, 1998. Springer-Verlag.

[189] Laurie Williams and Robert Kessler. *Pair Programming Illuminated*. Addison-Wesley, New York, 2003.

[190] Andrew D. Wilson. Touchlight: an imaging touch screen and display for gesture-based interaction. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, pages 69–76, New York, NY, USA, 2004. ACM.

[191] Andrew D. Wilson. Playanywhere: a compact interactive tabletop projection-vision system. In *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 83–92, New York, NY, USA, 2005. ACM.

[192] Andrew D. Wilson and Hrvoje Benko. Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 273–282, New York, NY, USA, 2010. ACM.

[193] Andrew D. Wilson and Aaron F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21:884–900, September 1999.

[194] Terry Winograd. *The design of interaction*, pages 149–161. Copernicus - Springer-Verlag, New York, NY, USA, 1997.

[195] Ying Wu and Thomas S. Huang. Vision-based gesture recognition: A review. In *GW '99: Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pages 103–115, London, UK, 1999. Springer-Verlag.

[196] Jie Yang and Yangsheng Xu. Hidden markov model for gesture recognition. Technical report, The Robotics Institute - Carngie Mellon University., May 1994.

[197] Robert Zeleznik, Andrew Bragdon, Ferdi Adeputra, and Hsu-Sheng Ko. Hands-on math: A page-based multi-touch and pen desktop for technical work and problem solving. In *UIST2010*, 2010.

[198] Robert C. Zeleznik, Kenneth P. Herndon, and John F. Hughes. Sketch: an interface for sketching 3d scenes. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Courses*, page 9, New York, NY, USA, 2006. ACM.

[199] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, pages 493–502, New York, NY, USA, 2007. ACM.

[200] Martin Zybur and Eskeland Gunnar. A. Olfaction for virtual reality. *Quarter Project, Industrial Engineering, University of Washington*, 1999.