Workshop on the Challenges of Engineering Multimodal Interaction: Methods, Tools, Evaluation

Wednesday 13 May

Fraunhofer Institute for Applied Information Technology

Sankt Augustin (Bonn), Germany
Workshop on the Challenges of Engineering Multimodal Interaction: Methods, Tools, Evaluation

Workshop Overview

The OpenInterface Project (http://www.oi-project.org/), an EU-funded STREP, has been developing a run-time environment and toolset designed to help multimodal interaction developers to explore interaction alternatives and to combine and configure them appropriately for different contexts of use. OI has been running since October 2006 and is now coming to an end. This workshop is intended both as a showcase for OI's work and as forum for discussing the current state of the art in multimodal interaction development and its current and future research challenges.

Programme Overview

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00-09:30</td>
<td>Welcome</td>
</tr>
<tr>
<td>09:30-10:15</td>
<td>Keynote</td>
</tr>
<tr>
<td>10:15-10:30</td>
<td>Break</td>
</tr>
<tr>
<td>10:30-11:00</td>
<td>Talks - (1) New Interaction Techniques</td>
</tr>
<tr>
<td>11:00-12:00</td>
<td>Talks - (2) Design and Tools</td>
</tr>
<tr>
<td>12:00-13:00</td>
<td>Lunch (Posters and Demos)</td>
</tr>
<tr>
<td>13:00-14:00</td>
<td>Talks - (3) Engineering</td>
</tr>
<tr>
<td>14:00-15:00</td>
<td>Talks - (4) Software Technologies and Evaluation</td>
</tr>
<tr>
<td>15:00-16:00</td>
<td>Challenges (Discussion session)</td>
</tr>
<tr>
<td>16:00-17:00</td>
<td>Report back and wrap up</td>
</tr>
<tr>
<td>17:00-19:00</td>
<td>Demo Session</td>
</tr>
</tbody>
</table>

Workshop Aims

The workshop will provide the opportunity to present and discuss existing and novel methods and tools for the design, development, and evaluation of modern multimodal applications and interfaces.

Topics addressed in the workshop include, but are not limited to:

- User centered methods for the design of multimodal applications
- Methods for evaluation of multimodal applications
- Tools to support multimodal interaction designers and developers
- Methods and tools to bridge the gap between research and industry
Workshop Format

The one-day workshop will be organized around a series of short presentations (8 mins per talk) of research work in the area of multimodal interaction. We will begin with a keynote on the motivations for researching methods and tools for supporting multimodal interaction design, development and evaluation. Each of the paper sessions will then focus on each of (1) New Interaction Techniques, (2) Design, and tools for multimodal interaction design and development, (3) Engineering multimodal interaction, and (4) Software technologies. The afternoon session will involve a discussion in groups on the challenges of engineering multimodal interaction. Finally, you are all invited to visit the demos by the Open Interface partners and workshop participants from 17:00-19:00.

Speakers - Each paper has been assigned to a paper session. Please get in touch with us by email (mcgeemr@dcs.gla.ac.uk) if for any reason the time or session does not work for you. Digital projection facilities will be provided. Please get in touch if you have any special requirements.

All participants - you are encouraged to bring a demo and/or poster to be set up before the lunch break in order to share and discuss your research (optional). Please get in touch if you plan to bring a demo and/or poster. These should be set up in the morning before the highlight on the demos and posters during the lunch break.

Expected outcomes

The first outcome of the workshop is community building for researchers and practitioners interested in developing methods and tools for the design, development, and evaluation of multimodal applications.

A selection of paper authors will be invited to submit extended papers to be published in a special issue of the Springer Journal on Multimodal User Interfaces to appear before the end of 2009.

Workshop Organizers

Sebastian Denef, Fraunhofer FIT, Germany
Philip Gray, University of Glasgow, UK
Marilyn McGee-Lennon, University of Glasgow, UK
Laurence Nigay, Universite Joseph Fourier, France

For further information please email Dr Marilyn Rose McGee-Lennon - mcgeemr@dcs.gla.ac.uk
Detailed Programme - Talks

09:30-10:15  **Keynote**  
*Interaction in the mobile context: some challenges for multimodality*  
Professor Luca Chittaro, Head of the HCI Lab, University of Udine, Italy.

10:15-10:30  **Break**

10:30-11:00  **Talks - (1) New Interaction Techniques**  
*Augmenting Multiparty Interaction: Archives, Context, Presence*  
Steve Renals, University of Edinburgh  
*HAPTIMAP: Multimodal Location Based Services*  
David McGookin, University of Glasgow

11:00-12:00  **Talks - (2) Design and Tools**  
*Design Support of Modalities for Mixed Interactive Systems*  
Guillaume Gauffre, IRIT, University of Toulouse  
*Tools to aid the capture of emotional responses to specifically designed activities in a location based game*  
Lynne Baillie, Glasgow Caledonian University  
*Visual Design of Multimodal Interaction - Bridging the Gap between Interaction Designers and Developers*  
Werner Konig, University of Konstanz  
*OpenWizard: A Wizard of Oz Component-Based Approach for Rapidly Prototyping and Testing Non-Fully Functional Multimodal Interfaces*  
Marcos Serrano, Universite Joseph Fourier Grenoble  
*Studying the Expression and Perception of Emotions for the Design of Embodied Multimodal Interfaces*  
Jean-Claude Martin, LIMSI-CNRS

12:00-13:00  **Lunch (Posters and Demos)**

13:00-14:00  **Talks - (3) Engineering**  
*Multimodal Components Integration*  
Diego Arnone, CALLAS Consortium, Engineering Ingegneria Informatica  
*Description Languages for Multimodal Interaction*  
Bruno Dumas, University of Fribourg, Switzerland  
*Towards a model-driven engineering approach for designing and developing multimodal services*  
Slim Ben Hassen, Orange Labs (Lannion - France)  
*Incorporating into OI platform : from simple interaction modalities to a complete multimodal interpreter*  
Kenia Sousa, UCL-BCHI

14:00-15:00  **Talks - (4) Software Technologies and Evaluation**  
*Framework for Interactive Services in Ambient Computing Environments*  
Andreas Lorenz, FIT, Sankt Augustin, Germany  
*Architecture of a Multimodal Fusion Framework for Interactive Applications*  
Hildeberto Mendonça, Université catholique de Louvain  
*Dynamic Design Using the Multimodal Browser*  
Nilo Menezes, Multitel, ASBL  
*The activation of modality in virtual objects assembly*  
Guillaume Riviere, ESTIA-Biarritz  
*Towards an extension of usability concepts to evaluate multimodal games,*  
Marc Mersiol, Orange Labs (Lannion - France)

15:00-16:00  **Challenges** (Discussion session)

16:00-17:00  **Report back and wrap up**

17:00-19:00  **Demo Session**
The Meeting Location

Fraunhofer Institute for Applied Information Technology FIT
Schloss Birlinghoven, 53754 Sankt Augustin, Germany

We suggest to stay in a Hotel in Bonn and take Bus 608 or 538 to the institute. Bus tickets can be purchased on the bus. Bus takes about 30 minutes, costs about 3 Euro. The stop is called: "Fraunhofer-Schloss Birlinghoven". It's the final stop of line 608 and an intermediate stop of line 538. In Bonn, Bus 608 has a direction sign "Hoholz" and Bus 538 reads "Hennef". Busses go every 20 minutes from Bonn central station and a station close to Kennedy Bridge named "Bertha-von-Suttner Platz".

http://www.fit.fraunhofer.de/profil/anfahrt_en.html
http://maps.google.de/maps/ms?ie=UTF8&hl=de&msa=0&msid=103262133797387383639.00046539598d2b2c858b6&ll=50.741344,7.158279&spn=0.091902,0.180416&z=13

Dinner – Wednesday at 20:00

20:00 - Zum Gequetschten

http://www.bredderbud.de/

Sternstr. 78, Bonn

http://maps.google.com/maps?hl=en&client=safari&ie=UTF8&cid=0.0.6804411323588602463&fb=1&splt=1&dq=zum+gequetschten+bonn&daddr=Sternstr.78,+53111,+Bonn,+Germany&geocode=14179695715267504147.50.735812.7.098412&ei=NOQESpDnHc7eDC-Ab7ramaAw&ll=50.73567.099872&spn=0.007972,0.016394&z=16
After the workshop

Papers will be made available as PDFs on the website – http://www.dcs.gla.ac.uk/~sjg/oi

You are invited to submit a long paper for a special issues of the Springer Journal on Multimodal User Interfaces.

Submissions should be 4 to 12 pages long and must be written in English.

Formatting instructions and templates are available on: http://www.jmui.org

Authors are encouraged to send to: mcgeemr@dcs.gla.ac.uk including a brief email indicating their intention to participate as soon as possible, including their contact information and the topic they intend to address in their submissions.
The Challenges of Engineering Multimodal Interaction:

DESIGN
Tools to aid the capture of emotional responses to specifically designed activities in a location based game

ABSTRACT

Location based games offer opportunities for us to learn more about peoples interactions and feelings towards the environment they are in as well as to understand more about the mental models and locations associated with known environments e.g. a university campus with its associations of learning. In our study we wanted to manipulate the activities in a game to take advantage of certain locations in the hope of producing certain emotional reactions. However, it is not enough to simply produce these reactions one must also have a way of capturing any emotions produced whether these are the ones expected or not. The objective of the work was to trial a new methodology for location based games that aims to capture the players emotional reactions to the activities in a game whilst in certain locations. In order to test the methodology we designed a location based game that can be played on any Bluetooth enabled mobile phone that has an accelerometer. The game has been designed to interweave with a persons’ normal activity, as a result there is little distinction between gaming time and non-gaming time.

1. INTRODUCTION

Pervasive and ubiquitous games are likely to become more important, because of the increasing pervasiveness, power and sensory capabilities of the main platform they need – mobile phones and similar devices. To date, most of the literature is devoted to technical feasibilities of various elements of pervasive and location-based games, with some publications also concerned with usability of the games. There is a recognition that established design considerations and methods for PC or console games are not fully adequate for newer pervasive, location-based games.

The nearly constant presence of the mobile device, that is carried and taken everywhere by its owner, means that games played with the device can take on a casual character, whereby the player can choose to suspend and resume the game at any time. Usually “casual games” (IGDA, 2006) include easy to learn games that are often played for relaxation, either as “snacks” or more intensively and obsessively. Here we refer particularly to the way that pervasive games can in principle be played at any time, and may be suspended to be resumed later at the same point, so that they may be integrated into or alternating with normal life activities. They are different from PC games, therefore, and can be expected to produce a different kind of player experience. Korhonen et al (2008) have found, for instance, that players may be curious about the game state if it continues while they are not playing it, and may log in several times a day merely to check out of curiosity.

Location-based games present technical challenges in order to activate locations with game-relevant information. This may be done with software and signals, such as satellite GPS data; or with fixed hardware, such as the Bluetooth stations we use in our game “Destination”. But location-based games also challenge the game designer, especially if they are also casual games. In that case the game could
be pervasive and possibly invasive in the player's life, becoming embedded into everyday activities. As such it may compete with work, study or social interaction. The game could be distracting to the player, or to other people in the same space. Activities in the game might cause embarrassment to the player, as other people in the area would not know that the player is playing a game at all. Current prototype pervasive and mobile games may be embarrassing in that they require players to wear an open laptop on the back, for example, or to wear other strange devices attached to their clothing, such as “pacman game name” (Magerkurth et al, 2005) or “Time warp” (Herbst, et al 2008). With more use of activity sensors, such as accelerometers in mobile devices, however, new modes of interaction are possible, including gesture-based interfaces. Destination has been designed to include various gestures that the user has to undertake from the repetitive to the more exciting e.g. fighting gestures with an invisible enemy at certain locations, some of which are more public than others, to discover whether public embarrassment and other emotions could be a significant factor in PUGs and other PU applications.

In order to investigate this question, a method is needed to measure the emotional reactions and play experiences of the participants in the study. Some kind of play testing is a normal part of the game design process in the games industry (Fullerton et al, 2004). It can be as simple as employing people to play prototype versions of the game and report any bugs or design flaws they find, in their opinion. But in HCI and in psychological research it is better practice to standardize participants' subjective reports with some structured questionnaire or similar instrument, preferably one that has been validated in earlier research in the literature.

In the relatively new field of affective computing (Picard, 1997), it is a research aim to enable machines to read a person's emotions through various sensors. This is difficult to do, although some limited progress has been made, also in application to PC video games (e.g. Mandryk et al, 2006; Moffat & K eigler, 2006). Ideally, we would like to apply these methods to automate and standardize the measurements of play experience to our PUG game. Such work is best done in a special purpose usability laboratory, however, with video cameras and other recording devices to observe people as they play the game. For obvious reasons, this is not feasible for PUGs because by their nature they take place in the real world, not in a laboratory.

The main purpose of our work is to respond to this last challenge, by taking existing methods for measurement of emotion and adapting them to the PUG context, to evaluate how well they apply there. The methods we are currently trying are comparatively cheap and easy to apply, and if found to be successful should help other researchers in the design and evaluation of PUGs and multimodal applications too and this is what we hope to discuss at the workshop.

References


Design Support of Modalities for Mixed Interactive Systems.

Guillaume Gauffre, Philippe Truillet, Emmanuel Dubois

gauffre@irit.fr, truillet@irit.fr, edubois@irit.fr

IRIT – University of Toulouse
FRANCE

1 Introduction

Mixed Interactive Systems (MIS) are interactive systems which merge physical and digital worlds to create rich interaction (Augmented Reality, Pervasive systems or Tangible User Interfaces). Their deployment is still in its infancy and mature tools addressing their development remain unavailable. Indeed, much works on MIS are based on ad-hoc development or dedicated SDK, following a bottom-up approach. We prefer a top-down approach which allows exploring the many facets of MIS and the complexity and richness of mixed interactive situations. With the perspective to address the different steps of their development, we define several models and the way to interlace them [3].

2 MIS development support

We focus in this paper on the design and implementation steps of our process to depict the design of interaction modalities. We describe such development steps with the creation of a system which enables to interact with the Google Earth software through the use of mixed input modalities. The system named “GE-SteeringBoard” [2], provides a board, representing a navigational compass, that the user holds and moves to modify its point of view on Earth. The three next sections present the three stages that our process involves to realize such a system.

2.1 Interaction modelling

The aim of this stage is to describe how the interaction takes place for a specific mixed interactive situation, presenting the user, the physical and digital resources and the interaction channels between them. The communications between entities are described considering the channel, the sensing mechanism – receiver – and the modification method – emitter. Furthermore, a channel is characterized with attributes such as the medium carrying the data, the language form to express the data, its dimension, etc.

2.2 Software Architecture modelling

This second stage is about the description of the software architecture as a component-based architecture. ASUR-IL is used to identify each component required, their role in the system and the communications between them.

2.3 Implementation

The third stage of this process consists in the implementation of the system, in our case using the WComp platform, a rapid prototyping platform based on the .NET framework and which has the advantages of being simple and efficient and offering low resource consumption. The WComp metamodel implies the description of an assembly, the components and their interfaces and the connectors between them. Once compiled, the assembly can be loaded with WComp in order to test the system.
2.4 Articulations
Each stage of the process is interlaced with the others thanks to Model-Driven Engineering tools and the model transformation ability. The aim is to consider the modeling results of each phase as a source for the next phase. Once the interaction described with ASUR, the model is transformed into a ASUR-IL model which initialize the design of the software architecture. Finally, an ASUR-IL model is transformed into a WComp assembly and component skeletons. Such a process allow us to rapidly prototype a MIS, by using each model as a support to each stage and as a source for further stages.

3 Multimodality and Mixed Interactive Systems
With the “GE-SteeringBoard” system, some gestures on the board may be difficult to realize, especially the tilt movement. The concurrent use of all the functionalities is also hard to control. We realize another prototype based on two complementary modalities: the motions of the board on 3 axes and a touch sensor to separate two modes: position/elevation and tilt/azimuth. When designing such a system, thinking about how integrating multimodality contributions to MIS design leads us to identify several concerns of MIS for applying multimodality. They are about three main purposes: the description of the interaction, the composition of a modality and the evaluation concerns.

3.1 Characterizing the interaction
MIS highlighted the use of physical resources in the interaction which makes the modalities more and more heterogeneous and plural. Multimodality must take into account this diversity, forget that a device is a dedicated electronic device and accept every perception sense and every action means as potential modalities. Consequently, a user may interact through a modality constituted of heterogeneous physical artefacts instead of a specific set of peripherals. Each physical object has a particular shape, weight, colour, lightning properties,... i.e. each one has specific physical properties, that affect the way to combine them. Furthermore, several objects can be involved for the same modality. To deal with such an evolution, the definition of a modality must be adapted.

3.2 Describing a modality
The works around multimodality gave one point of view on what is a modality, establishing the duet: device and language, but may consider several “devices”, or sensing mechanisms/modification of a specific set of peripherals. Each physical object has a particular shape, weight, colour, lightning properties,... i.e. each one has specific physical properties, that affect the way to combine them. Furthermore, several objects can be involved for the same modality. To deal with such an evolution, the definition of a modality must be adapted.

3.3 Evaluating a system
MIS have also the particularity to be involved in low-constraint spaces. Mobility and wide spaces can characterize some of the situations related to MIS. In order to evaluate such systems, the description of modalities may have to be completed by spatial and focus level considerations. Ambient spaces are such situations where modalities are used to spatialize information, their source and their importance. With such considerations added, expressing a multimodal relationship with for example the CARE properties, may require additional concepts to address the MIS specificities.

4 Conclusion
Several MIS concerns affect the implementation of multimodality in MIS context. They constitute anchor points for multimodality tools and methods, with a focus on the definition of a mixed modality. It cannot be described as a duet: device and language, but may consider several “devices”, or sensing mechanisms/modification methods, and different forms of language to consider each information path between physical and digital artefacts: < D x L > \rightarrow < \{D_{1},..,D_{n}\} x \{L_{1},..,L_{n}\} > . The representation of a concept has yet been decomposed when Ishii
et al. [4] introduced the MCRit for TUI and expressed the intangible and tangible representations. With such elements identified, the question is now to know if multimodal tools and methods have to be adapted or modified in order to consider mixed interactive situations. Is the CARE set complete to address the different kinds of modalities combination?

5 References
**Visual Design of Multimodal Interaction - Bridging the Gap between Interaction Designers and Developers**

Werner A. König, Roman Rädle, and Harald Reiterer  
HCI-Group, University of Konstanz  
Box-D73, Universitätsstraße 10, 78457 Konstanz  
{koenigw, raedler, reiterer}@inf.uni-konstanz.de

**ABSTRACT**

In contrast to the pioneers of multimodal interaction e.g. Richard Bolt in the late seventies, today’s researchers can benefit of a wide variety of existing interaction techniques, devices and frameworks. Although these tools are available, the usage of them is still a great challenge particularly in terms of usability. A major issue results from the trade-off between the functionality of the system and the simplicity of use. We introduce a novel visual user interface concept which is especially designed to ease the design and development of post-WIMP user interfaces including multimodal interaction. It provides an integrated design environment for our interaction library “Squidy” based on high-level visual data flow programming combined with zoomable user interface concepts. The user interface offers a simple visual language and a collection of ready-to-use devices, filters and interaction techniques. We specifically address the trade-off between functionality and simplicity by utilizing the concept of semantic zooming which enables dynamic access to more advanced functionality on demand. Thus, developers as well as interaction designers are able to adjust the complexity of the Squidy user interface to their current need and knowledge.

**VISUAL INTERACTION DESIGN WITH SQUIDY**

In the last years diverse frameworks evolved which offer great opportunities to implement and realize novel interaction techniques in the field of multimodal interaction and post-WIMP user interfaces [8]. These frameworks such as ICON Input Configurator [1], its successor MaggLite [2], the OpenInterface Framework [7], MAX/MSP [5] and vvvv [9] provide high flexibility and great functionality, but interaction designers with less or no programming experience are mostly overstrained by the complexity of these tools and their user interfaces. However interaction designers have to iteratively test, improve and evaluate their techniques especially when doing research.

We address this issue with our interaction library “Squidy” which unifies various device toolkits and frameworks in a common library and provides an integrated user interface for visual dataflow management as well as device and data filter configuration. Squidy thereby hides the complexity of the technical implementation from the user by providing a simple visual language and a collection of ready-to-use devices, filters and interaction techniques. This facilitates rapid prototyping and fast iterations during the process of design and development. However, if more functionality and profound customizations are required, the visual user interface reveals advanced information and operations on demand by using the concept of semantic zooming. Thus, users are able to adjust the complexity of the user interface to their current need and knowledge (ease of learning).

![Figure 1: View of a simple pipeline in Squidy.](image)

The basic concept which enables the visual definition of the dataflow between the input and output is based on a pipe-and-filter concept (Figure 1). This offers a very simple but powerful visual language to design the interaction logic. The user thereby selects the input device or hardware prototype of choice as “source”, connects it successively with filter nodes for data processing such as compensation of hand tremor or gesture recognition and routes the refined data to the “sink”. This can be an output device such as a vibrating motor for tactile stimulation or LEDs for visual feedback. Squidy also provides a mouse emulator as output node to offer the possibility to control standard WIMP-applications with unconventional input devices. Multipoint applications (e.g. for multi-touch surfaces or multi-user
environments) and remote control are supported by an output node which transmits the interaction data as TUIO messages over the network. TUIO is a widely used protocol for multipoint interaction based on OSC. The internal dataflow between the nodes in Squidy consists of a stream of single or multiple grouped data objects of well-defined data types based on the primitive virtual devices introduced by Wallace [10]. In contrast to the low-level approaches used in related work, such abstracting and routing of higher-level objects has the advantage that not every single variable has to be routed and completely understood by the user. The nodes transmit, change, delete data objects, or generate additional ones (e.g. if a gesture was recognized).

Semantic Zooming
According to the assumption that navigation in information spaces is best supported by tapping into our natural spatial and geographic ways of thinking [6] we use a zoomable user interface concept to navigate inside the Squidy visual user interface. When zooming into a node, additional information and corresponding functionalities appear, depending on the real estate available (semantic zooming). Thus, the user is able to gradually define the level of detail (complexity) according to the current need for information and functionality.

Interactive Configuration & Evaluation
In contrast to existing frameworks the user does not have to leave the visual interface and switch to additional applications and programming environments in order to get additional information, to change properties, or to generate, change or just access the source code of device drivers and filters. In Squidy, zooming into a node reveals all parameters and enables the user to interactively adjust the values at run-time (Figure 2a). The changes take place immediately and thus neither requires a restart nor a recompilation of the source code. This is especially beneficial for empirically testing different suitable parameters (e.g. adjusting the noise levels of a Kalman filter) because of the possibility to directly compare these settings without introducing any (e.g. temporal) side effects. This process of interactive configuration and evaluation is much needed during the design of multimodal user interfaces especially when using uncommon interaction techniques and user interfaces. Squidy therefore facilitates rapid development iterations.

Details on demand
Further beyond the access and manipulation of parameters, Squidy provides illustrated information about the functionality, usage and context of the node directly embedded in the node. By zooming into the information view marked by a white “i” on a blue background (Figure 2a) the information is shown without losing the context of the node. This information view (Figure 3a) may contain code documentation (e.g. automatically generated by javadoc), user-generated content (e.g. from online resources such as wikipedia.org or the Squidy-Wiki) or specifically assembled documentation such as a product specification consisting of textual descriptions, images and videos.

Embedded Code and on-the-fly compilation
The user has the ability to even access the source code (Figure 3b) of the node by semantic zooming. Thus, code changes can be made directly inside the design environment. Assistants such as syntax highlighting or code completion support the user even further. If the user zooms out, the code will be compiled and integrated on the fly, again without the need of a system restart. Users may even add new input and output devices or filters by adding an empty node and augmenting it with applicable code. In order to share the new node with the community the user can publish it into the knowledge base. The design rationale is not to replace the classical development environments such as Microsoft Visual Studio or Eclipse, but rather to integrate some of their functionality directly into Squidy. Thereby, we provide a unified but easy to use design environment which seamlessly integrates the most relevant tools and functionalities for the visual design and interactive development of post-WIMP user interfaces and multimodal interaction.

CONCLUSION
The trade-off between functionality and simplicity is at least as old as the idea of visual programming. In previous systems the agreement concerning this trade-off was made in advance of their usage. The design rational behind
Squidy and its visual user interface concept is the idea to give the decision to the current user. Thus, we address the trade-off by a dynamic, user-controlled approach. This concept also bridges the gap between different users and their knowledge and tasks. Interaction designers may choose and combine input devices and interaction techniques at a high abstraction level whereas developers may zoom in and modify technical details programmatically. Squidy therefore provides a unified design environment bridging the gap between diverse techniques and users. You may find some more information about Squidy in [3] and [4].

REFERENCES
Current human-computer interfaces are limited compared to the everyday multimodal communication we use when engaging with other people in normal day-to-day interactions or in emotionally-loaded interactions. In our human-human interactions, we combine several communication modalities such as speech and gestures in a spontaneous and subtle way.

Whereas current human-computer interfaces make an extensive use of graphical user interface combining keyboard, mouse and screen, multimodal human-computer interfaces aim at an intuitive interaction using several human-like modalities such as speech and gestures which require no training to use.

An Embodied Conversational Agent (ECA) is a multimodal output interface in which an animated character displayed on the screen combines several human-like modalities such as speech, gesture and facial expressions. Using an ECA is expected to lead to an intuitive and friendly interaction, for example via the display of emotional expressions that can be useful for applications in education, for example as a means of motivating students.

A bidirectional interface aims at combining multimodal input and ECA, thus involving intuitive modalities on both sides of the interaction. It raises several challenges for computer scientists. Designers of the multimodal input interface need to know how users will combine their speech and gestures when referring to objects so that they can define appropriate fusion algorithms. Designers of the ECA need to know how the character should combine its communication modalities so that the user gets the message right. This calls for appropriate input and output computational models of multimodal behaviors.

With respect to evaluation, the fact that users and the system can choose among several modalities could be an advantage over classical graphical user interfaces.

The literature in social sciences provides results of numerous studies on nonverbal communication. But, using these results to design multimodal interfaces is not straightforward: in order to be able to analyze or to generate multimodal behaviors we need detailed descriptions of observed behavior. Furthermore, the situation for which we want to design an interface might be quite different from the ones studied in the literature, and the applicability of their results to another situational context requires domain-specific investigations. Our goal is to contribute to the design of bidirectional multimodal interfaces. Our approach is threefold. First, we build representations of situated (e.g. collected in a specific context) multimodal communicative behaviors. Second, we consider the two sides of the multimodal human-computer interface (e.g. multimodal input from the user, and multimodal output via the ECA). Third, we evaluate such representations and interfaces via experimental studies involving different user profiles.

We define a methodological framework for the collection of videotaped multimodal behaviors, their annotation, their representation, their generation, including the perception of the collected and generated behaviors. This framework features a typology for analyzing combinations between modalities in multimodal behavior (complementarity, redundancy, conflict, equivalence, specialization, transfer).
We will summarize several studies achieved with this methodological framework. We focus on multimodal deictic and emotional behaviors. We selected these two behaviors because 1) they involve several modalities, 2) they are required both in input and output, and 3) they are especially useful in pedagogical applications.

With respect to deictic behaviors, we studied how adults and children combine their speech and gestures when they interact with 2D and 3D animated agents in conversational edutainment applications. We used a simulated version of the system, we analyzed the observed combinations of modalities produced by users, we designed an input fusion module, and we evaluated the fusion of users’ speech and deictic gestures in the final system. On the output side, we specified and compared the perception that different users have of different multimodal strategies and graphical look of animated agents when they refer to graphical objects during technical presentations [2].

With respect to multimodal expressions of emotion, we studied how people combine different modalities in spontaneous expressions of emotions. Most emotion studies in social sciences focus on acted mono-modal behaviors instructed by single labeled emotions, although some studies also observed blends of emotions. We used our approach to explore how people express and perceive such blends of emotions in non-acted multimodal behavior. We collected videos of multimodal emotional behaviors during TV interviews. We defined schemes for manually annotating these expressive emotional behaviors. We investigated the use of image processing techniques for validating these manual annotations and we computed representations of multimodal expressive behaviors. We defined a copy-synthesis approach using these representations for replaying the behaviors with an expressive agent. Perceptual studies were conducted for analyzing individual differences with respect to the perception of such blended emotional expressions [1]. The understanding of conflicting facial expressions of emotion and textual dialogues was also assessed with autistic users in an experimental study for which a multimedia platform was designed [3].

These studies provide three contributions to the design of multimodal human-computer interfaces. First, these studies revealed several individual differences between subjects with respect to age, gender, cognitive profile and personality traits with regards to the perception and production of deictic and emotional behaviors. Second, we iteratively defined a methodological framework for studying multimodal communication including the annotation, representation, generation, and perception of situated multimodal behaviors. Third, this framework was used to define representations such as combinations of modalities in user's multimodal input, and expressive emotional profiles for the annotation and replay of emotional behaviors.

Our future directions of research aim at providing further contributions to inform the design of bidirectional multimodal human-computer interfaces including facial and postural expressions of emotions, virtual characters and humanoid robots, and their applications for artistic and pedagogical applications. For this purpose, we will extend our experimental approach to large and multisource multimodal corpora [4].

References
OpenWizard: A Wizard of Oz Component-Based Approach for Rapidly Prototyping and Testing Non-Fully Functional Multimodal Interfaces

Marcos Serrano, Laurence Nigay
University of Grenoble, CNRS, LIG
BP 53, 38041, Grenoble cedex 9, France

ABSTRACT
OpenWizard allows the designer and the developer to rapidly test a non-fully functional multimodal prototype by replacing one modality or a composition of modalities that are not yet available by wizard of oz techniques. OpenWizard is based on our conceptual component-based approach for rapidly developing multimodal interfaces, an approach first implemented in the ICARE software tool and more recently in the OpenInterface tool. We present a set of wizard of oz (WoZ) components that are implemented in OpenInterface. While some wizard of oz (WoZ) components are generic to be reused for different multimodal applications, our approach allows the integration of tailored WoZ components. We illustrate OpenWizard using a multimodal map navigator.

General Terms
Algorithms, Design, Human Factors, Standardization, Theory.

Keywords

1. Introduction
Recent years have seen the advent of a variety of new interaction techniques. User interfaces have evolved from the classical screen-mouse-keyboard to the integration of several innovative interaction devices. Examples include tactile surfaces and augmented physical objects. Even if such techniques open a vast world of new interaction possibilities, they can be expensive and/or difficult to integrate. Not to mention the complexity due to the combination of such modalities into multimodal applications.

In this context, tools for prototyping and developing multimodal user interfaces are then needed. One such attempt is our conceptual component-based approach for rapidly developing multimodal interfaces, an approach first implemented in the ICARE graphical tool ([1]) and more recently in the OpenInterface framework ([3]).

Based on our experiences with OpenInterface, while the rapid development of multimodal prototypes is possible, testing such prototypes raised further technical and experimental issues.

To accelerate the cycle prototype-test at the early stage of the design, non-fully functional prototypes are the solution we adopted with OpenWizard. As pointed out in [2], wizard of oz (WoZ) studies have been shown to allow fast prototyping of interactive applications by allowing the evaluator (wizard) to simulate missing functions. We focus here on non-fully prototypes in terms of interaction modalities.

2. OpenWizard Approach

Underlying component-based approach

In ([3]) we presented a component-based approach for multimodal interaction. Our approach defined a characterization space of software components for multimodal interaction constructed along three dimensions. In the present work we use two of those three dimensions. The first dimension is related to the data-flow from input devices to an interactive task. Along this axis, we identify four types of components: Device, Transformation, Composition and Task. The second dimension describes the genericity of the components. Our space includes both generic and tailored components. Generic components represent high-level reusable abstractions. Tailored components implement operations for specific devices or for application-dependent tasks.

Puzzle metaphor: building multimodal interaction with WoZ components

Based on our component-based approach, a multimodal interaction prototype can be seen as a puzzle, where each piece (i.e. component) corresponds to a physical device, a transformation algorithm, a fusion algorithm or a task. The designer and the developer assemble those puzzle pieces in order to create the complete puzzle, their multimodal interface prototype.

In order to avoid losing time and effort in waiting for missing pieces and in creating and implementing pieces that will then be discarded during evaluations with end-users, designers and developers should be able to fill that puzzle with WoZ pieces in order to complete a non-fully functional prototype that could then be evaluated. Figure 1
illustrates the puzzle metaphor along with a component assembly that includes WoZ pieces (i.e. WoZ components).

In order to evaluate that interaction, we can use a “two-dimensional position” WoZ component, as shown in Figure 3. This WoZ component allows the wizard to click on an image in order to generate a position (x, y). This WoZ component enables us to replace a DiamondTouch table (i.e., an expensive device) or a vision-based finger tracking not fully functional.

4. Conclusion and Future Work
OpenWizard is a component-based approach for the rapid prototyping and testing of input multimodal interaction. Such an approach will enhance the creation of innovative multimodal applications by allowing a designer/developer to quickly build and test non-fully functional prototypes based on wizard of oz (WoZ) techniques for simulating some parts of the multimodal interaction. Moreover for defining the WoZ experimental setting, we can dispatch each WoZ component to a dedicated wizard (e.g., multi-wizard experiment) or we can consider only one wizard assigned to simulate different components. In this latter case the workstation of the single wizard includes an integrated user interface for simulating several components. Such integration of user interfaces as well as various tools including the visualization of inputs (Input console of Figure 2) and automatable tools for detecting temporal relationships for the case of fusion simulation are currently under study for assisting the human wizards in their tasks.

5. References
The Challenges of Engineering Multimodal Interaction:

ENGINEERING
An Open Source Integrated Framework for Rapid Prototyping of Multimodal Affective Applications in Digital Entertainment

Position Paper

Diego Arnone, Massimo Bertoncini
Engineering Ingegneria Informatica.
via San Martino della Battaglia n°56, Roma
diego.arnone@eng.it, massimo.bertoncini@eng.it

Abstract

The development of applications relying on multimodal interfaces is becoming more and more an emerging area of interest in the Digital Art and Entertainment Domain. This paper aims at proposing a new approach to the integration of multimodal modules that are able to gather the emotional states of the audience, to process them and to perform an emotional output. The proposed approach is being developed inside the FP6 EU co-funded CALLAS Project and finds its concretization in a Framework and a set of integrated multimodal components.

Keywords: Multimodal Interaction Framework, multimodal affective applications development.

1. Introduction

Multimodality is a widespread feature of the new emerging software applications in the Digital Art and Entertainment domain especially those ones that aim at collecting the emotional state of the interacting users and at performing an emotional reaction. Emotions and moods can be indeed conveyed through several yet simultaneous modalities like body movements, speech, hand and head gestures, facial expressions which all together are able to better describe the affective state of people interacting with digital media (like a technological artwork, or an artistic installation, an educational interactive application, a next generation TV programme).

So, in order to gather and to perform the non verbal aspects of the human-machine communication, like the ones related to paralinguistics or gesture expressivity, a wide set of components is needed and must be placed at disposal of the multimodal affective applications designers. In this context, the CALLAS Project proposes a library of input components to detect user affectiveness (mostly related to user interest and participation level), and a set of powerful output components, able to create sounds, speech, augmented reality that emotionally react to the spectators. The peculiarity of the CALLAS technologies is that for the first time, as from our knowledge, the multimodality is addressed at higher semantic level, rather than at earlier time, based on the signal processing theories (DSP), and by proposing complex algorithms for removing the ambiguity often expressed through different modalities.

In the domain of multimodal interaction, several frameworks have been proposed and are being currently developed: SmartKom [1], Embassi [2][3], Galaxy C. [4], JASPIS [5] are mainly focused on the development of Dialog Systems that implement a human-machine interaction model very different from the one considered in CALLAS, where a flow of emotions, more than a structured dialog, lead the participation of spectators at the artwork. Moreover, even if multimodal frameworks like the ones proposed by CHIL [6], ICARE [7], OpenInterface [8][9][10] represent key references for the CALLAS Project, they differ from the proposed Framework in many significant features:

- none of them manages the emotional states of the audience, whereas the CALLAS Framework must provide fusion and interpretation algorithms whose main goal is the emotion recognition and interpretation;
- the interconnection among components often follows the pipelining approach so that the flow of data is defined at design time, whereas a key functionality required by who designs
multimodal affective applications is the dynamic runtime configuration of the final artistic installation, that must evolve over time depending on the information circulating among components;
- software modules can be integrated only if they provide a well defined API, whereas the CALLAS Framework allows the integration of a wider range of components on condition that they are able to communicate through TCP/UDP connections. The components are often coded in different programming languages, run on different software platforms or are released as black boxes;
- applications developers are mainly programmers, while the CALLAS framework target includes also artists that are often not confident with technology.

On the basis of its peculiarities, the proposed Framework differs from the existing software solutions so that a proper architecture has been designed from scratch.

The remainder of the paper is organized as follows. Section 2 illustrates the overall architecture and is split in 3 subsections: 2.1 defines the main functional blocks of the overall system and explains the mechanism that rules the communication among components; 2.2 describes the way a component can be integrated and 2.3. depicts the Framework graphical interface. Conclusions are drawn in Section 3.

2. The CALLAS general architecture

A preliminary remark concerns the distinction between Components and Core. The former encapsulate all the functionalities related to the recognition (input), the processing (framework) and the synthesis (output) of emotions by interfacing the interacting users through a multimodal interaction, the latter mainly deals with the communication mechanism among the components. The general architecture is depicted in figure 1. Input and Output components belong to the CALLAS Shelf, whereas the CALLAS Framework consists of the Core (Shared Repository and Control), the framework components, and the Authoring Tool.

2.1. A blackboard-based solution

The multimodal affective applications developed in CALLAS involve the interacting user even from the emotional point of view. Input components collect data coming from devices pointed at the users, extract a proper set of features from the raw signals and separately translate the extracted features in emotions, often referring to different emotional models. By doing so, input components supply a shared repository that stores a stack of information units (called Semantic Items). In their turn and according to particular preconditions, framework components access the repository, read the Semantic Items they are interested in, process them and write the resulting items in the same stack. Three main classes of processing are foreseen:
- disambiguation of conflicting emotions and fusion (fusion components);
- translation from an emotional representation model to another or interpretation of the same emotion in different contexts (interpretation components);
- partial or total encapsulation of the application logics (intermediate components).
At the output side, other components are in charge of performing an affective reaction to the public on the basis of the emotional behaviour that the application designer decides to apply.

All the components that cooperate inside an application can write in the repository or read from it if and only if some specific preconditions are satisfied: preconditions can depend on the peculiarities of the component or be defined by the programmer/artist at design time by means of a proper graphical interface (see section 2.3).

The interaction and cooperation process described above does perfectly implement the Blackboard model [11] in which all the components are Knowledge Sources that access the shared repository, i.e. the Blackboard, according to a policy managed by a Control. The figure 1(a) also shows a Configuration Manager and an Interconnection Manager: the former is in charge of translating in a control strategy the conditions and constraints that the designer defines by means of the Authoring Tool, whereas the latter leads with the flow of data at the feature level by implementing a dynamic pipelining manager. The need of an interconnection manager emerges because some input and output components are also able to respectively produce and accept data at the feature level and so a bus of direct connections from the input to the output side needs to be dynamically managed.

The figure 1 also shows in (b) how the preconditions related to the peculiarities of components are managed: all the software modules can deal with well specified classes of semantic items, called topics. This means that the Blackboard is subdivided in distinct containers labeled with different topics, and that the access mechanism implements a publish and subscribe pattern. The example in the figure involves only eight components.

In short, the proposed approach is a data-driven solution and makes the CALLAS Framework to differ from the most of the current and past multimodal frameworks (based, for instance, on pipelining or on the hub and spoke paradigm). More, modelling the framework as a blackboard system allows developers to include new functionalities by encapsulating them inside a knowledge source: so, for instance, existing augmented reality toolkits or storytelling engines or anthropomorphic virtual agent etc can be wrapped and brought in.

2.2. Integration approaches

In CALLAS, depending on the adopted integration approach, a component can be:
- native, if it accesses the Blackboard by directly invoking the Framework API;
- compliant, when it implements a well defined integration (OSC-based) protocol;
- external, when a proper wrapper is needed to translate its proprietary communication protocol in the invocation of the proper method in the Framework API.

In details, together with the Framework, also three skeletons are provided, one per type of components: input, output and framework. These skeletons allow a developer to code whatever native component so that it can access the blackboard through a well defined set of API. The result is a high level of performance and the access to a larger set of functions.

On the contrary, when a component is released as black box, it can exchange data with other software modules through a CALLAS compliant communication protocol or through a proprietary but properly wrapped interaction channel, external to CALLAS.

2.3. An Authoring Tool to support designers

Artists and, in general, multimodal-emotional applications designers are often non-programmers and so not confident with technology, code, configuration files, socket connections and so on. In order to provide them with an easy-to-use graphical interface, a proper Authoring Tool (AT) is being developed that will allow designers to:
- configure semantic connections to the blackboard. The way a semantic component connects to the blackboard, depends on its activation (for writing modules) and on its own subscription to the Control (for reading modules);
- configure the raw data connections among components to draw the pipeline;
- manage the runtime reconfiguration of an application, i.e. describe how the aggregation of components change over time;
- manage the runtime reconfiguration of components, i.e. define which configuration messages

1 The term “dynamic” is used to underline that the pipelined connections can change over time and depending on the semantic data that approach the Blackboard.
must be sent to the component that are able to perform a runtime auto-reconfiguration.

Starting from all the abovementioned functionalities, a CALLAS application has been modelled as a finite state machine, where each state corresponds to an aggregation of components: among the defined states, transitions are triggered by a set of preconditions defined by the author and imply actions that can be implicit (i.e. deduced from the differences between final and initial state) and explicit, when mainly concern the runtime component reconfiguration and are manually defined.

So the AT foresees two main panels: a **state panel**, allowing the aggregation of available (integrated) components and the definition of both raw data connections and semantic subscriptions, and an **application panel**, allowing to define transitions among states. In order to support this model, two proper XML-based description languages has been defined to describe components (**CCDL**, CALLAS Component Description Language) and applications (**CADL**, CALLAS Application Description Language).

### 3. Conclusions

In the Digital Art and Entertainment Domain, multimodality is a key requirement of the application interface, especially in affective artworks. Although many multimodal frameworks have been proposed by the scientific literature in the last fifteen years, the solution developed in the CALLAS Project, due to its peculiarities, required a novel design. As result, the proposed multimodal Framework consists of:

- an Authoring Tool, allowing designer to manage the application lifetime;
- a set of Components that process the semantic information gathered by the input components and produce semantic information for the output components;
- a Core, that manages the communication among components by adopting the pipelining approach at the raw data level and proposing a blackboard-based solution at the semantic level.

### 4. References


Towards a Model Driven Engineering Approach for Designing and Developing Multimodal Services

Slim Ben Hassen\(^1\) and Mariano Belaunde\(^1\)

\(^1\) Orange Labs, 8 Avenue Pierre Marzin
22300 Lannion, France
{slim.benhassen, mariano.belaunde}@orange-ftgroup.com

Abstract. Designing and developing multimodal interactive services is one of the challenges aspects facing interaction technologies. In Human Computer Interaction, multimodality denotes the capacity of a User Interface (UI) to offer more than one modality to interact with a computing system. Multimodality raises many issues for both the design and development of UIs. In this paper, we discuss why Model Driven Engineering concepts can be used in this area.

Keywords: Human Computer Interaction, Model Driven Engineering, Multimodal interface

1 Introduction

Multimodal human computer interaction (HCI) lies at the crossroads of several research area including computer vision, psychology and many others [1]. Human beings perceive the external world using the five senses of touch, taste, smell, hearing and sight. In most situations, natural communication between human beings is multimodal, as it combines several modes and modalities.

As computers become integrated into everyday objects, effective natural human computer interaction becomes critical: in many applications, users need to be able to interact naturally with computers the way human human interaction takes place.

The need of models in human computer interaction has been recognized for long. What we call a Model is a simplified formal representation of an entity, system or phenomenon in the real world (abstraction), with a certain correspondence, for a certain purpose.

Models can be of very different natures according to what kind of concepts they represent: for example, there are mathematical models, structural models (database structure, architecture), dynamic models (for computer simulation), economical models, behaviour models, mental models and so on. In this paper, the focus is set on models that represent engineering concepts and data used or produced by engineering processes.
Model driven engineering (MDE) [2] refers to the systematic use of models as primary engineering artefacts throughout the engineering lifecycle. Full engineering processes are described by explicit networks of models connected by the mappings and transformations [3]. Such a model of a modelling language is referred to as a meta-model since it is in effect a model of all models expressed in that modelling language. Therefore, a meta-model of a language is a formal specification of a class of models expressed in that language. A model is said to conform to its meta-model like a program conforms to the grammar of the language it was written in.

In this paper, a result of an extensive literature overview, we discuss why it is promising to investigate model driven engineering for the design and development of multimodal interactive services, in particular when many variants of a same UI are to be produced for different platforms.

2 Model Driven Engineering of Multimodal UIs: Why?

A promising approach to address multimodal UI complexity is to develop model driven engineering technologies that combine the following advantages:

- Advantages in terms of communication: It improves communication between stakeholders (designers, developers, HCI experts, etc.) by providing a common vocabulary as well as graphical notations to facilitate common understanding. They can share a common reference and ambiguities are reduced.

- Advantages in terms of methodology: It is a widely accepted software engineering principle to start a software development cycle with a specification stage.

- Advantages in terms of handling and mastering complexity: This approach supports precise analysis of complex softwares. By adopting a model driven approach, a multimodal complex UI can be represented at different levels of detail and over multiple granularities.

- Advantages in terms of reusability and capitalization: In a multi-target context [4], MDE tools can provide automatic portability across the different targets. For example, developers can apply user-interface models at various abstraction levels to develop efficient tools for repurposing existing user interfaces to other platforms [5]. Capitalization of models can save the know how in HCI and enterprises. Knowledge is now recognized as a major and strategic subject for enterprises in an economical context of world-wide competition.

- To solve problems in a particular domain, MDE features include the development of Domain Specific Languages (DSLs) [6]. A Domain Specific Language is a specialized and problem oriented language. DSLs are typically small, highly focused languages, designed to be useful for modeling and solving clearly identified problems belonging to a specific domain. What separates DSLs from General Purpose Languages (GPLs) is the degree
of closeness to specific domain concepts. Their main goal is to increase productivity of domain engineers by targeting their specific domain needs. A DSL user concentrates her/his effort on domain description while complexity, design and implementation decisions are hidden. By working at a level away from implementation technology, multimodal HCI experts themselves will be able to understand, validate, modify and even develop DSL programs to model and solve specific HCI problems.

All of these advantages are the result of greater visibility within and beyond the scientific community. The last years have seen significant advances from the software engineering community, both conceptual and technological, in the field of model driven engineering. While the code-centric approaches might be suited for the development of simple UIs, they raise several problems for maintaining and testing multimodal UIs. The fundamental promise of model driven engineering is that the development, evolution and maintenance effort can be reduced by working at the model level instead of the code level.

3 A Review of Experiences from Applying MDE in UIs

The main goal of this section is to overview the experience gained by existing model driven approaches for developing advanced UIs.

We consider that the start point for investigating the actual status of model driven approaches for developing multimodal UIs is the UsiXML [7] (USer Interface eXtensible Markup Languages) initiative.

UsiXML is more than a XML-compliant language that describes the UI for multiple context of use such as graphical UIs and multimodal UIs, it also considers models to support model-based UI development. Task and domain models are first from requirements. One or many abstract user interfaces are produced by model-to-model transformation [8]. Subsequently, one or many concrete user interfaces are produced by model-to-model transformation in order to cover multiple interaction modalities (vocal, tactile, multimodal interfaces, etc.). Finally, each concrete UI then gives rise to a final user interface for a target computing platform.

A second interesting work is the UIML [9] (User Interface Markup Language) language. UIML is a declarative XML-based language that can be used to define user interfaces. It allows a multi-platform description of UIs and to generate them for different platforms. It provides a multi-step transformation-based framework containing the Logical Model, Physical Model and the Platform-Specific UI. The logical model has to be transformed to the physical model and the physical model has to be transformed to the Platform-specific UI. This framework utilizes concepts form the model-driven engineering literature and applies them to the area of multi-platform UI development.
4 Conclusions

In this paper, we identified and discussed motivations for applying MDE to design and develop multimodal services.

MDE is assumed to lead to higher productivity (by increased automation in the development process) and increased standardization and formalism. Developers can apply user-interfaces models at various abstraction levels and using them to consult multimodal HCI experts or to work with analysis to evaluate general decisions.

In the near future, we will be trying to articulate research/development efforts around an MDE approach to create a generic framework for modeling multimodal services. This will imply looking at the best ways for modeling different multimodal interaction patterns including alternative input ("say a word or click the button") - or synergic (touch a location in the screen saying "put the result here") as well as their implication in terms of the generated service logic code. In order to increase productivity and multimodal UI quality we are thinking on extending existing DSLs for service development [10].

References

Description Languages for Multimodal Interaction: What for, for Whom and How?

Dumas, B., Ingold, R., Lalanne, D.

University of Fribourg, Switzerland

Tools and frameworks such as OpenInterface, target at easing the creation of multimodal interfaces. The development of these tools open a broad question: how to best model and represent multimodal human-machine interaction? This position paper focuses on one of the possible ways to address this problem: description languages, and some of their features. In particular, this position paper seeks to open discussion on three specific questions: who would be the users of description languages for multimodal interaction? What would be these languages used for? How should such languages be able to describe best multimodal interaction and its distinctive features?

<table>
<thead>
<tr>
<th>Language</th>
<th>Dialog level</th>
<th>Events level</th>
<th>I/O level</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMMA (W3C Multimodal Interaction activity)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>XISL (Katsurada et al.)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ICO (Ladry et al.)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>MultimodaliXML (Stanciulescu et al.)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>TeresaXML (Paternò et al.)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>MIML (Araki et al.)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>NiMMiT (De Boeck et al.)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. State of the art languages and guidelines.

Languages For Multimodal Toolkits and Their Focus

The World Wide Web Consortium (W3C) Multimodal Interaction Activity proposes a theoretical framework describing major components involved in multimodal interaction, as well as potential or existent markup languages used to relate those different components, such as the W3C EMMA markup language, or modality-focused languages such as VoiceXML or InkML. The works of the W3C inspired Katsurada et al. for their work on the XISL XML language. XISL focuses on the synchronization of multimodal input and output, as well as dialog flow and transition. Another approach of the problem is the one of Araki et al., who propose MIML (Multimodal Interaction Markup Language). One of the key characteristics of this language is its three-layered description of interaction, focusing on interaction, tasks and platform. Ladry et al. proposed the ICO notation for the description of multimodal interaction. This approach is closely related to a visual tool allowing edition and simulation of interactive systems, while being able to monitor at a low level a systems operation. Stanciulescu et al. followed a transformational approach for developing multimodal web user interfaces based on MultimodaliXML, a set of tools that all use the UsiXML language. Four steps are achieved to go from a generic model to the final user interface. This transformational approach is also used in Teresa XML (see Paterno et al.). Finally, at a higher level of modeling, NiMMiT (see De Boeck et al.) is a graphical notation associated to a language.
used for expressing and evaluating multimodal user interaction. Table 1 above summarizes at which levels each of these languages act. Readers will find all these references in (Dumas et al. [2009]).

Table 1 above summarizes at which levels each of these languages act. Readers will find all these references in (Dumas et al. [2009]).

SMUIML

SMUIML is our solution to cover the three levels presented in Table 1 and stands for Synchronized Multimodal User Interaction Modeling Language. As its name implies, the language seeks to offer developers a language for describing multimodal interaction, expressing in an easy-to-read and expressive way the modalities used, the recognizers attached to a given modality, the user-machine dialog modeling, the various events associated to this dialog, and the way those different events can be temporally synchronized.

Description languages can be used as configuration scripts for tools allowing creation of multimodal interfaces. The SMUIML language is able to configure HephaisTK, a toolkit for creation of multimodal interfaces, designed to plug itself in a client application that wishes to receive notifications of multimodal events received from a set of modality recognizers.

References

TOWARDS AN EXTENSION OF USABILITY CONCEPTS TO EVALUATE MULTIMODAL GAMES

Marc Mersiol
France Telecom

ABSTRACT
This paper puts forward some thoughts about the ergonomic evaluation of entertainment technologies. The concepts of usability to assess man-machine interfaces are extended to evaluate the specificities of games using multimodal interactions.

Keywords: Ergonomics, satisfaction, affective evaluation, game, multimodality.

1 INTRODUCTION
Multimodal interaction opens new users’ experiences especially in the domain of entertainment. For example, what emotional feeling can entertainment technologies activate when used by a user in a given context? To answer this question, new methods should be proposed to estimate the emotional dimensions inferred by the interaction of the users with entertainment applications and games (Chateau and Mersiol, 2005). In ergonomics, user tests traditionally focus on the study of the usability of the interfaces. Usability studies measure how a product can be used by given users to reach specific objectives with effectiveness, efficiency and satisfaction in a particular context of use (ISO 9241-11). Usability includes the dimension of satisfaction which is often estimated by a questionnaire and/or a Lickert scale. To complement the usability evaluation of a product and/or a service including multimodal interaction, the concept of satisfaction should be widened. The usability property of a product is associated with pleasure. Difficulties of use can generate feelings as boredom, frustration or stress (Jordan, 2000). An “emotional evaluation” should be conducted. Issues relative to pleasure, affects and feelings aroused by an interaction should be raised (Jordan, 2000 ; Norman, 2004) in particular in the case of multimodal interaction. To estimate playful experiences implies to consider not only the usability and the playability of the games but also the feelings felt by the players (Mandryk et al., 2006). The concepts of pleasure, affect and emotion would deserve to be considered for the design of new multimodal interfaces and for their assessment. These concepts seem to be particularly suited to study playful applications. The “funology” term is defined as the evolution of the usability standards towards a set of wider problem concerning the "fun", the pleasure, the aesthetics and the usage experiences (Blythe et al., 2003). Extensions of the concepts used in usability studies are proposed in this position paper to adapt them to the assessment of multimodal entertainment applications and games.
2 PROPOSITIONS OF AN EXTENSION OF USABILITY CONCEPTS TO EVALUATE MULTIMODAL ENTERTAINMENT APPLICATIONS AND GAMES

The criteria used to evaluate interfaces should be reconsidered in the scope of the evaluation of multimodal entertainment applications and games. The criteria of utility could be extended by the one of motivation. The criteria of usability could be extended by the one of playability. And the criteria of satisfaction could be extended by the one called “fun factor”.

2.1 From utility to motivation

The concept of utility appropriate for the definition of needs in ergonomics can evolve towards that of motivation in the field of games. The motivations of the players are determining elements for the success of a game (Joyner and TerKeurst, 2003). Motivations are generated by challenges. They refer to all the factors of the game connected to the difficulty and to the completion or to the realization of the objectives (Pagulayan, Keefer, Wixon, Romero & Fuller, 2003). Some motivations are intrinsic (challenge, control, curiosity, imagination), others are in the scope of the need (auto-fulfilment, esteems, feeling of membership, safety, physiology) and finally, some motivations are interpersonal (competition, gratefulness, cooperation). The study of these motivations and their relations are necessary to estimate their adequacies with the wishes of the players. A "good" game should have relevant challenges. Those generating frustration and difficulties of usability must be eliminated (Pagulayan et al., 2003). The levels of difficulty and reward (positive reinforcement) have to vary not to dull the motivation of the players. The level of difficulty has to increase to maintain a minimum and increasing interest. Besides the fact of being able to get over at the superior level, the motivation of the players can be maintained by activities of collections of objects, items, bonus... (Pagulayan et al., 2003). The levels of difficulty must be adapted to players' various profiles (Pagulayan et al., 2003). The game balance consists in finding the appropriate proportioning between objectives (challenges, obstacles of the game and the controls that the player can have on them). The player must be able to feel some pleasure by managing to reach his/her objectives and by avoiding feeling some frustration. On the contrary, he/she also has to feel the adversity, a confrontation; otherwise a feeling of boredom can arise. To learn how to use new multimodal interfaces can be a thrilling goal. The impact of using multimodal interactions should be evaluated according to the interest they add to the game. Finally, a game has to convey a system of values which must be studied, which is faiths with regard to what must be made in a context by basing itself on a comparison of the known options (Barr, Noble & Biddle, 2007).

2.2 From usability to playability

From an ergonomist point of view, an interface must be easy to understand and to use. The number of actions necessary for the realization of a task must be reduced. But this expected simplicity for an interface is not necessarily desirable for a game. Simplicity and game are not compatible (Hassenzahl, Platz, Burmester & Lehner, 2000). Certainly the game must be easy to learn by way of successive and progressive levels of difficulties. The controls which the player can have on the game have to be in adequacy with its perceptive and driving capacities (Pagulayan et al., 2003). User-friendliness must be possible and the physiological and driving answers should be without danger even if there are possible risks (recently, a player broke himself the foot by playing football in his lounge in a virtual environment and the term of "Wii-
elbow\(^1\) made its appearance). To understand the success of the system of control or of the interface of a game requires to understand how they affect its playability (Barr, Noble & Biddle, 2007). Playability\(^2\) is the degree to which a game is fun to play and usable, with an emphasis on the interaction style and plot-quality of the game; the quality of gameplay. The term of playability appoints all the rules and the possibilities governing the control of the player over his character (either machine or entity) via input devices (joystick, keyboard or mouse) in the video games. Playability considered as "good" evokes an interface and intuitive controls of game, a certain ease of handling (or at least a possible apprenticeship with a margin of progress) and a sufficient speed of display of the answers to the commands made by the player. Finally, the interface should be personalized (for example, allocation of specific actions in particular buttons of the interface and/or using specific modalities). How multimodal interfaces can give opportunities to achieve these goals should be evaluated. Heuristics are proposed to estimate the playability (Desurvire, Caplan & Toth, 2004). The proposed criteria concern 1) the playful experience in terms of fatigue, coherence, clarity of the objectives, 2) the story of the game in terms of understanding, interest, 3) the mechanisms of the game as answers suited to the actions of the players, the effects, the intuitive controls… and 4) the usability of the game: immediate returns on the actions of the users, the coherence, the menu integrated into the game… As games can be implemented on mobile devices, the playability will also have to be estimated in mobility. For that purpose, new heuristics are proposed (Korhonen and Koivisto, 2006).

2.3 From satisfaction to the “fun factor”

"The satisfaction is a feeling of well-being; pleasure which results from the fulfilment of what we wait, wishes, or simply of a desirable matter" (Le Petit Robert, in French). The satisfaction would not only be estimated by a cognitive point of view (judgment) but also an emotional one. To estimate playful experiences implies to consider not only the usability and the playability of the games but also the feelings felt by the players (Mandryk, Atkins & Inkpen, 2006). Without emotion, there is no game. An "emotional evaluation" of video games should thus be led. The story and the interactivity of the game are important. The scenario (or the story of the game) defines the role of the player. The created context and possibly the relation with the other players participate in its interactivity. This one aims at allowing the player to think of controlling the course of the game (presence, flooding) even if the computer can realize actions without possible intervention of the player. This last one must know which actions are possible and which are their possible consequences. Multimodal interaction can contribute to improve this “fun factor” by giving specific and new ways to interact. Some modalities can be less efficient than others to achieve a goal but very funny to use. The impact of multimodal interactions on the “fun factor” has also to be evaluated. The assessment of global quality of a video game declines by the factors "fun" and felt pleasure for a player. This pleasure is possibly generated by the quality of the images (for example, in terms of realism), animations, and sounds but also by original multimodal interactions. The feeling to improve his/her skills during the successive games is also important: the player must be gratified (positive reinforcement). This gratification is to a large extent connected to the possibility of being able to progress and to get over from a lower level of difficulty to a superior one. That is why the process of apprenticeship is a key element of a video game. A successful game is based on levels of difficulties allowing the player to learn how to overcome the encountered difficulties. In that case, the player strengthens his ego. For example,

---

1 In reference of « tennis-elbow », elbow inflammation of high level tennis player
players can learn how to use new multimodal interactions. On the contrary, if the game is too difficult, the player can lose confidence (Natkin, 2004). Other factors may contribute to the success of a game. In a video game, the player is active, his/her experience can vary in time: boredom and frustration in a session; excitement, enrichment and rewards in another one (Norman, 2004). But in the end, the experience of game must be positive. If the pleasure of the player must be the higher possible, its attention should also be maintained at a high level (Pagulayan et al., 2003). An immersion in the game can contribute to it. The player is taken in a flow which is a state of concentration, intense pleasure and total absorption in an activity (Johnson and Wiles, 2001). In this state, the player is taken in the game, the world disappears, the time stops. The player is only conscious of his playful activity. He/she feels boss of his/her actions and feels an intense enjoyment (Csikszentmihalyi, 1990). To reach this objective, the rhythm of the game can be raised to get the attention of the player and generate a tension, a stress, or even rushes of adrenaline. The intensity of the game allows focusing the attention of the player. The quantity of the information to be treated is not necessarily limited. The criterion of conciseness used in ergonomics is not necessarily respected. Therefore, the workload can be raised… Cognitive loads excess can arise during a game. Some high workload can put difficulties even to expert players. Strategies are necessary to exceed them. But these cognitive excess loads are also desirable to make the game fascinating (Ang, Zaphiris & Mahmood, 2007). Multimodal interactions can also improve this feeling of being overwhelmed in a game (for example, when using gesture and tactile modalities) and will have to be evaluated.

3 CONCLUSION AND FURTHER WORK

To conclude, this article presents first elements to which the ergonomists can bring answers within the framework of the evaluation of multimodal video games and entertainment applications. The position of the ergonomist in the evaluation of such applications remains to be exactly defined. This research theme is emergent in ergonomy. The research perspectives are numerous. The domain of the games asks theoretical and methodological issues for which the answers are not at present cut. The work of the ergonomists can, on one hand, contribute to the release of the knowledge connected to playful activities. For example, what are the impacts of a sustained playful activity on the behaviour of the players? In what measure a game can contribute to apprenticeships? How to design a multimodal game likely to generate optimal experiences ("state of flow")? In evaluation, a role of the ergonomists can be to define and implement new methodologies. To begin, they can consist in continuing the review of the relevant criteria in the field of video games and to confront them to those proposed in ergonomics for the evaluation of entertainment technologies. These criteria could then be established in a classification and new criteria proposed. They could be assessed during the evaluation of multimodal entertainment application and games (assessments, user tests…). In situ usage of these applications can also be envisaged to evaluate multimodal interactions over long periods.
REFERENCES


Incorporating into OI platform: from simple interaction modalities to a complete multimodal interpreter

Jean Vanderdonckt

UCL-BCHI

Abstract

One of the main goals of the OpenInterface platform consists of incorporating new interaction modalities into a plug-and-play component that can be then be offered and exploited by any interactive multimodal application. For this purpose, we developed two interaction modalities that became OI components: hand gesture recognition by off-the-shelf camera and sketch and gesture recognition on graphic tablet. Not only those two components have been integrated into OI but also they have been exploited in InterpiXML, a Java interpreter that interprets a multimodal user interface specified in UsiXML so as to render it dynamically. In this way, it is possible to enter a same piece of information either by traditional modalities such as mouse and keyboard but also by the two developed components in a way that is transparent to the end user. This resulting interpreter has been developed in a way that it itself becomes an OpenInterface component that can be also used as a simple multimodal interface for simple input/output. An empirical experiment has been conducted by both end users and developers that suggest that this new component is useful and that its OI implementation is superior to its previous one as estimated by developers.
The Challenges of Engineering Multimodal Interaction:

INTERACTION TECHNIQUES
ABSTRACT
In this paper we outline work currently being undertaken as part of the EU FP7 Haptimap project. Haptimap is focused on investigating how multimodal feedback and interaction techniques can be applied to improve navigational services for both sighted and visually impaired users. We outline two aspects of our current work: Touchmap, provides a virtual tactile map to allow visually impaired users to access map based information in a convenient and dynamic way. WedgeTouch attempts to overcome issues in relating visual 2D maps to the built environment, by providing an effective way for a user to non-visually “scan” the environment around them, to find points of interest such as shops, museums and cash machines.

INTRODUCTION
The EU FP7 Haptimap project is investigating how multimodal interaction can be incorporated into the design and development of mobile maps and location based services. The growth of such services has been rapid in recent years, with devices such as the Apple Inc. iPhone, bringing GPS and “always on” data connectivity to mainstream mobile devices. However, the interfaces that these devices use to accept input from, and output to, the user often rely on a solely visual modality, failing to consider the wide range of user activities and scenarios where location based services will be accessed. Haptimap aims to investigate how multimodal feedback can be used to best provide information when the user is engaged in “eyes busy” scenarios, such as running, cycling, or when the user has an impairment to his or her ability to interact with visual based location services. In this paper we outline two applications that are being developed at Glasgow University as part of the project. Both primarily address pedestrian navigation and awareness of services in the built environment.

TOUCHMAP
Presentation of geographical information presents significant challenges to people with a visual impairment. Visually impaired (VI) users often learn about geographical information by understanding routes: sequences of directions that can be learned and followed to allow navigation between two points [3]. Whilst these routes can be learned and navigated successfully, route learning does require explicit training that is often unavailable for all but the more commonly used routes. For example, a VI person may be taught how to navigate from his or her home to the local supermarket, but will not be trained on how to navigate from a hotel when on a trip. Several researchers have investigated how to provide navigational directions to a visually impaired user, using a combination of GPS and speech feedback [5]. Whilst these can be successful, they fail to provide both an overview of an area, as well as supporting the formation of a cognitive map, which allows different routes to be mentally integrated and a broader knowledge of an area to be obtained. A common way in which this information is provided to a sighted user is through the use of maps [2]. Whilst tactile maps are available to visually impaired users, they are paper based and tend to be cumbersome. Most are typically printed on A3 sheets of paper, and must be specially prepared so that information is presented large enough to be felt by the VI user. Therefore, an area may need to be represented on multiple sheets of paper. This makes it impractical for a user to take a map on holiday, or “out and about” on a trip. To investigate how maps may be better presented to a VI user, we are developing the TouchMap system.

TouchMap runs as an application on a Nokia N810 internet tablet (see Figure 1). A tactile map is drawn on the screen of the device using high contrast colours. As the user moves his or her finger across the screen, the system checks if the currently touched pixel is part of a map feature. If so, the user feels a vibration from an attached C2 Tactor (see Figure 1). In this way, the user can spatially integrate the information from the touchscreen, and feel around the edges of a building or road. To make this task easier, the vibration intensity is varied. At the edges of a feature, the intensity is presented at one half of the total intensity, and rises to full intensity when the user moves from the edge to the central part of the feature. In addition, we categorise several different kinds of map feature (such as roads, buildings and rivers), each of these is presented using a different waveform of vibration, with the standard waveform of the tactor (250Hz sine), being amplitude modulated to provide distinct discriminable
et al. developed the visual Wedge technique for off-screen visualisation as an alternative to incorporate multimodal feedback into their location-based services. In this paper, we have outlined two aspects of our work as part of the EU HaptiMap project. Both of these systems are currently being evaluated in lab-based scenarios to determine if users can extract useful information. However, after these primary evaluations, our intention is to carry out more long-term and in-depth evaluations, to identify how the applications fit around current user practice, and what they afford to the user in terms of interaction possibilities. However, to do this, both of our existing prototypes will need to be substantially modified to incorporate GPS positioning and more “real-world” evaluations.

### WEDGETOUCH

Another focus of our work is how to make visual maps more useful through the incorporation of multimodal feedback. A common map problem is when the user wishes to see a map containing both the current and destination points, but cannot do so with a sufficient map scale to provide enough information to show the correct walking direction. In addition, much of the information a user may require in an urban or unfamiliar environment may be unobvious (e.g., a coffee shop in the next street). To overcome these problems, several researchers have proposed visualisation techniques to bring currently off-screen data on-screen [4]. However, these techniques may still have issues when the user is trying to relate a position on the map to his or her current position in the environment [1], and still require the user to look at the device for all interactions.

To overcome some of these issues we are developing the WedgeTouch system. WedgeTouch, is an augmentation of the visual Wedge technique for off-screen visualisation as developed by Gustafson et al. [4]. Visually, wedges represent an off-screen point on-screen, by extending a triangle anchored to the off-screen point onto the visual display, the user can therefore use the size of the extension of the triangle to determine the direction and distance of an off-screen location (see Figure 2 left). We extend this concept by applying a spatialised auditory space that the user can interact with non-visually using the touchscreen of the device. Usually, the device shows a standard map with wedges showing off-screen points. When the user wishes an overview of the off-screen points, he/she taps the touchscreen of the iPhone. On the tapped point, a pie menu is rendered (see Figure 2 right). As the user moves his or her finger around the touchscreen and enters a segment of the pie menu, all of the off-screen features that lie in the extension of the pie menu segment are concurrently presented. Each off-screen point is represented by a single musical note, the timbre of which represents the type of the off-screen feature. The sounds are spatialised on a 2D plane, collocated with the user’s ears, to aid in the discrimination, and a 300ms onset-to-onset gap is applied between notes in the same segment of the pie menu. This allows the user to gain a quick overview of the features and facilities around them in the environment. Currently, we are comparing our WedgeTouch interface to that of the visual wedge system. Future developments in this area will include the incorporation of GPS positioning and more “real-world” evaluations.

### DISCUSSION

In this paper we have outlined two aspects of our work as part of the EU HaptiMap project. Both of these systems are currently being evaluated in lab-based scenarios to determine if users can extract useful information. However, after these primary evaluations, our intention is to carry out more long-term and in-depth evaluations, to identify how the applications fit around current user practice, and what they afford to the user in terms of interaction possibilities. However, to do this, both of our existing prototypes will need to be substantially modified to incorporate GPS positioning and more “real-world” evaluations.

### ACKNOWLEDGMENTS

This work is supported by EU FP7 HaptiMap (ICT-224675).

### REFERENCES


Augmenting Multiparty Interaction: Archives, Context, Presence

Steve Renals
University of Edinburgh
26 April 2009

1 Introduction

In the AMI and AMIDA projects (http://www.amiproject.org), we have been concerned with supporting human interaction and collaboration in both face-to-face and remote communication situations. In particular, we have focused on the development of algorithms, models and prototype systems that support interaction in meetings and access to meeting-related information. Much of this work has been based on the use of instrumented spaces, from which we capture multimodal recordings of the “communication scene”.

Interpreting the interaction between a group of people (whether co-located or remote, synchronous or asynchronous) is a major research challenge. To address this challenge requires a diversity of disciplines ranging from signal processing (“who spoke when”, tracking individuals and objects) and computer vision to speech and language processing (speech recognition, discourse processing) and models of group dynamics. Within AMI and AMIDA, we can factor the technology we have developed to understand interaction along three dimensions:

Archives: Enabling efficient access (indexing, search, browsing) to multimodal meeting recordings, captured from instrumented spaces, shared workspaces and videoconferencing

Context: Automatically finding and incorporating material related to an ongoing meeting or interaction, based on the current content of that interaction.

Presence: The realtime communication of state, to enable remote interactions, in particular to become richer.

2 Archives

Environments instrumented with multiple cameras and microphones, along with capture of whiteboards, handwriting, and data capture, provide a rich multisensory record of a communication scene. But to make any sense of such multimodal archives it is necessary to add semantic information to the raw signals, using audio-video recognition technologies such as:

- Unconstrained speech recognition;
- Object localization and tracking (e.g. head, hand tracking);
- Identification of the focus of attention;
- Person identification.

Each of these poses a challenging problem, requiring the state of the art to be extended, owing to the facts that the environments are typically not calibrated, people display natural and spontaneous behaviours, and the people themselves are not instrumented. In AMI and AMIDA we have addressed these problems in the constrained domain of business meetings.

The recognisers discussed above add significant value to multimodal recordings (the words spoken, the identities of the people involved, the actions they made) and it is indeed possible to use this information to index, search and browse recorded archives of such interactions.

State-of-the-art: With varying error rates it is possible to recognize conversational speech from room-based microphone arrays, track the focus of attention and identify the people in an interaction. And there are several international evaluations (e.g. NIST RT, CLEAR). Several demonstration systems (e.g. meeting browsers) have been constructed and are being trialled with commercial users.
Challenges: (1) Generalisation: current systems are typically well-adapted to a particular domain (e.g. busi-
ness meetings), but we need technologies that work well in a variety of settings, without need for extensive
calibration; (2) Scaling: can we develop systems which work in larger instrumented spaces in which a many
more people are meeting. Some specific smaller challenges might be to detect and track how people join and
leave subgroups etc. (3) Social signals: much of the important content in human interaction is not given simply
by the meanings of the words but is concerned with patterns of influence, agreement/disagreement, subjective
viewpoints, etc. We have started to systematically address some of these issues in a new Network of Excellence
on Social Signal Processing (http://www.sspnet.eu).

3 Context

State-of-the-art: To track context in realtime, recognizers must run in realtime. This has been achieved, with
relatively little degradation, for speech recognition, it is harder for computer vision. The AMIDA Content
Linking system is an example system

Challenges: One of the biggest problems is presenting contextual information in a way that does not result in a
cognitive overload. Finding an interface that people can use without disrupting their participation in a meeting
is a major issue. One interesting direction is developing virtual worlds which do not just mimic physical spaces
but are also information and contextual spaces, thus enabling realtime context to be part of the fabric of the
virtual world.

4 Presence

The notion of presence has become ubiquitous in the past few years, but it often has a very shallow meaning,
such as the communication state (and preferred modality) of a messaging client. We keep the notion of realtime
communication of state, but have focussed on a much richer User engagement and floor control

State-of-the-art: Transmitting information about what co-located people are attending

Challenges: How can social signals (interest/boredom, wanting to take the floor) be transmitted to a remote
participant in a meeting?

5 Outlook

Converging trends

- Cloud computing and storage
- Computational social signal processing
- Virtual worlds
- Low cost, high quality sensors

Challenges for

- Signal processing
- Machine learning
- Human-computer interfaces
- Social scientists
The Challenges of Engineering Multimodal Interaction:

SOFTWARE EVALUATION
Framework for Interactive Services in Ambient Computing Environments

Andreas Lorenz
andreas.lorenz@fit.fraunhofer.de

For decades, the setting of the user operating mouse and keyboard in front of a Personal Computer (PC) was standard. With the advent of mobile technology, this setting was enhanced by three aspects: users became location-independent, users interact with remote services embedded in the environment, interaction is missing mouse and keyboard hardware. While desktop computing is accompanied by ubiquitous computing, we stand at the border of the transition from local interaction with a computer system to location-independent remote interaction with systems available in the environment of the user. Powerful mobile computer and high speed wireless networking enable for enhanced interaction with services.

In a world of ambient computing environments, the technology disappears into the surroundings until only the user interface remains perceivable by users. "The user interface is that part of a system that the user comes in contact with physically, perceptually, or conceptually." [4] Most preferably, this interface is separated from the device hosting the services, e.g. the user interface is running on a portable device connected with the devices in the environment. The user interface depends on the context of the user, in particular it should not have a defined physical shape or prescribed interaction metaphor. The selection of input devices will be performed in an ad-hoc manner, either of using own devices the user carries with him, devices found in the environment, or the user implicitly interacts with the computer systems using (parts of) his body or just concluded from the user’s behavior. This work investigates to allow that the user can at his option use any physical device to generate semantically and syntactically correct input. The applied approach enhances the idea of separating the user interface from the application logic, leading to the definition of virtual or logical input devices physically separated from the controlled services.

Olsen pointed to the need of system models providing the fundamentals for supporting developers in building interactive applications [5]: "Lots of good research into input techniques will never be deployed until better system models are created to unify these techniques for application developers." The main objective of this work is to elaborate a generic solution enabling input devices to express user input to services of ambient computing environments. The key concept used in this work is separation of the user interface on the user input
device from the internal logic of the service of ambient computing environments. The central concern of separable user interfaces is the assumption that “it is possible to identify components that perform functions that should not be seen as part of the interface.” [1]

Developers of interactive systems must create a logical separation between application and user interface, enabling higher specialization in development and flexibility of use [2]: “Separation lets specialists develop the user interface and the application independently, promotes interface consistency across applications, and allows application functions to be added or combined in new ways.” The separation makes user interface development more efficient because the design, building, and evaluation of the user interface are separated from the code of the application. It allows for implementation of different interaction methods, potentially using different styles and modalities interchangeable for expressing similar user input. It implies that the user interface must have sufficient access to application internals in order to keep the user aware of the application semantics (the application objects, operations and effect of the interaction).

The research described in [3] elaborates the characteristics of a system and derives the fundamental structure of a framework. The main requirements to the design of the framework are abstraction, architectural design, and being independent from hard- and software. The fundamental idea is that the user interface should be defined not in terms of syntax or states of the application but in terms of the meaning of the interaction to the application. The specification of the framework identifies the components, defines the relationships between the components and illustrates the data flow within an intended system.

The presentation will introduce the framework using virtual input devices to specify the input of the user interface without constraints regarding metaphor, shape, location, or modality. It covers three main components: The definition of the input event, a client interface and a service interface.

References


A Multimodal Fusion Framework For Interactive Applications

Hildeberto Mendonça, Olga Vybornova, Benoit Macq
Université catholique de Louvain - UCL
Bâtiment Stévin – 2, Place du Levant
Louvain-la-Neuve, 1348 Belgium
+32 10 47 80 75
{hildeberto.mendonca, olga.vybornova, benoit.macq}@uclouvain.be

ABSTRACT

This research aims to propose a multimodal fusion framework for high-level data integration between two or more modalities. It takes as input extracted low level features from different system devices, analyzes and identifies intrinsic meanings in these data through dedicated processes running in parallel. Extracted meanings are mutually compared to identify complementarities, ambiguities and inconsistencies to better understand the user intentions when interacting with the system. The whole fusion lifecycle will be described and evaluated in an ambient intelligence scenario, where two co-workers interact by voice and movements, demonstrating their intentions and the system gives advices according to identified needs.

Keywords

Multimodal fusion, ambient intelligence, speech recognition, context-sensitive interaction.

1. INTRODUCTION

Multimodal systems have progressively evolved towards more robust semantic interpretation and they all include some fusion mechanism to combine the separate data streams by reducing uncertainty. To understand and formalize the coordination between modalities involved in the same multimodal interface, it is required to extract relevant features from signal representations and thereafter to proceed to high-level (semantic-level) fusion.

To provide natural and robust interaction with the user(s) the system must be able to:

a) interpret as a whole the various modalities used by a human;
b) extract information arriving simultaneously from different devices and other components; and
c) combine them into one or several unified and coherent representations of the user’s behavior.

A challenge is to fuse these different input modalities effectively to augment their natural strengths and use redundancy to increase robustness. Moreover, it is critical to decide which information should be fused and which should not.

Many fusion engines have been developed so far with different approaches to allow data fusion [2] [5] [10], but they have specific purposes, solving particular scenarios of fusion. When a new scenario is identified, a common practice is to implement the fusion from scratch due to the lack of tools or frameworks providing the basis for multiple implementations. We believe that the field can be substantially improved if a general support for fusion is provided, taking into consideration common features of existing fusion mechanisms. This research aims to bridge this gap by proposing a fusion framework, based on a common understanding about multimodal fusion, to allow researchers to focus on data processing and integration instead of complex problems such as concurrency, synchronization, distribution and state management.

2. APPROACH

We observed the need for a fusion framework when a fusion scenario was identified, modeled and implemented considering two main modalities, which were speech recognition and human behavior analysis through image processing tools. We have conducted experiments in a hypothetical context, with two or more people interacting with each other and with objects in the scene [13] [14]. We observed that when intending to perform an action, the user(s) might:

- speak about it describing their actions or intentions (the ideal case, then the linguistic and action recognition data are complementary);
- be doing things, but speaking about something absolutely irrelevant to those actions (in this case, conflict resolution techniques must be applied to detect which data should be considered);
- be just silent when performing actions (in this case the unimodal operation mode is required and no data fusion is needed).

Everything that is said or done is meaningful only in the particular context. To accomplish the task of semantic fusion we should take into account the information obtained at least in the following three types of context [1]:

- domain context: meaning prior knowledge of the domain, semantic frames with predefined action patterns, user profiles, situation modeling, a priori developed and dynamically updated ontology defining subjects, objects, activities and relations between them for a particular person;
- conversation context: derived from natural language semantic analysis; and
- visual context: capturing the user’s gesture/action in the observation scene and allowing eye gaze tracking to enable salience models.

In our approach the high level fusion of modalities is performed in three stages:

- early stage merges data coming from low level data processing or fusion to provide reliable reference resolution;
- late stage integrates all the information to give the final interpretation of the human behavior [13]; and
- **reinforcement stage** executes one or more cycles of verification, reevaluating all identified meanings to check complementarities, redundancies, perform disambiguation and conflict resolution [14].

During the early stage, some data received from each modality should be synchronized in order to strengthen the link between data and optimize the late stage. An example of early stage done in our experiment is person identification from the speech and from image features. It helps us to complement a recognized intention from the speech with the behavior of the speaker on the late stage, predicting which decision is more appropriate for the situation.

When the early stage finishes, the search for semantics continues with the application trying to identify plurality of meanings. This is the input of the late stage, which deals with higher symbolic elements abstracted by the modality recognizers. These elements are first matched between the different modalities and then integrated during the final decision stage to resolve uncertainty and produce the user's intention interpretation. One example of plurality of meanings on speech recognition is to detect the human intention linking the person (subject) with concrete or abstract things (object) through an action or relation (predicate). In the sentence “I want to call Nick” there is a clear intention represented by the links between “I”, “want to call” and “Nick”, which are subject, predicate and object, respectively. These meanings are combined with other meanings during the late stage.

In addition to the early and late stages, we are contributing with the **reinforcement stage**, which is a cyclical process that uses all detected meanings to reinforce or redefine them, thus producing better inputs for the late stage. The need for reinforcement came from a reflection about how an existent modality can help the analysis of other modalities preserving the modality cohesion, where a modality must not be aware of the presence of other modalities in order to avoid some level of coupling. We could observe this need in our experiment, when we implemented n-best re-ranking functions [5] for the speech recognition analysis to find the best hypothesis about what was said. These functions need direct access to the list of concepts from the ontology and what was analyzed from other modalities in order to probabilistically rank the best hypothesis possible.

The general schema of the high level fusion is presented in the Figure 1.

3. **SCENARIO**

In order to develop and validate the whole experiment, we defined five different scenarios where two people interact mutually in a room, talking in a natural way and behaving without restrictions. Each scenario tries to explore specific combinations of speech and behavior to increase the robustness of the system. This system is able to track people, analyze their behavior, movements, speech, and takes decisions about how to prompt them necessary information when required or provide any other assistance.

The main challenges that we face in this application are unrestricted human language and free natural behavior within home/office environment. In order to analyze behavior efficiently, the system has to correctly process, interpret and create joint meaning of the data coming from speech analysis and video scene analysis. We consider that human behavior is goal-oriented, so our main aim is to recognize the users’ plan and to see what they want to achieve.

The system manages the data streams arriving from two sources – video scene and speech. In particular, we show a technique distinguishing between the data from different modalities that should be fused and the data that should not be fused, but analyzed separately.

4. **FUSION MECHANISM**

The fusion mechanism [15] instantiated on the fusion framework employs various components integrated into a single system: speech recognition (Sphinx-4 [16]), speaker identification (implemented), syntactic parsing (C&C Parser [4]), natural language semantic analysis (C&C Boxer [4]), video analysis (Open CV [7]), human behavior analysis component (implemented) and a cognitive architecture that serves as the context-aware controller of all the processes, manages the ontology (Protégé [11]) and provides the decision-making mechanism (Soar [12]).

Figure 2 depicts these components integrated and coordinated by the fusion framework, which is based on OpenInterface (OI) [9], a platform to interconnect, in a pipeline, components developed in different languages. The process starts when an audio or a video signal is detected by the first time. There is no restriction about which signal should start first because all modalities can be processed in parallel and independently. When an audio stream is received, the speech recognition component processes it, generating a string of what was said. The same signal is sent to the speaker identification component, which will associate what was said with who said that. The string is sent to the syntactic parsing component to identify the syntax of each word, which is important for the natural language semantic analysis component, responsible for the identification of the subject, the agent, the predicate, the object of interest and other elements. From the semantic analysis, it is possible to extract semantic structures very similar to the structure of the knowledge base, represented by ontology. If we find the identified semantic in the ontology then it means that the sentence is valid inside the context and can be useful to fuse with other meanings coming from other modalities.

---

Figure 1: General schema of high level fusion

Inspired by several approaches of modality fusion and reinforced by our own experience, this proposal of a multimodal fusion framework has been designed to address a large variety of cases. The first validation of this framework is described in the next section, where our approach has been applied.
On the other hand, when a video stream is detected, the image is processed by an image processing component called OpenCV. This component analyzes some image features to calculate the position of each person on the horizontal plan of the scene, their movements direction and identifies who is each person according to a predefined profile. It is also important to identify people through this modality because we have to know the position of who is talking in order to associate the user intention with his/her actions. The synchronization of detected users in each modality is done during the early stage.

The next step is to analyze the human behavior, comparing the movements of the user with a set of rules. The behavior is relative to fixed objects in the scene, which are defined in the context domain and are directly associated with the aid to be given by the system. The rules define the boundaries of what is near or far from a certain object. Then the result of the rule processing at this stage is: [person] is near the [telephone], [person] is far from the [computer] or [person] is moving to the [library]. This result is produced for each person in each frame of the video. Individually, these results are not significant enough for fusion. We have to analyze the movements in many frames in order to have final conclusions. For instance: if in the last 80 frames, the rule engine produced “[person] is moving to the [library]” then we can conclude that there is a real intention to achieve the library, considering some variables of the environment, such as area of the room. The late stage occurs when we identify a person moving to the library and we also detect the intention to find a book in a sentence like this: “I can find a book about it in the library”. Therefore, we can conclude the person is moving to the library because he/she wants to find a certain book there.

5. RESULTS

The composition of many different tools and components was definitively a challenge, achievable due to the framework that has been developed on top of OI. We used Java and C/C++ programs and OI allowed the communication between all of them without any particular change in the source code.

We are using only open source software to compose each part of the fusion mechanism. Unfortunately, these tools follow the state of the art very slowly and we could not get results that are possible to expect from proprietary or inaccessible tools developed by companies and recognized research centers. One of the strongest impacts of this slow evolution was in the speech recognition modality, where Sphinx could not provide precise results and C&C Tools were imprecise on semantic analysis when ignoring interrogative sentences and other punctuation. On the other hand, the computer vision modality provided precise results about people positioning and movement directions, due to the rich framework OpenCV.

The decision making part was strongly impacted by the poor results of the speech modality. However, we could provide two advices during the experiment: 1) Nick’s phone number because Ronald’s intention was easily recognized together with his movements towards the telephone; and 2) the localization of the book about French wines in the library because of good recognition results. This fusion mechanism instance contributed to validate the life cycle of the fusion framework, represented in the Figure 3. Each channel represents a modality. Two devices were connected to capture signals. These signals are processed in the recognition phase. For each detection, the framework creates a segment in the segmentation phase, delimitating useful data. Segments are processed and meanings are extracted from them. These meanings are annotated in the annotation phase. During the modality life cycle, an event-oriented architecture executes threads in parallel to perform the low level fusion and the three stages of high level fusion as represented in the Figure 1.
6. CONCLUSION
We have posed and solved several problems during this challenging research. Using the current proposal of the framework, we could instantiate a fusion mechanism to:

- manage spatial relationships based on the fixed objects in the room;
- make semantic fusion of events not coinciding in time;
- achieve good results in speaker identification - synchronization between image and speech identification;
- create an open framework to manage fusion between two (in our case) or more modalities (in enhanced future work); and
- design the system so that each component can run in a separate machine due to the distribution mechanism interchanging data through a TCP/IP network.

However, we have more problems to solve in our future work. To name just a few, we should:

- implement an effective learning mechanism;
- perform efficient decision making, even from information fragments;
- perform 3D video analysis;
- identify detection of orientation of the people in the scene;
- add at least one more modality, e.g. eye gaze tracking;
- recognize various types of gestures; and more.

7. REFERENCES
ABSTRACT
In this paper, the Multimodal Browser is presented as an interactive tool to enable dynamic design of multimodal configurations. As the message forwarding can be managed using simple rules and components can be added or removed independently, the Multimodal Browser can be used to interactively enable or disable connections between different components, changing interactions techniques or adding special scriptable components started as independent applications to modify the message flow. These dynamic design features were developed as a side effect when the first version of the Multimodal Browser was developed for mobile phones due to the restriction to start external applications and to the migration from library based approach to a networked one in order to allow end user customization and the storage of special profiles. These features have been explored during the development of the Multimodal Browser itself, facilitating experimentation and the communication between multiple devices suggesting it as an interesting dynamic design tool.

Categories and Subject Descriptors
H5.2 [Information interfaces and presentation]: User Interfaces – Prototyping. D2.2 [Software Engineering]: Design Tools and Techniques – Modules and interfaces; user interfaces. D2.m [Software Engineering]: Miscellaneous – Rapid Prototyping; reusable software.

General Terms
Management, Design, Experimentation.

Keywords
Multimodal middleware protocol, multimodal browser, multimodal hub, dynamic design, multimodal design.

1. INTRODUCTION
The Multimodal Browser was developed during the European STREP project OpenInterface[1] to enable the evaluation of multimodal applications on mobile phones. The original intent was to develop a management tool for a network of multimodal components running on these devices. The Multimodal Browser was then used to configure end user profiles previously designed after experimentation using the OI Framework and tools.

As the target environment was mobile phones, a new network based protocol was developed to enable the data exchange between components and it was called the Multimodal Middleware Protocol. To manage the connections between different components, another tool called the Multimodal Hub was developed. The Multimodal Hub acts like a network hub gathering all points (components) to build a single network.

These tools combined with multimodal components enabled the successful run on mobile phones using the Java Micro Edition and demonstrated interesting side effects as features for designing new applications. One of those features was the independence on the activation order of components and also the support for components disconnections and reconnections. Another interesting feature is the fact that the Multimodal Browser can be used to inject new rule sets into the Multimodal Hub message routing, changing the component network behavior during runtime. The open source implementation of the protocol allows the rapid prototyping of new components using the Python programming language.

An architecture overview of the Multimodal Hub and Multimodal Browser is presented on figure 1. The Multimodal Hub is the center of the network, receiving and forwarding messages to all components. The Multimodal Browser has a direct link with the Multimodal Hub, allowing fine control of internal structures, but can also be decoupled as a standalone component.

Figure 1 - Architecture Overview

2. Connection order independence
Once a new component is connected to the Multimodal Hub, it announces a list of events it is capable to produce and another list with events it can consume. A component produces an event when it generates a message with specific event identification number. An event is consumed when it is received by a component. The Multimodal Hub collects this event information from all components, continually updating its lists based on component connections or disconnections.
The event lists are matched using an event identification number. When the Multimodal Hub detects a component can produce an event another one can consume, it starts a subscription forwarding messages from the producer to the consumer. The Multimodal Hub is also able to copy a message to many consumers. This dynamic behavior allows component connection order independence, as the final message flow will be the same.

This feature allows new components to join or leave the network anytime. Even similar components can be connected at any time, allowing multiple forms of control to work simultaneously, like a joystick and mouse control in parallel with speech recognition to trigger special application events.

3. Dynamic Rule Sets support
The standard behavior to connect components can be controlled using a dynamic rule set. The rule set is specified by a XML file into the Multimodal Browser. Once applied, the rule set is immediately effective and independent of the components already connected. A rule set can be specified even if no components are connected, as it is applied after every message reception and before forwarding.

The rules define which events a component can consume or produce, but they also support the destination identification or source component. The rule set can be used as an application profile and be changed as the user moves to another context. Using the same principle, a component can be used to expose the rule management capabilities of the Multimodal Browser to control which rule set should be active or to simply change user preferences without stopping any other component.

4. Scripting Components
As new components can join the network, the use of components written in a script language like Python can easy the prototyping of complex applications. These new components would have access to the expression power of a full programming language and not only a limited subset of functionality normally exposed in a dialog interface.

The open source project Python Multimodal Hub[2] provides a complete implementation of the Multimodal Hub and the Multimodal Middleware Protocol. This implementation provides a simple example to use library and documentation to easily write new components with few lines of code. For operations like translation and filtering of messages, new components can be developed with few lines of code. The network complexity is completely managed by the library itself, allowing a script writer to focus on his component and not on programming details.

Using the Multimodal Middleware Protocol and the open source reference library, new scripting languages can be added and visual tools can be used to generate simple code or stubs for new components. The aim of an open source project is to collect and share experiences from its user community, allowing further development to happen naturally and guided by collective need.

5. Free to experiment
With connection order independence, dynamic rule set changing and easy component writing, the Multimodal Browser and the Multimodal Hub becomes a dynamic design engine, where components can be changed freely and experiments can be performed with different configurations. The Multimodal Middleware Protocol network capabilities also enable a network of devices to build larger configurations in a distributed environment. This resource also enables the research and use of heavy processing algorithms without impacting the performance of other components located in different computers or devices. The multiplatform features of the protocol enable the combination of software running in different operating systems and connected by different network protocols to build a single multimodal component network. An example for this scenario would be parallel image recognition engines acting on the same image source, but running with different algorithms. This flexibility allows researchers to change components and experiment new combinations of components to build richer applications in less time.

6. Known limitations
Components running on independent applications are easier during the prototyping of final application, but they have to be very well managed to avoid deployment issues later. The use of a proxy component responsible for starting all other components is a solution for this problem, but it also increases the complexity of the component network as it turns the proxy into a middleman.

A dialog based user interface can offer a more user friendly approach to a new multimodal application design and experimentation but compromises the power of adaptation and changes to new requirements.

Another important limitation is the use of network services for message exchange, adding latency to the component network, but this collateral effect can be overcome by the benefits of application isolation, flexibility and reduced multithreading and multiprocess issues.

7. Conclusions
The Multimodal Browser and the Multimodal Hub can be used in an experimentation environment where the designers have access to basic programming techniques. This approach can be enhanced with better visual tools and a complete integration with OIDE, providing a better mix of scripting and visual design techniques leading to higher levels of experimentation and customization of components output and input flows.

The Multimodal Browser is also an interesting environment for experimentation and not only final execution on mobile phones as initially planned. The flexibility gained by dynamic component connections management, support of easily scriptable components and custom rule sets compose a new tool chain to develop multimodal applications.

8. ACKNOWLEDGMENTS
The work described on this paper was mainly founded by the European FP6-35182 OpenInterface Project and the Walloon Region of Belgium.

9. REFERENCES
The activation of modality in virtual objects assembly

Nadine Couture\textsuperscript{1,2} and Guillaume Riviere\textsuperscript{1,2}

\textsuperscript{1} ESTIA, Technolole Izarbel, F-64210 Bidart, France
\textsuperscript{2} LaBRI, UMR 5800, CNRS, Université Bordeaux 1, IPB, 351, cours de la Libération, 33405 Talence, France
{n.couture|g.riviere}@estia.fr

Abstract. In the specific case of the assembly of virtual 3D objects, using tangible bimanual interaction, we discuss the choice of modality in order to activate another modality.

Keywords: CHI, TUI, Multimodality, Virtual Assembly.

1 Context and related works

We place our analyse in the specific case of the assembly of virtual 3D objects using tangible bimanual interaction. In that situation, both hands are requisitioned and the user can not drop the props to activate a new modality without changing his action or the expected results. Various solutions exist. It is this point that we wish to discuss at the workshop: the design and development of multimodal interaction, for virtual objects assembly, by the activation of a modality by another modality.

Given \( m \) the modality such that \( m=<\text{prop}, (x,y,z)> \), where to move the prop leads to move the corresponding virtual object in the \((x,y,z)\) 3D coordinates. The problem is to determine a modality in order to activate/inactivate the \( m \) modality. When the modality \( m \) is inactivated, moving props has not consequences on the 3D position of virtual objects. We focus on three potential modalities: the vocal modality, the modality embedded physically in the principal devices (the props) and the modality embedded physically in a foot pedal device.

The vocal modality is a solution commonly suggested by the users. The voice can actually be used to confirm, cancel or trigger an operation. In 1980, Bolt \cite{Bolt1980} first combined voice and gesture to interact with data projected onto a large screen. In \cite{PassiveProps}, the manipulation of Passive Props is combined with the voice modality in a situation similar to our case study. Authors explain that the latency due to the voice recognition system is a source of embarrassment and they highlight the interference that may exist between the speech and the short-term memory. Despite the wishes of users, the voice modality is not a suitable solution in this case. This is mainly due to the fact that a discrete operation must be triggered by a brief and discrete action. The modality embedded physically in the principal devices (the props) is illustrated in I/O Brush \cite{I/OBrush}. The change of modalities exists in order to modify the use of the brush: acquire an image, acquire a colour and acquire a sequence of images. The change of
modalities is performed by turning a knob attached to the handle of the brush. Quoting the example of the declutching mechanism [2] a button was also embedded to a rectangular sheet. This button, activated by the thumb of the hand, allows user to start controlling the position of cutting plane represented by the rectangular sheet. These apparent equipments affect the conventional use of props. Thus, even if embedded physically actuators on the props are a solution that may seem "self-evident", this is not a good solution in an ubiquitous oriented design. Furthermore, from the handling point of view, these equipment buttons leads to problems of discomfort, loss of freedom and loss of accessibility and may also be activated unintentionally. It is therefore difficult to integrate this kind of modality efficiently in the interaction.

The modality embedded physically in a secondary device, a foot pedal, appears both in the previous example [2] and in [4]. The process of declutching has been delegated to a foot pedal. The users controlled the movement of the virtual objects by keeping the foot pedal pressed. Thereby, the user keeps a total freedom of movement with his/her hands.

2. Exploration of solutions

Our aim is the design and the development of a new tangible user interface, based on multimodal interaction, for the efficient assembly of 3D virtual objects. The key idea of the system is to use two props as physical representation and control for the virtual objects. In each hand, the user manipulates an electromagnetically tracked prop, and the translations and rotations are directly mapped to the corresponding virtual objects on the display. During the process of assembly, the user needs to reposition his/her hands or to switch props. Hence, the system has to provide a modality to clutch hand movements.

By relying on the state of the art, illustrated above, we have eliminated the voice modality due to the fact that it is not necessarily the technology that limits the use of the voice modality, but the user's capacities. We then conducted a user study to confirm the superiority, in our case study, of the modality <foot pedal, activate/desactivate> and <button on props, activate/desactivate>. The subjects had to accomplish a series of 6 assemblies of various shapes. The conditions were between-subject, and the preference between pedals and buttons was evenly distributed with a slight preference for the foot pedals. The user study revealed that the movement of the hands is more similar to real-world assembly scenarios when using the two foot pedals and that the users can keep on concentrating on the actual assembly task.

In conclusion, we have displayed the fact that to achieve some tasks, a modality has to be activated to realise the principal modality. Thus, it is often necessary to add equipment in the system. In the specific case of the assembly of virtual 3D objects, using tangible bimanual interaction, we shown that to activate the modality <prop, (x,y,z)> the best modality is <foot pedal, on/off>.
Acknowledgments. Thanks to P. Reuter who has the leadership on the ArcheoTUI system used in the experimentations.

References

Workshop on the Challenges of Engineering Multimodal Interaction: Methods, Tools, Evaluation

Wednesday 13 May
Fraunhofer Institute for Applied Information Technology
Sankt Augustin (Bonn), Germany

Workshop Organized by
Sebastian Denef, Fraunhofer FIT, Germany,
Marilyn McGee-Lennon, University of Glasgow, UK
Laurence Nigay, Universite Joseph Fourier, France,
Philip Gray, University of Glasgow, UK
<table>
<thead>
<tr>
<th>Full Name</th>
<th>Affiliation</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diego Arnone</td>
<td>CALLAS Consortium</td>
<td><a href="mailto:diego.arnone@eng.it">diego.arnone@eng.it</a></td>
</tr>
<tr>
<td>Lynne Baillie</td>
<td>Glasgow Caldedonian University</td>
<td><a href="mailto:L.Baillie@gcal.ac.uk">L.Baillie@gcal.ac.uk</a></td>
</tr>
<tr>
<td>BELAUNDE Mariano</td>
<td>Orange Labs (Lannion - France)</td>
<td><a href="mailto:mariano.belaunde@orange-ftgroup.com">mariano.belaunde@orange-ftgroup.com</a></td>
</tr>
<tr>
<td>Slim Ben Hassen</td>
<td>Orange Labs (Lannion - France)</td>
<td><a href="mailto:slim.ben.hassen@orange-ftgroup.com">slim.ben.hassen@orange-ftgroup.com</a></td>
</tr>
<tr>
<td>Stefano Bigi</td>
<td></td>
<td><a href="mailto:S.Bigi@cbt.it">S.Bigi@cbt.it</a></td>
</tr>
<tr>
<td>Roberto Cavalcante</td>
<td>Multitel ASBL</td>
<td><a href="mailto:cavalcante@multitel.be">cavalcante@multitel.be</a></td>
</tr>
<tr>
<td><strong>Luca Chittaro</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marcello Coiana</td>
<td>Arcadia Design</td>
<td><a href="mailto:marcello.coiana@archadiadsgn.it">marcello.coiana@archadiadsgn.it</a></td>
</tr>
<tr>
<td>Alex Conconi</td>
<td>TXT</td>
<td><a href="mailto:alex.conconi@txt.it">alex.conconi@txt.it</a></td>
</tr>
<tr>
<td>Jean-Baptiste de la</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Riviére</td>
<td>Immersion, France</td>
<td><a href="mailto:jb.delariviere@immersion.fr">jb.delariviere@immersion.fr</a></td>
</tr>
<tr>
<td>DE TURENNE François</td>
<td>France Telecom</td>
<td><a href="mailto:marc.mersiol@orange-ftgroup.com">marc.mersiol@orange-ftgroup.com</a></td>
</tr>
<tr>
<td>Sebastian Denef</td>
<td>FIT</td>
<td><a href="mailto:sebastian.dene@fit.fraunhofer.de">sebastian.dene@fit.fraunhofer.de</a></td>
</tr>
<tr>
<td>Nicolas Dittlo</td>
<td>Immersion, France</td>
<td><a href="mailto:nicolas.dittlo@immersion.fr">nicolas.dittlo@immersion.fr</a></td>
</tr>
<tr>
<td>Cindy Dorfmann</td>
<td>FIT</td>
<td><a href="mailto:cindy.dorfmann@fit.fraunhofer.de">cindy.dorfmann@fit.fraunhofer.de</a></td>
</tr>
<tr>
<td>Bruno Dumas</td>
<td>University of Fribourg, Switzerland</td>
<td><a href="mailto:bruno.dumas@unifr.ch">bruno.dumas@unifr.ch</a></td>
</tr>
<tr>
<td>Guillaume Gauffre</td>
<td>IRIT, University of Toulouse Design</td>
<td><a href="mailto:gauffre@irit.fr">gauffre@irit.fr</a></td>
</tr>
<tr>
<td>Pinar Gorsev</td>
<td>PCLI</td>
<td></td>
</tr>
<tr>
<td>Phil Gray</td>
<td>University of Glasgow</td>
<td><a href="mailto:pdg@dcu.gla.ac.uk">pdg@dcu.gla.ac.uk</a></td>
</tr>
<tr>
<td>Clio Kakoulli</td>
<td>FIT</td>
<td><a href="mailto:elio.kakoulli@fit.fraunhofer.de">elio.kakoulli@fit.fraunhofer.de</a></td>
</tr>
<tr>
<td>Werner König</td>
<td>University of Konstanz</td>
<td><a href="mailto:werner.koenig@uni-konstanz.de">werner.koenig@uni-konstanz.de</a></td>
</tr>
<tr>
<td>Andreas Lorenz</td>
<td>FIT, Sankt Augustin, Germany</td>
<td><a href="mailto:andreas.lorenz@fit.fraunhofer.de">andreas.lorenz@fit.fraunhofer.de</a></td>
</tr>
<tr>
<td>Jean-Claude Martin</td>
<td>LIMSI-CNRS</td>
<td><a href="mailto:MARTIN@LIMSI.FR">MARTIN@LIMSI.FR</a></td>
</tr>
<tr>
<td>Marilyn McGee-Lennon</td>
<td>University of Glasgow</td>
<td><a href="mailto:mcgeemr@dcu.gla.ac.uk">mcgeemr@dcu.gla.ac.uk</a></td>
</tr>
<tr>
<td>David McGookin</td>
<td>University of Glasgow</td>
<td><a href="mailto:mcgookdk@dcu.gla.ac.uk">mcgookdk@dcu.gla.ac.uk</a></td>
</tr>
<tr>
<td>Hildeberto Mendonca</td>
<td>UCL</td>
<td><a href="mailto:hildeberto.mendonca@uclouvain.be">hildeberto.mendonca@uclouvain.be</a></td>
</tr>
<tr>
<td>Nilo Menezes</td>
<td>Multitel ASBL</td>
<td><a href="mailto:menezes@multitel.be">menezes@multitel.be</a></td>
</tr>
<tr>
<td>Laurence Nigay</td>
<td>Universite Joseph Fourier Grenoble 1</td>
<td><a href="mailto:Laurence.Nigay@imag.fr">Laurence.Nigay@imag.fr</a></td>
</tr>
<tr>
<td>Reinhard Oppermann</td>
<td>FIT</td>
<td><a href="mailto:reinhard.oppermann@fit.fraunhofer.de">reinhard.oppermann@fit.fraunhofer.de</a></td>
</tr>
<tr>
<td>Leonardo Ramirez</td>
<td>FIT</td>
<td><a href="mailto:leonard.o.ramirez@fit.fraunhofer.de">leonard.o.ramirez@fit.fraunhofer.de</a></td>
</tr>
<tr>
<td>Andrew Ramsay</td>
<td>University of Glasgow</td>
<td><a href="mailto:adr@dcu.gla.ac.uk">adr@dcu.gla.ac.uk</a></td>
</tr>
<tr>
<td>Steve Renals</td>
<td>University of Edinburgh</td>
<td><a href="mailto:s.renals@ed.ac.uk">s.renals@ed.ac.uk</a></td>
</tr>
<tr>
<td>Guillaume Riviere</td>
<td>ESTIA-Biarritz</td>
<td><a href="mailto:g.riviere@estia.fr">g.riviere@estia.fr</a></td>
</tr>
<tr>
<td>Marcos Serrano</td>
<td>Universite Joseph Fourier Grenoble 1</td>
<td><a href="mailto:Marcos.Serrano@imag.fr">Marcos.Serrano@imag.fr</a></td>
</tr>
<tr>
<td>Kenia Sousa</td>
<td>UCL-BCHI</td>
<td><a href="mailto:kenia.sousa@gmail.com">kenia.sousa@gmail.com</a></td>
</tr>
<tr>
<td>Philippe Truillet</td>
<td>IRIT, University of Toulouse Design</td>
<td><a href="mailto:truillet@irit.fr">truillet@irit.fr</a></td>
</tr>
</tbody>
</table>