Project ref. no.	FP6-507609
Project acronym	SIMILAR
Deliverable status	R
Contractual date of delivery	31 May 2005
Actual date of delivery	August 2005
Deliverable number	D18
Deliverable title	Current practice description and evaluation: Towards a framework for usability evaluation
Nature	Report
Status & version	Final
Number of pages	36
WP contributing to the deliverable	SIG7
WP / Task responsible	NISLab
Editor	Laila Dybkjær
Author(s) (alphabetic order)	Enrique J. Gómez Aguilera, Niels Ole Bernsen, Laila Dybkjær, Pedro Correa Hernandez, Benoit Macq, Georgios Nikolakis, Pablo Lamata de la Orden, Fabio Paternò, Carmen Santoro, Daniela Trevisan, Dimitrios Tzovaras, and Jean Vanderdonckt
EC Project Officer	Mats Ljungqvist
Keywords	Usability, evaluation, current practice, framework.
Abstract (for dissemination)	This deliverable provides an overall description and evaluation of current practice in usability evaluation of multimodal and natural interactive systems within the areas of expertise of the partners. It thus covers the areas of spoken multimodal dialogue systems, vision-based systems, haptics-based systems, mixed-reality systems in surgery, and tools for remote usability evaluation. Moreover, the report outlines steps towards a usability evaluation framework to be further investigated.

Deliverable D18

Current Practice Description and Evaluation: Towards a Framework for Usability Evaluation

GBT, Polytechnic University of Madrid, Spain
ISTI-CNR, HIIS Laboratory, Pisa, Italy
ITI-CERTH, Greece
NISLab, University of Southern Denmark
Tele, Université catholique de Louvain, Alterface SA

August 2005

Section responsibilities

Section 1: NISLab, Denmark: Laila Dybkjær and Niels Ole Bernsen
Section 2: NISLab, Denmark: Laila Dybkjær and Niels Ole Bernsen
Section 3: Tele, Université catholique de Louvain, Alterface SA, Belgium: Pedro Correa
Section 4: ITI-CERTH, Greece: Giorgos Nikolakis and Dimitrios Tzovaras
Section 5: Tele, Université catholique de Louvain, Belgium: Daniela Trevisan, Benoit Macq and Jean Vanderdonckt
GBT, Polytechnic University of Madrid, Spain: Pablo Lamata and Enrique Gomez
Section 6: ISTI-CNR, HIIS Laboratory, Pisa, Italy: Fabio Paternò and Carmen Santoro
Section 7: NISLab, Denmark: Laila Dybkjær and Niels Ole Bernsen

Contents

1 Introduction

This deliverable gives an overall description and evaluation of current practice in usability evaluation of multimodal and natural interactive systems within the areas of expertise of the partners. The areas covered include spoken multimodal dialogue systems, vision-based systems, haptics-based systems, mixed-reality systems in surgery, and tools for remote usability evaluation. Moreover, the report outlines steps towards a usability evaluation framework to be further investigated. The present deliverable builds on SIMILAR Deliverable D16 which provided an overview of the state-of-the-art in usability evaluation within the partner areas mentioned.

Section 2 on spoken multimodal dialogue systems evaluation outlines current practice usability evaluation methods and criteria and discuss problems related to, in particular, criteria and their definition and use. Future challenges concerning methods and criteria are outlined. The proposal for work towards a framework includes a thorough description of well-defined methods and criteria as well as a further investigation of new methods and criteria needed for new system types, and of criteria which are not well-defined yet.

Section 3 on vision-based systems evaluation briefly presents state-of-the-art and main challenges in this area. When evaluating vision-based systems focus is on performance in a very broad sense, including, e.g., ease of use, but it is not necessarily straightforward how to measure performance, and it is not the same set of criteria that is equally relevant for all vision-based systems. Evaluation methods are discussed, including experimental methods as well as more formal testbed evaluation. With respect to a framework, also this section points out the need to list known methods and criteria along with a description of how and when to use them.

Section 4 focuses on usability evaluation of haptics-based systems. Different categories of current practice evaluation criteria and methods are presented. A problem with many criteria is that they are not well defined. Weaknesses and limitations of current practice in evaluation are pointed out and new practices are proposed to overcome some of the current problems. Like the previous sections, this section stresses the need for a framework which allows evaluators to select a subset of criteria for their next evaluation job. The selection of criteria depends, among other things, on the development stage of the system to be evaluated.

Section 5 concerns usability evaluation of mixed reality systems in surgery. It highlights the problem that current practice in evaluation of such systems does not sufficiently address human-computer interaction problems and integration in the clinical context. Important issues proposed for an evaluation framework include perceptual properties, cognitive properties and the naturalness of the interaction process (e.g., functional properties). Also the importance of didactic resources for teaching basic and more advanced skills must be evaluated and may be included in an evaluation framework.

Section 6 on tools for remote usability evaluation outlines current practice techniques in remote usability evaluation. As challenges for remote usability evaluation it mentions the reduction of cost and effort involved in evaluation and that it is important to capture the user's context of use from many data sources, including emotional state data, to make an appropriate evaluation. Issues related to the evaluation of multi-device, migratory, and multimodal interfaces are mentioned as important for further investigation with the increasing use of mobile devices.

Section 7 discusses similarities and differences between the areas presented in Sections 2-6 and discusses whether a single joint framework is likely to be a good idea.

2 Usability evaluation issues in spoken multimodal dialogue systems

2.1 Introduction

Foolproof usability evaluation of spoken multimodal dialogue systems (SMDSs) does so far not exist and maybe never will. Nevertheless we should try to improve the current state-of-the-art. The main problem is that we still know too little about exactly which parameters contribute to usability and with which weights. On the other hand, what we know already should certainly be used and open issues should be further explored. Usability have in recent years moved into focus as an important competitive parameter and is likely to remain important since more and more ordinary people with no particular computer skills are using an increasing amount of electronic devices which can run spoken multimodal dialogue applications. At the same time, applications are becoming increasingly sophisticated and powerful which implies new challenges for usability evaluation.

2.2 Current practice in usability evaluation of spoken multimodal dialogue systems

Usability evaluation should be done throughout the development life cycle and as an integrated part of development. When performing a usability evaluation of a system one normally uses a particular method and a set of evaluation criteria which specify what to look for and evaluate.

2.2.1 Current practice usability evaluation methods

A number of different evaluation methods are available, including, e.g., expert reviews, heuristic evaluation, cognitive walkthroughs, prototype testing, Wizard-of-Oz, controlled laboratory tests, and field tests, see, e.g., http://usabilitynet.org/tools/methods.htm. Which method(s) to use depends, among other things, on where in the life cycle the system is and which resources are set aside and are available for usability evaluation. It also depends on the kind of evaluation one wants to carry out. Diagnostic evaluation has its focus on identifying and diagnosing errors in order to repair the system so that they will not appear again. Performance evaluation concentrates on measuring the user's performance with the system. Adequacy evaluation has its main focus on how well the system fits its purpose and meets actual user needs and expectations. For example, expert reviews and heuristic evaluation may work well for diagnostic evaluation but are not very suitable if the focus is on user performance. For realistic measurements of many aspects of performance an implemented system is needed whereas adequacy may very well be measured by using, e.g., a Wizard-of-Oz simulated system. The above usability evaluation methods belong to current practice and do not only apply to SMDSs but are general to software systems.

As mentioned in SIMILAR deliverable D16, it is common with respect to task-oriented systems to run a test involving users on two different prototypes or systems and then compare their usability based on the test results. A related exercise is to test two system versions, e.g., with and without an animated agent, to see which effect the agent has on the usability results. A fairly rarely used method is theory-based evaluation. The problem with theory-based evaluation is that it requires a theory and for many issues no theory is (yet) available.

In our opinion there exists a fair range of evaluation methods which can help designers and developers improve the usability of their systems. The main problem related to methods is probably a resource problem where 'resources' refer to both people, time and money. There must be people with the right skills to select one or more suitable methods and people who are familiar with the method(s) to be applied. There must be time to conduct usability testing, and there must be a reasonable budget in terms of money to do it.

2.2.2 Current practice usability evaluation criteria

Evaluation can be quantitative or qualitative, subjective or objective [Dybkjær and Bernsen 2000]. *Quantitative* evaluation consists in quantifying some parameter through an independently meaningful number, percentage etc. which in principle allows comparison across systems. *Qualitative* evaluation consists in estimating or judging some parameter by reference to expert standards and rules. *Subjective* evaluation consists in judging some parameter by reference to users' opinions. *Objective* evaluation produces subject-independent parameter assessment. Ideally, we would like to obtain quantitative and objective progress evaluation scores for usability which can be objectively compared to scores obtained from evaluation of other SMDSs. However, many important usability issues cannot be subjected to quantification and objective expert evaluation is sometimes highly uncertain or non-existent.

There are general standards for measuring the usability of software and these standards may also be used in the evaluation of SMDSs. For example ISO 9241-11 on usability defines usability as [http://www.tau-web.de/hci/space/i7.html]:

> "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"

Effectiveness, efficiency and satisfaction are then defined as follows:

> *Effectiveness:* The accuracy and completeness with which specified users can achieve specified goals in particular environments.
>
> *Efficiency:* The resources expended in relation to the accuracy and completeness of goals achieved.
>
> *Satisfaction:* The comfort and acceptability of the work system to its users and other people affected by its use.

Effectiveness is often measured as task completion rate, efficiency is measured as time to task completion while satisfaction data are obtained via a questionnaire given to users. However, these three measures are very overall and several parameters contribute to each of them. Therefore, many other criteria or parameters should be considered for use as well when evaluating the usability of a system. For example, at http://www.tau-web.de/hci/space/x12.html usability is also defined as having to do with:

> *Learnability:* The ease with which new users can begin effective interaction and achieve maximal performance, including predictability, synthesizability, familiarity, generalizability, and consistency.
>
> *Flexibility:* The multiplicity of ways in which the user and system exchange information, including dialogue initiative, multi-threading, task migratability, substitutivity, and customizability.
>
> *Robustness:* The level of support provided to the user in determining successful achievement and assessment of goals, including observability, recoverability, responsiveness, and task conformance.

However, there are several definitions of what usability exactly is and correspondingly of how to measure it. A second (and newer) ISO definition (ISO/IEC 9126-1 on product quality), cf. http://www.hostserver150.com/usabilit/tools/r_international.htm, is:

> *Usability:* the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions.

Thereby parameters such as attractiveness, understandability, and operability are introduced as issues that should be considered in the evaluation of usability.

Looking at the SMDS literature, many other criteria than those already mentioned are also used, including, e.g., modality appropriateness, adequacy (of, e.g., modality understanding, output phrasing, output representation, error handling, feedback, and emotion expression), quality (of output such as voice, graphics, and animation), naturalness (of, e.g., interaction and embodied agent), ease of use (of, e.g., system and devices), frequency of interaction problems, sufficiency (of, e.g., domain coverage, reasoning capabilities, and user modelling), task success rate, error correction rate, and many more. These should be considered for use as well when evaluating the usability of a system.

One major problem with criteria is to make the right selection for a given usability test of a given system. For an evaluator it is important to know the whole range of criteria which may influence usability in order to make a proper selection of criteria for his/her specific evaluation purpose. However, it is difficult to make an exhaustive list. One reason is that people sometimes use different terms for more or less the same issue and the same term is not always equivalent with one criterion measured in one and the same way. Another reason is that we probably don't know them all.

A second major problem is that many usability criteria are vaguely defined which makes it hard to evaluate a system with respect to these criteria because it becomes difficult to determine if the result was successful. For example, one may ask what is "adequate output phrasing"? How does one determine, when looking at test results, if output phrasing was adequate or not? This problem is not made easier with new system types that require new criteria. Such new criteria also require a solid definition.

2.3 Future challenges in usability evaluation of spoken multimodal dialogue systems

Today's challenges for usability evaluation of SMDSs relate both to how to carry out an evaluation (method) and to the choice of what to evaluate, i.e., which criteria to use and possibly how to define these criteria if they are new, and to figuring out what the evaluation results actually tells about the usability of a given system.

2.3.1 Challenges in usability evaluation methods

The methods mentioned in Section 2.2.1 are very general but nevertheless well-suited for SMDS evaluation even though they are not tailored to SMDSs. The most specialised method of those mentioned is perhaps the Wizard-of-Oz method which is very well suited when the speech modality is involved. Many usability evaluation methods require testing with users followed by data analysis. Such methods are resource heavy but they also provide input which it may be impossible to obtain from other evaluation types, such as an expert evaluation. On the other hand, e.g., an expert review or a heuristic evaluation may provide useful information which one would not get from a test with users. In other words, no existing usability evaluation method is able to capture all aspects of usability. Thus it is normally preferable to apply more than one method in the course of the software development life cycle.

Tailored methods may be said to exist in the form of guidelines which provide detailed suggestions for what (not) to do in a particular kind of system interface, e.g., in web interfaces. Theory-based methods may also provide a tailored approach. In the area of SMDSs, evaluation based on properties from modality theory [Bernsen 2002] has been suggested [Elting et al. 2002] but is far from having been fully explored. Future challenges consist both in pursuing approaches like modality theory-based evaluation further and to look into new theories which may form the basis for new theory-based or guideline-based evaluation approaches. Theory-based and guideline-based methods are cheap to apply because they don't require user interaction. On the other hand, a theory-based method can only provide feedback within the scope of what the theory covers. Thus, modality theory-based evaluation would give input on the use of modalities and modality combinations but not, e.g., on how prompts should be phrased or which colours and font types to use on the screen.

2.3.2 Challenges in usability evaluation criteria

Measuring usability is difficult because we don't know exactly how to measure it. At the end of the day, it is the users who decide if they find a system more or less usable, but as developers and vendors we would of course like to be able to predict the users' reactions with a high probability. However, some basic problems are that we don't know exactly which parameters makes a user happy – maybe it even differs across users - we don't know what the contribution of each parameter is to user satisfaction, some parameters we don't know how to measure apart from asking users, and for some new or upcoming application types we may not know for sure what it is important to measure. We may even have to redefine old criteria and invent new ones.

Very many usability evaluation criteria are of a qualitative nature. Most of these may be measured objectively as well as subjectively. However, when measured in both ways, results may often differ. What may be optimal from an objective point of view is not necessarily identical to what all users find optimal when asked. Users often have different *preferences* which, e.g., may be due to old habits or resistance to learn something new. Thus, user preferences are important and may be a key parameter in the measurement of usability.

There are not yet many results that show what happens to users' perception of usability *over time*. Truly, people may get accustomed to basically anything, but this does not mean that they become happy with the usability of a system. They may just use a system because there is nothing better or because they are required to use it. However, there is no doubt that time may change users' perception of a system to the better or to the worse. Their preferences for, e.g., modalities may change.

In some recent SMDSs, parameters such as *educational value* and *entertainment value* have become of interest. To measure these parameters one may of course ask the user but one may also look at, e.g., *learning effect* over time (for educational value) or *time spent* with a system (entertainment value), although one should be careful here because much time spent does not necessarily imply that the users had a good time. Maybe they just struggled to solve a problem.

Task success rate is a quantitative measure which has been used for many years in task-oriented systems. However, even for task-oriented systems, the definition of how to measure task success rate is not straightforward, cf. the discussion in [Dybkjær and Dybkjær 2004]. The problem lies in what to count as a task. When we move away from task-oriented systems and towards more free conversational systems without any particular task we may define a parameter called *conversation success rate*. No doubt this parameter is important but there is no good definition yet of how to measure it.

2.4 Conclusion

We have briefly outlined and evaluated current practice in usability evaluation of SMDSs and we have discussed new challenges.

When working towards an evaluation framework which is among the next steps to pursue, our proposal will be to do what we can to outline possibilities explicitly. This includes, among other things, to explain at least the most commonly used usability evaluation methods, including when to use them and what their advantages and disadvantages are. We will also describe a multitude of evaluation criteria, possibly grouped according to some categorisation. Each criterion should be explained and possible ways in which to use the criterion in actual evaluation should be proposed. Examples of use may also be included.

Furthermore, it should be kept in mind that a framework should be open for continuous changes and additions in a systematic way.

3 Usability evaluation issues in vision-based systems

3.1 Introduction

Currently, the development of human-computer interfaces which enable a more natural communication mode for human beings is a very active area of research [Carroll 2002].

People naturally communicate through gestures, expressions, movements. Research work in natural interaction is to invent and create systems that understand these actions and engage people in a dialogue, while allowing them to interact naturally with each other and the environment. People shouldn't need to wear any device or learn any instruction, interaction is intuitive. Natural interfaces follow new paradigms in order to respect human perception [Valli 2004].

Different techniques are emerging in order to create new natural (and thus non invasive) communication approaches with the machine world: whole body gesture based, point-at gesture based and facial based. They all have in common the fact that the innovation and attractiveness of this new way of interacting makes up with a certain lack of precision that make them still unsuitable for sensitive areas (i.e. medicine, engineering…). For the time being, there is still an inevitable trade-off between precision and natural interaction to be made.

Human-Computer Interaction (HCI) in three dimensions is not well understood, and there are few 3D applications in common use. Moreover, the complications of 3D interaction are magnified in immersive virtual environment (VE) applications: characteristics such as inaccurate tracking and lack of access to traditional input devices cause the design of user interfaces (UIs) and interaction techniques (ITs) for immersive VEs to be extremely difficult. Despite these difficulties, we maintain that there are complex applications for which immersive VEs are desirable, so special attention needs to be paid to the design and implementation of ITs for these applications.

3.2 Current practice in vision-based systems

The goal of gesture recognition is not to measure metrical parameters of a motion, but to recognize the intention that the action signifies. The same action may have different meanings in different contexts. To make machines able to recognize purposeful motor activities, processing algorithms need to deal with the great variety of shapes and styles a gesture can assume.

Since Myron Krueger's visionary work in the mid-1970s [Krueger 1983], this research area has been mainly aimed at a whole new branch in the interactive world: virtual immersion and mixed reality, and thus to the fields of the multimedia arts, entertainment and edutainment industry. These are three areas where precision is not a priority, and where visual feedback can be used as a very good backup in order to make this approximativeness go virtually unnoticed. Furthermore, end-user satisfaction is very much linked to how the application is comparable to previous similar ones. All these factors combined, and since this realm of vision-based applications is completely novel, we can say that user's satisfaction is often met, even when using simple techniques.

The Eye Toy © [Sony 2002. http://www.eyetoy.com] is the most straightforward example. With its very simple technique: frame to frame comparison, it has been able to achieve a massive success combining it with creative and intuitive content (this technique requires only

10 percent of the PlayStation 2's processing power, leaving a hefty 90 percent to render all the other graphic features). A vast majority of visual-based artistic and/or commercial displays don't need much more complex algorithms (other than their own sensibility and creativity) in order to achieve stunning effects [http://www.reactrix.com], [http://www.playmotion.com].

Only in the next stage, the one that uses human feature extraction, is some kind of tracking possible. At that stage, exact human features are robustly detected and tracked [Correa et al. 2004, Umeda et al. 2004], enabling more evolved and demanding applications, such as three-dimensional navigation and more realistic immersive environments.

The main challenge with whole body feature extraction is that new tracking methods are needed. Indeed, this kind of motion correspondence problem needs to match different points (often five different points) that can have very irregular trajectories, and that are very dependent, thus generating frequent auto-occlusions and/or fusions. Some probabilistic methods have been needed at this point to tackle this problem [Cox 1993].

The following sections are largely inspired by Doug A. Bowman's Thesis: *Interaction Techniques for Common Tasks in Immersive Virtual Environments: Design, Evaluation, and Application.*

3.3 Evaluation issues in vision-based systems

The goal of a usability methodology is to test the performance of vision-based systems. But what is performance? We could focus almost exclusively on speed, or time for task completion. Speed is easy to measure, is a quantitative determination, and is almost always the primary consideration when evaluating a new processor design, peripheral, or algorithm.

Another performance measure that might be important is accuracy, which is similar to speed in that it is simple to measure and is quantitative. But in human-computer interaction, we also want to consider more abstract performance values, such as ease of use, ease of learning, and user comfort.

Indeed, more than any other computing paradigm, virtual environments involve the user – his senses and body – in the task. Thus, it is essential that we focus on user-centric performance measures. If a vision-based system does not make good use of the skills of the human being, or if it causes fatigue or discomfort, it will not provide overall usability despite its performance in other areas.

3.3.1 Range of evaluation methods

Research in Human-Computer Interaction (HCI) has introduced a wide range of interface evaluation techniques. Evaluators have a choice regarding the statistical validity of their tests, the number of users involved, the time and effort required, and the results they wish to achieve. In this research, we feel that many of these techniques are appropriate for various stages of evaluation.

Initially, we come to look at these interaction tasks and techniques with very little concrete information, except our experience with them in applications, and in a few cases the published evaluations of others. Our first goal is to establish a taxonomy and perform categorization, but this is difficult given limited information. Therefore, in many cases it is appropriate to perform some informal evaluation at the beginning to gain a base of understanding of both the task and techniques. This may take the form of a guideline-based evaluation, where one or more usability experts try the techniques and note obvious problems and successes. In many cases, since there are few guidelines or experts in this field to draw from, an informal user

study would be useful, in which a few users try out the techniques on some representative tasks, and their general performance and comments are noted. Finally, if the techniques have already been implemented as part of an application, a usability study with some quantitative measures may provide some good information.

3.3.2 Testbed evaluation

The experimental methods and other evaluation tools discussed above can be quite useful for gaining an initial understanding of interaction tasks and techniques, and for measuring the performance of various techniques in specific interaction scenarios. However, there are some problems associated with using these types of tests alone. First, while results from informal evaluations can be enlightening, they do not involve any quantitative information about the performance of interaction techniques.

Without statistical analysis, key features or problems in a technique may not be seen. Performance may also be dependent on the application or other implementation issues when usability studies are performed.

On the other hand, formal experimentation usually focuses very tightly on specific technique components and aspects of the interaction task. Techniques are not tested fully on all relevant aspects of an interaction task, and generally only one or two performance measures are used.

Finally, in most cases, traditional evaluation takes place only once and cannot truly be recreated later. Thus, when new techniques are proposed, it is difficult to compare their performance against those that have already been tested.

Therefore, we propose the use of *testbed evaluation* as the final stage in our analysis of interaction techniques for universal virtual environments interaction tasks. This method addresses the issues discussed above through the creation of testbeds, i.e. environments and tasks that involve all of the important aspects of a task, that test each component of a technique, that consider outside influences (factors other than the interaction technique) on performance, and that have multiple performance measures.

As an example, consider a proving ground for automobiles. In this special environment, cars are tested in cornering, braking, acceleration, and other tasks, over multiple types of terrain, and in various weather conditions. Many quantitative and qualitative results are tabulated, such as accuracy, distance, passenger comfort, and the “feel” of the steering.

The testbeds will allow us to analyze many different ITs in a wide range of situations, and with multiple performance measures. Testbeds are based on the formalized task and technique framework discussed earlier, so that the results are more generalizable. Finally, the environments and tasks are standardized, so that new techniques can be run through the appropriate testbed, given scores, and compared with other techniques that were previously tested.

3.3.3 Application of results

Testbed evaluation produces a set of results that characterize the performance of an interaction technique for the specified task. Performance is given in terms of multiple performance metrics, with respect to various levels of outside factors. These results become part of a performance database for the interaction task, with more information being added to the database each time a new technique is run through the testbed.

The last step in our methodology is to apply the performance results to VE applications, with the goal of making them more useful and usable. In order to choose interaction techniques for applications appropriately, we must understand the interaction requirements of the

application. We cannot simply declare one best technique, because the technique that is best for one application will not be optimal for another application with different requirements. For example, a VE training system will require a travel technique that maximizes the user's spatial awareness, but this application will not require a travel technique that maximizes point-to-point speed. On the other hand, in a battle planning system, speed of travel may be the most important requirement.

Therefore, applications need to specify their interaction requirements before the correct ITs can be chosen. This specification will be done in terms of the performance metrics which we have already defined as part of our formal framework..

3.4 Conclusions: Summary of methodology

For each universal interaction task, the process begins with informal evaluation techniques: observation, user studies, and/or usability evaluations. These should lead to an understanding of the task and the space of possible techniques, which allows us to create a taxonomy and to categorize existing and proposed ITs, and may also inspire the creation of new techniques. We can also list outside factors influencing performance and performance measures at this time. Once this formal framework is in place, we can perform more formal experiments, involving specific task and technique components and performance measures.

These results, along with our design framework, may lead to the design and implementation of novel techniques for the task. Also, experimentation may cause some reworking of the initial taxonomy. When the formal framework is judged complete, we can move to the final analysis step: testbed evaluation. Use of the testbed with a range of techniques and performance measures produces a dataset of results for the given task, which can then be used to make an informed choice of ITs for the target application(s), given their performance requirements.

In practice, these are some examples of aspects that can be taken into account in our testbeds:

- The Learning curve: Time needed for the user to perform a first useful action.
- Fatigue: number of interactions a user can conduct one after the other without being tired/fed-up.
- Friendliness: The ease of use of the particular interaction mode, its compatibility with the physical abilities of the user.
- Effectiveness: The accuracy and completeness with which specified users can achieve specified goals in particular environments.
- Efficiency: The resources expended in relation to the accuracy and completeness of goals achieved, notably overall time spent.

4 Usability evaluation issues in haptics-based systems

4.1 Introduction

The performance of a haptic interface depends on the mechanical characteristics of the haptic devices used and the control system architecture [Sjostrom2001a, Sjostorm2001b]. Currently there is a number of widely accepted specifications that haptic-based systems need to follow in order to be considered usable. Measurements [Burdea1996] have resulted in maximum exertable forces that can be controlled by a human and frequencies that can be perceived satisfactorily. It is crucial that haptic devices and the controlling software provide feedback in the range of the forces that can be controlled by the users and with update frequencies that can be perceived by them. These measures need to be taken into account so as to design and develop usable haptic interfaces. However, following the above-mentioned requirements does not ensure the creation of user-friendly haptic-based systems.

There are several methodologies used to perform usability evaluation of haptic-based systems (HS). Usability evaluation of HSs is based both on objective and subjective criteria. In some cases evaluation of a haptic system may vary significantly when different scenarios are considered, since haptic interaction applications involve in most cases 3D environments. Changing the scenario (i.e. the 3D environment) is similar to changing position and/or size of buttons in a GUI and thus changes the usability of the overall system. Additionally, in some cases, haptic-based systems are developed as simulations of real environments (either as training environments or as intermediate evaluation steps before creating real prototypes). In these cases, the usability evaluation of the haptic-based system does not aim to measure the ease of use of the system but how similar it is to its real replica.

In order to estimate the usability of existing HSs several procedures have been proposed. Currently, the evaluation procedures utilize a number of criteria to measure the user acceptance to specific haptic systems and to extract useful information for the added value, if any, produced by the use of haptics.

This section is organized as follows. Subsection 4.2 presents criteria and methodology currently used for the evaluation of haptic systems. Uncertainties and other issues that could be improved are discussed in Subsection 4.3. Finally, Subsection 4.4 aims to outline an evaluation framework for haptic-based systems as a conclusion to the presented study.

4.2 Current practice in haptics – based systems

Usability in haptic-based systems involves several issues such as anatomy, physiology, psychology and design. Usability evaluation is performed to ensure that products and environments are comfortable, safe and efficient for people to use, as well as to ensure that a product fits the target users' needs. This section aims to classify evaluation criteria currently used for the evaluation of HSs to four categories: the well-defined, standard, frequently used and "known to be useful" criteria.

Well-defined evaluation criteria: This category includes evaluation criteria, where the measured values can be defined accurately and relations between the values and the system usability are also well defined. The criteria might be application specific or not. Such criteria are based on the measurement of specific metrics, such as distance, angles and time. Using such metrics it is possible to measure position, shape and rotation accuracy, and timing. These criteria provide useful information on the usability of haptic-based environments and their

performance on specific actions or tasks, such as object size discrimination using haptic devices. Well-defined criteria can be used to estimate the added value provided by integrating haptic-based interaction into an existing system as well as to compare different haptic interfaces that aim to perform the same task. Usually this kind of evaluation takes place in the first steps of the implementation in order to allow designers and developers to proceed securely in the creation of user-friendly haptic-based systems. However, it is also important for the evaluation of complete systems, since it can provide specific results and measures for their usability [Feygin et al. 2002], [Emery2003], [WangMacKenzie 2000], [Burdea1996].

Standard evaluation criteria: This category includes evaluation criteria that are considered as standard in the evaluation of HSs but are not well-defined criteria. The criteria might be application specific or not.

Such criteria include psychophysical methods and questionnaires involving users before, during and after the evaluation test procedure in order to review what the users are expecting, doing and what is their impression of the examined system [Tzovaras 2004], [Sener et. al. 2002] [Esch-Bussemakers & Cremers 2004], [Yu2002].

The questionnaires usually include questions about the mental demand, physical demand, performance level achieved, fatigue, etc. The main drawback of this evaluation procedure is that it is not objective and requires a large user group to provide reliable results.

Frequently used evaluation criteria: This category includes evaluation criteria that are frequently used in a variety of HSs. This category is similar to the standard evaluation criteria category, however, criteria in this case are not yet accepted as standard in the area, but are widely used by many researchers.

Such criteria include measures of user error rates and retries (times a user is performing the same task to complete it successfully) [Kaster et. al.2003]

"Known-to-be useful" evaluation criteria: This category includes criteria that have been proved useful in cases of haptic-based systems. However, they are not widely accepted and usually are application or task specific.

The known-to-be useful criteria are used to identify specific advantages or disadvantages of a system. Indicatively time delay is measured in a distributed haptic system and evaluation is performed to estimate acceptable delays in the system [Shen et. al. 2004].

4.3 Evaluation issues in haptic – based systems

Usability evaluation of HSs using the criteria described provides information required to design and develop such systems. Although criteria have been proven useful in the past in many cases, they are not sufficient to measure the performance of the systems. This subsection aims to identify weak points of the currently used criteria and propose new in order to overcome the problems caused by them.

The use of specific metrics, such as **size discrimination**, for the usability evaluation of HSs allows identifying the functionalities that a system can support. However, these criteria focus only on specific measures and not to the end user preferences. Thus, although they are suitable to perform evaluation for HSs they do not allow the evaluation of the overall usability of the system since they do not record the users' preferences using direct input from them.

Psychophysical criteria, on the other hand, perform the evaluation of the system based on questionnaires filled by the users. This allows recording of users' preferences and recording of the overall acceptance of the system. However, this kind of evaluation is subjective and requires a large number of users in order to provide reliable results. Psychophysical methods used for usability evaluation are suitable for testing systems that are approaching a final form

but not for systems at an early development stage. The main reason is that during early development stages data relevant to the functionalities that the system supports are more important, while during final stages of development the way that supported functionalities should be used is more important.

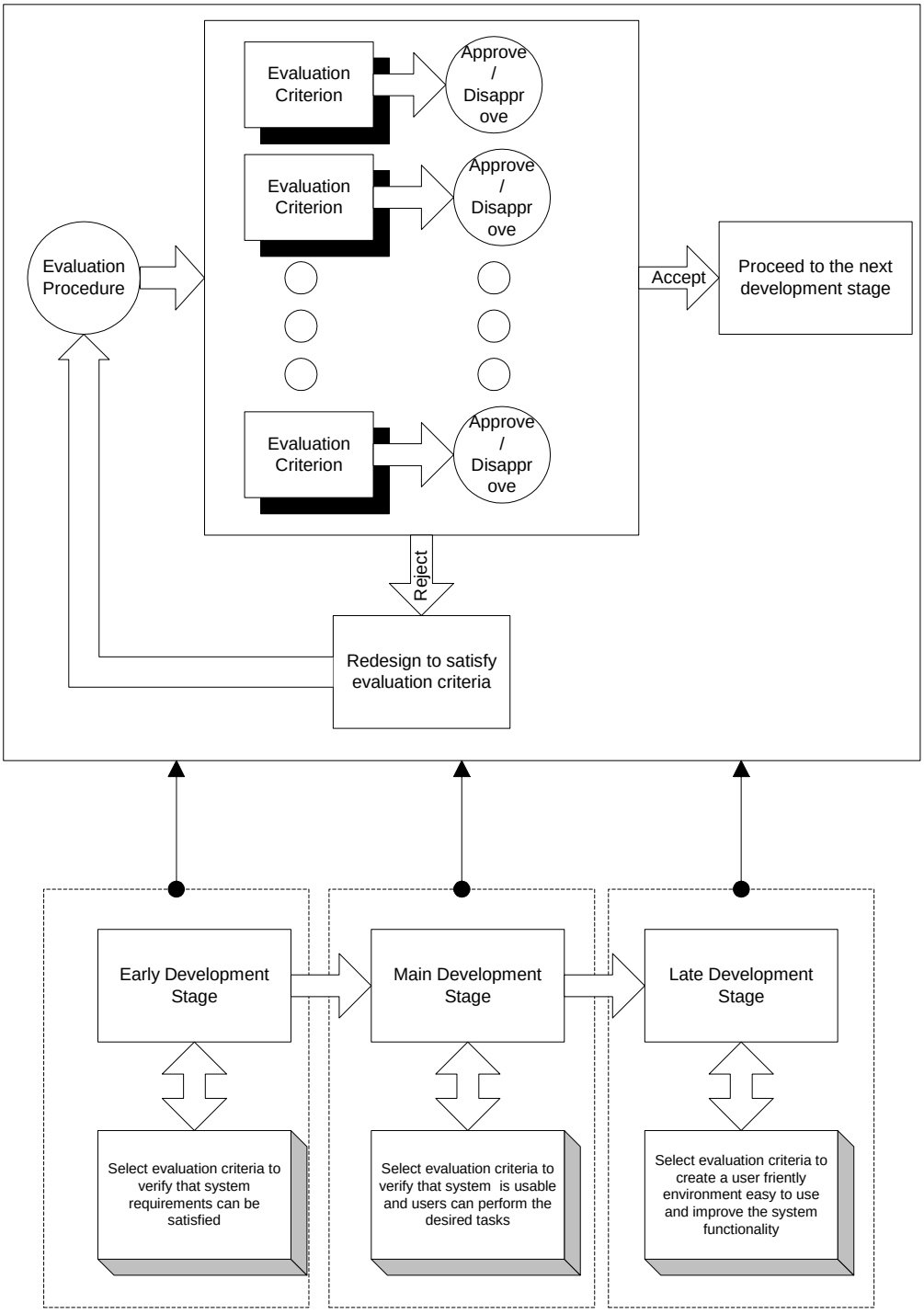

Figure 1. Usability evaluation procedure.

Measuring of users' error rate is another important criterion used for usability evaluation. Since most users are not familiar to haptics it is expected that they will make errors while using a HSs. This measure can be used to compare the usability of different interfaces that

aim to perform the same tasks as well as to identify how easy it is for users to get used to the interface of the HS.

Similar to that is the **measure of retries performed before a successful task completion**. Both criteria can be assumed objective when applied to a large number of end users. Since most users are not familiar to haptics, it is possible that error rates and retries may have relatively high values compared to the use of conventional user interfaces. Thus, in this case, it is important to study the evolution of the measures through time in order to get results concerning the usability of a system and how easy it is for users to get used with it.

The aforementioned weaknesses of the criteria described, prove that usability evaluation requires following a framework in order to avoid ambiguities in the results. The following diagram (Figure 1) proposes a general framework to perform usability evaluation of HSs. The selections of the evaluation criteria to be satisfied depend on the development stage of an application. However in order to proceed to a next development stage it is important to satisfy all the criteria selected for each stage.

4.4 Conclusions

In this section current practice and criteria in usability evaluation of HSs and identified weaknesses of the currently used criteria are described. Also methods and criteria used for the usability evaluation of HSs are classified in categories depending on their characteristics. Weak points are identified in each case and practices that can be used to overcome these weaknesses are proposed. Finally, an abstract framework is proposed in order to support the design of usability evaluation tests for HSs.

5 Usability evaluation issues in mixed reality systems in surgery

5.1 Introduction

With the advent of mixed reality systems into many surgical and training specialties interactions based on traditional input and output devices are not effective in a mixed scenario as it distracts the user from the task at hand and may create a severe cognitive seam. Having multiple sources of information and two worlds of interaction (real and virtual) involves making choices about what to attend to and when. New interaction paradigms and visualization techniques centred on the user's task focus need to be investigated and evaluated.

On the other hand VR simulators are enhancing surgical training and enabling new ways for skills assessment. The key question is how these skills are transferred from the virtual world to the operating room, and how the different trainees and experts in surgical training believe this transfer occurs. The influence of fidelity, augmented virtuality capabilities and other simulation resources needs to be investigated for optimal simulation design.

5.2 Augmented reality systems

5.2.1 Current practice in image-guided systems

In D16 we have presented validation criteria for Image-guided surgery systems. Such validation criteria are related to a device or a process and involve the following criteria: accuracy, precision, robustness, consistency, fault detection, computation time and functional complexity.

In fact just validating these criteria we will not address problems found during man-computer interaction and integration in the clinical context. Such issues will be discussed as follows.

5.2.2 Evaluation issues in augmented reality systems

(Case study: Image-guided application described in D17)

Besides the validation criteria that are applied to a device or a process we should take into account the usability criteria to assess the interaction technique of this new paradigm of interaction.

At most basic level a multimodal-augmented reality system should contain at least four components:

1. Sensors, for determining user, platform or environment state (for instance how to perceive the real world)
2. Inference engine or classifier to evaluate incoming sensor information (for instance registration techniques to align virtual information to the real scene)
3. Adaptive user interface (for instance visualization of the augmented scene according to the user's view)
4. Underlying computational architecture to integrate these components.

In reality a fully functioning system would have many more components, but these are the most critical for inclusion as an augmented interaction system. Independently, each of these components is fairly straightforward. Much of the ongoing augmented interaction research

focuses on integrating these components to ”close the loop,” and create computational systems that adapt to their users, tasks or environment.

As has been discussed in many previous works, humans have well documented limitations in attention, memory, learning, comprehension, sensory bandwidth, visualization abilities, qualitative judgments, serial processing and decision making. For an augmented system to be successful it must identify at least one of these bottlenecks in real time and alleviate it through a performance enhancing mitigation strategy. These mitigation strategies are conveyed to the user through the adaptive interface and might involve: modality switching (between visual, auditory, & haptic), intelligent interruption, task negotiation and scheduling, and assisted context retrieval via book marking. When a user state is correctly sensed, an appropriate strategy chosen to alleviate the bottleneck, the interface adapted to carry out the strategy and the resulting sensor information indicates that the aiding has worked – only then has a system “closed the loop” and successfully augmented the user’s interaction.

In this sense our first proposed framework should be able to evaluate the interaction in terms of:

How, where and when the information will be delivered to the user (e.g. perceptual properties) will reflect how easy or hard the information will be interpreted by the user (e.g. cognitive properties) and how natural the interaction process (e.g. functional properties) will be.

5.2.3 Designing for continuous interaction

We have proposed a first evaluation framework entitled: “Designing for continuous interaction”. This framework takes into account all properties related to the development of an augmented reality system to support continuous interaction.

Continuous interaction is the capability of the system to promote a smooth interaction scheme during task accomplishment considering perceptual, cognitive and functional properties.

Perceptual property

The perceptual property of continuity is defined as an ability of the system to make all data involved in the user’s task available in one perceptual environment in order to avoid changes in the user’s focus. The following design aspects may have influence while evaluating the perceptual property:

- *User’s interaction focus*: is sorted by degree of realityvirtuality. The user could be either performing a task in order to manipulate or modify an object in the real world (when the task focus is on the real world), or an object in the virtual world (when the task focus is on the virtual world).
- *Insertion context*: it is sorted according to the distance at which each device displaying the interaction space is inserted in the environment relative to the user’s position and the user’s task focus.
- *Depth Cues*: involves perceptive issues such as accommodation, convergence, binocular parallax, motion parallax, occlusion, shades, shadows, perspective, colours and brightness, tactile sense, texture, …
- *Spatio-temporal links*: fusion mechanisms related to the spatial arrangement and temporal synchronisation.

Cognitive property

The cognitive property is defined as an ability of the system to ensure that the user will correctly interpret the perceived information and that is correct with regards to the internal

state of the system. The following design aspects may have influence while evaluating the cognitive property:

- *Language*: defined by device + modality chosen. It is related to the sensory channels used to interpret the information. The modality is sorted by level of complexity and dimensionality, starting with basic modalities such as text (1D) and image (2D) and finishing with more complex and structured modality type such as those found in 3D animation and immersive environments.
- *Consistency*: close to or far from the real concept.
- *Spatio-temporal links*: fusion mechanisms related to the spatial arrangement and temporal synchronisation.

Functional property

The functional property is related to the effort of the user in experiencing a new interaction mode. This property is quite related to the functional complexity criteria defined in D16 where:

> Functional complexity concerns the steps that are time-consuming or cumbersome for the operator. It deals both with man-computer interaction and integration in the clinical context and has a relationship with physician acceptance of the system or method. The degree of automation of a method is an important aspect of functional complexity (manual, semi automatic or automatic).

We have identified two different levels of functional property: interaction and task levels.

- *Interaction level:* it is related to the interaction complexity and it involved the following design aspects:
  - *Connection type*: it is sorted by level of complexity in registering information. Environments with static links (i.e. where links between the real and virtual world are established during design time) are considerably less complex than environments in which all links are established during execution time.
  - *Transform type*: is arranged according to users' level of familiarity with the tuple action/effect. Thus real action with real effect is highly familiar, while virtual action with virtual effect is highly unfamiliar.
  - *Adaptation type*: open (need another process to properly adapt) or self-adaptive.
- *Task level* – it is related to the continuity of the task trough the different contexts and it involves the following design aspects:
  - *Temporal links*: intra-context task continuity
  - *Saving/Recovering task content and context*: inter-context task continuity.

5.3 Surgical VR simulators

Surgical simulators have nowadays two main goals: training and skill assessment. Technology and scientific knowledge are still immature for other applications like mission rehearsal or surgical credential. Usability is then understood as the capability of the simulator to achieve training or skill assessment with effectiveness, efficiency and satisfaction, a concept very close to validation.

5.3.1 Current practice in VR surgical simulators

The main validation tool for surgical simulators is the study that proofs how surgical skills are transferred from VR to the operating room. It is done with prospective, randomized and blinded surgical trials where novice surgeons are trained in different ways. In the field of laparoscopic surgery, the MIST-VR simulator has been recently validated with this kind of studies [Seymour et al. 2002; Grantcharov et al. 2004], what has been considered as a landmark [Fried 2004]. On the other hand acceptance of trainees and experts in surgical training is usually assessed by face validity studies [Schijven and Jakimowicz 2002], whose main drawback is its subjectivity. More details about current practice are explained in D16.

5.3.2 Evaluation issues in VR surgical simulations

Tension often exists between the design and evaluation of surgical simulations. A lack of high quality published data is compounded by the difficulties of conducting longitudinal studies in such a fast-moving field [Kneebone 2003]. Although the evidence for the inherent validity and reliability of several simulators is satisfactory, evidence for their ability to predict future operating room performance is lacking. Some common problems with the studies include the lack of universally agreed metrics, the lack of a ‘‘gold standard’’ for operating room performance, the variety of simulators used with differing levels of validity and reliability, the differing skill levels of the trial participants, and the small sample sizes seen to date [Feldman et al. 2004].

One of the most controversial dilemmas in simulation design is the incorporation of force feedback (FF). Trocar friction can hide tactile information [Picod et al. 2005], but perception is enhanced with FF both in grasping [Tholey et al. 2005] and pulling [Lamata et al. 2005a] manoeuvres. More studies comparing the effectiveness of simulation with and without FF [Kim et al. 2003] are needed for assess its importance for training transfer.

On the other hand acceptance of simulators is very influenced by its fidelity. Surgeons have a very high expectation about what VR represents. They imagine a system that emulates perfectly the behaviour of a human body. There is a need of a shift from this conception to what VR technologies can really offer, with their strengths and limitations.

Summarising there is a need of identifying which individual resources and their combinations available in VR simulation technologies are most important for laparoscopic training. The next section presents a conceptual framework for the analysis, design and evaluation of surgical simulators which offers guidelines to formulate and contrast hypotheses about the effectiveness, efficiency and satisfaction of surgical simulation.

5.3.3 Framework of simulation resources

We propose a taxonomy for the different resources available in VR simulation, which is considered as a didactic means to meet different training needs, using several didactic resources. Basically these resources are defined and classified in three main categories: Fidelity, Virtual and Evaluation resources. Fidelity resources refer to the different levels of realism offered by a simulator in its interaction and behaviour. They can be further divided into sensorial, mechanical and physiological. Computer resources are features unique to a computer simulated environment that can enhance training, like cues and instructions given to the user to guide a task, or to manage a training program. Evaluation resources are metrics to evaluate performance, follow up progress and ways to deliver constructive feedback to the user.

This taxonomy is a first step to assess the relationship between simulation design and training effectiveness. Our taxonomy reveals in detail how the different didactic resources are used by different commercial simulators [Lamata et al. 2005b]. Future research will concentrate on a thorough evaluation of the importance of different didactic resources to teach basic and more advanced laparoscopic skills. Randomized and blinded surgical trials where novice surgeons are trained in different ways is the best methodological approach. [Lamata et al. 2005].

5.4 Conclusions

As conclusion of the analysis of augmented reality systems we have proposed a first evaluation framework entitled: “Designing for continuous interaction”. This framework takes into account all properties related to the development of an augmented reality system to support continuous interaction. On the other hand, didactic resources offered by surgical simulators have been categorized as a framework to formulate and contrast hypotheses about the effectiveness, efficiency and satisfaction of surgical simulation.

6 Issues in tools for remote usability evaluation

6.1 Introduction

In remote usability evaluation users and evaluators are separated in time and/or space. There are many reasons for remote evaluations. Often evaluators have not available a usability laboratory because it is expensive and requires the user availability to move to it for the tests. In addition, the evaluation can provide more meaningful results if users interact with the application in their daily environment.

Usability evaluation is an increasingly important part of the user interface design process, but, it can be expensive in terms of time and human resources, and automation is therefore a promising way to augment existing approaches which are using more and more remote testing techniques, made more and more affordable also by the evolving technology. Web applications are one example of applications that can benefit from remote evaluation. While a Web site can easily be developed using one of the many tools available able to generate HTML from various types of specifications, obtaining usable Web sites is still difficult. Indeed, when users navigate through the Web they often encounter problems in finding the desired information or performing the desired task. With over 30 million Web sites in existence, Web sites have become the most prevalent and varied form of human-computer interface. At the same time, with so many Web pages being designed and maintained, there will never be a sufficient number of professionals to adequately address usability issues without automation [Ivory and Hearst 2001] as a critical component of their approach.

In the next sections we provide an overview of remote usability evaluation techniques.

6.2 Current practice in remote usability evaluation

With the refinement of instrumentation and monitoring tools, user interactions are being captured on a much larger scale than ever before. In order to obtain meaningful evaluation it is important that users interact with the application in their daily environment. Since it is impractical to have evaluators directly observe users' interactions, interest in remote evaluation has been increasing.

In [Castillo et al. 1998] the authors introduce the user-reported critical incident method (originally called semi-instrumented critical incident gathering) for remote usability evaluation, and describe results and lessons learned in its development and use. The findings indicate that users can, in fact, identify and report their own critical incidents, which reveal useful for the evaluation.

Some work has highlighted that through logging keystrokes and web pages on a given site, we could infer patterns of user behaviour that indicate usability problems or other design deficiencies. This possibility has obvious attractions for web designers, but in the HCI usability research it has been argued that it is not possible to identify usability problems without access to the use context, to the users tasks and goals and to the user's own reports of what counts as a problem for them.

Among other methods for remote usability evaluation we can also cite remote questionnaire or surveys.

In the paper of Ivory and Hearst [2001] authors present an extensible survey of usability evaluation methods, organized according to a new taxonomy that emphasizes the role of automation. The survey analyzes existing techniques, identifies which aspects of usability

evaluation automation are likely to be of use in future research, and suggests new ways to expand existing approaches to better support usability evaluation. From this paper, within the usability testing class, automated capture of usage data is supported by two method types: performance measurement and remote testing. Remote testing methods enable testing between a tester and participant who are not co-located: the evaluator is not able to observe the participant directly, but can gather data about the process over a computer network. Remote testing methods are distinguished according to whether a tester observes the participant during testing.

Same-time different-place and different-time different-place are two major remote testing approaches [Hartson et al. 1996]. In same-time different-place or remote control testing the tester observes the participant's screen through network transmissions and may be able to hear what the participant says via a speaker telephone or a microphone affixed to the computer. Software makes it possible for the tester to interact with the participant during the test, which is essential for techniques such as the question-asking or thinking-aloud protocols that require such interaction. The tester does not observe the participant during different-time different-place testing. An example of this approach is the journaled session [Nielsen 1993], in which software guides the participant through a testing session and logs the results.

Remote testing approaches allow for wider testing than traditional methods, but evaluators may experience technical difficulties with hardware and/or software components (e.g., inability to correctly configure monitoring software or network failures). This can be especially problematic for same-time different-place testing where the tester needs to observe the participant during testing. Most techniques also have restrictions on the types of UIs to which they can be applied. This is mainly determined by the underlying hardware (e.g., PC Anywhere only operates on PC platforms) [Hartson et al. 1996].

As far as Web UIs are concerned, the Web enables remote testing and performance measurement on a much larger scale than is feasible with WIMP interfaces. There are three solutions to log user interactions: server-side, proxy, client-side. Similar to journaled sessions, Web servers maintain access logs and automatically generate a log file entry for each request. However, server logs cannot record user interactions that occur only on the client side (e.g., use of within page anchor links or back button), and the validity of server log data is questionable due to caching by proxy servers and browsers. Client-side logs capture more accurate, comprehensive usage data than server-side logs because they allow all browser events to be recorded. Such logging may provide more insight about usability. On the downside, it requires every Web page to be modified to log usage data, or else use of an instrumented browser or special proxy server. In WebRemUsine a solution to ease such modification has been applied by just automatically including a JavScript in all the pages that have to be evaluated. Using these client-side data, the evaluator can accurately measure time spent on tasks or particular pages as well as study use of the back button and user clickstreams. Proxy-server based solutions are even less intrusive and not require any modification in the Web application to evaluate but they limit their analysis to the page accessed and are not able to capture the local user interactions.

In addition, due to the increasing diffusion of mobile devices, it has been put more and more attention to the need of remotely testing UI for mobile devices. For instance, in the paper [Waterson et al. 2002] the authors discuss a pilot usability study using wireless Internet-enabled personal digital assistants (PDAs), in which they compare usability data gathered in traditional lab studies with a proxy-based clickstream logging and analysis tool, and found that this remote testing technique can more easily gather many of the content-related usability issues, but device-related issues are more difficult to capture.

More recent studies [Tullis et al. 2002] have confirmed the validity of remote evaluation in the field of Web site usability. Some work [Lister 2003] in this area has been oriented to using audio and video capture for qualitative usability testing. In [Paganelli and Paternò 2003] authors provide more quantitative data for supporting their analysis. In this paper they present a tool for performing remote usability evaluation of Web applications without requiring expensive equipment. Indeed, the tool is able to automatically analyse the information contained in Web browser logs and compare it with task models specifying the designer model of the possible users behaviours when interacting with the application in order to identify whether and where users interactions deviate from those envisioned by the system design and represented in the model. An improved version of WebRemUsine has been developed [Paternò, Piruzza, Santoro, 2005] in order to consider also multimodal information regarding the users interactions with the system obtained through browser logs, eye-tracking, and WebCams.

6.3 Challenges in remote usability evaluation

As we have seen in previous sections, several methods have been developed for conducting remote usability evaluation, but each suffers from some drawback - e.g., time-consuming data capture, costly data analysis, inapplicability to real users doing real tasks in their normal work environment, or need for direct interaction between user and evaluator during an evaluation session. Future challenges in remote usability evaluation methods should try to develop a cost-effective method for remotely evaluating usability of real-world applications that overcomes these drawbacks, trying to maximise the information gathered on the users and the associated context, so as to interpret better user actions, which is really critical for remote evaluation due to the fact that the evaluator may be distant in space and time from the users. Another challenge is to minimise the effort for performing the evaluation, both in terms of hardware/software required and their configuration, maximising the flexibility of the automatic tools used for the evaluation, which should automatically identify the best options to be configured depending on the current user's context of use, network capabilities/limitations etc. Such tools should be able to identify the most appropriate techniques to be applied especially when more and more natural interactions are expected by the users, then different devices and modalities (and often also combinations of) might be concurrently used. For instance, the tools should be able to integrate and appropriately weight the data collected so as to order to better identify the context of use and better interpret the user's actions. To give an example, in this situation it is clear that logging pen/key-strokes for mobile devices might give information useful, but, taking into account that the user might be on the go, even more useful should be to record the surrounding environment, which might affect the mobile users' actions more than the user at the desktop system.

6.3.1 Remote evaluation for multi-device user interfaces

The growing diffusion of devices has opened a lot of issues for evaluating multi-device user interfaces.

Denis and Karsenty [Denis and Karsenty 2003] focus on the usability of a multi-device system and introduce the concept of inter-usability to designate the ease with which the users can reuse their knowledge and skills for a given functionality when switching to other devices. In this paper a framework for achieving inter-usability between devices is proposed. It is based on two components: (i) a theoretical analysis of the cognitive processes underlying device transitions, and (ii) an exploratory empirical study of the problems in using functionalities across multiple devices.

Another issue to be considered for remote evaluation of multi-device user interfaces is mobility. Different factors impact the user experience when users are mobile using their PDA to access the web. Some of these factors are external to the experience, such as noise, distractions and movement and they might have an impact on the user. Effective remote usability evaluation techniques should be able to gather information about the possibly changing contexts. In addition, depending on the remote usability evaluation technique used, the information should be customised depending on the specific device/modality used. For instance, if a logging tools is used to record the users' actions, different information should be provided depending on the different device used, and such information should be accompanied with information about the current context of use, so as to better interpret the recorded information, especially because with mobile devices the contexts of use may extensively vary.

6.3.2 Remote evaluation for migratory user interfaces

As far as criteria for evaluation of migratory user interfaces, it is important to note that a user's familiarity with a web page is important from the point of view of the usability. When a user first uses a web page, they establish a mental model of the page based on the structural organization of the information, such as visual cues, layout and semantics [Albers and Kim 2000; Danielson 2003; Spence 2001]. A primary objective when transforming a web page for different devices is to minimize the user effort in re-establishing the existing mental model of the original page. Danielson [2003] introduced the concept of transitional volatility and described two ways the web is volatile: web sites can change over time and within sites users can experience different navigation structures. Danielson [2003] found that a highly volatile session increased disorientation and decreased user navigation abilities. When users switch between devices to use the same web page, this introduces a new type of volatility: transformation volatility [Watters & MacKay 2004]. Transformation volatility results from changes to the look, design, layout and even content when using the same web page on different devices. Transformation volatility is a measure of change to navigation, layout, content and readability from one device to another. When a user accesses a web page on a desktop and uses the same web page on their laptop, the transformation volatility is small. But when the user uses the same web page on their PDA the transformation volatility is substantial, so, evaluating the impact of such switch might be useful for evaluating how previous experience (gathered from the version which the user feels familiar with) might be re-used in a different one, also evaluating how much the design respects a similar design through the different versions. Then, the evaluation should consider the effectiveness of adapting analogies found in systems with which users are already familiar (the presence of a common coherent framework of the pages through the different devices, should enable users to re-use their previous knowledge even with they visit a site through a new device).

Another dimension that should be evaluated with migratory user interfaces is the interaction continuity of the user interface when the user switch devices, which is at the basis of the theory of migratory user interfaces. The extent to which the change between the devices is transparent to the user should be evaluated, together with the degree of user's awareness of the fact that the migration process is occurring.

6.3.3 Remote evaluation of multimodal information regarding the user behaviour

Remote usability evaluation should also consider multimodal user interfaces. For instance, in (MultiModal WebRemUsine) authors discuss what information can be provided by automatic tools able to process multimodal information on users gathered from different sources, so as

to provide the most effective remote usability evaluation of websites. The collected information ranges from browser logs to videos to eye-tracking data, and the approach proposed tries to integrate such data in order to derive the most complete information for analysing, interpreting and evaluating the users while visiting a website, by taking into account the factors that might affect the performance of the users. The proposed approach is supported by a tool – MultiModal WebRemUsine, which has been improved over the years in order to include and handle more and more information and provide additional features. In one of its first versions [Paganelli and Paternò 2003], the tool was just able to automatically analyse the information contained in Web browser logs and compare it with task models specifying the ideal behaviour of users interacting with the application and representing the actual Web site design in order to identify where users interactions deviate from those envisioned by the system design and represented in the model. However, such information revealed soon rather limited because when users visit a webpage, their attention can be captured by different areas of the same page and this information cannot be derived just analysing log files, which are only able to track just physical interactions of the user with the application (e.g. scrolling, clicking, etc.). Eye tracking is a technique able to allow for deriving the current area of interest of the user by following the user's gaze. Thus, it helps evaluators in discovering the navigation strategies of the users visiting the web site and analysing the impact of different areas of the page. This enables easy identification of possible problematic parts of the page. However, there are situations in which even the eye-tracker data may result inadequate to provide sufficient information for effective evaluation. Indeed, a user may look at the same portion of the page for quite different reasons, and such reasons could not be discovered by just analysing eye-tracking information. For instance, users might delay in staring at a certain point of the page because they might not be aware of having found the requested information and still look for it or, alternatively, they are aware of having achieved the goal and thus are interested in further reading the information found. In both cases, a video-based analysis might provide useful information for interpreting the different impacts of the same page on users: for instance, it might highlight situations where, although the logged information and user's gaze might make evaluators conclude that the user has successfully completed the expected task, a puzzled expression of the user should force the evaluator to re-interpret the collected data and derive, more realistically, that the user has completed the task but might not realised it, which is still a signal of a usability problem in the page. So, this simple example shows how important is integrating all the data that is possible to capture on user (and possibly also on the environmental/contextual conditions in which the user interaction takes place) in order to perform the most comprehensive evaluation.

Moreover, it is worth pointing out that, apart from the data provided by the eye-tracker, which still remains a rather expensive technology, the approach proposed has the remarkable advantage to allow evaluators to identify usability problems even if the analysis is performed remotely, which might contribute to keep at minimum the evaluation costs and allows the users to remain in their familiar environments during the evaluation, improving the trustworthiness of the evaluation itself.

6.3.4 Remote evaluation tools

It is worth pointing out that the previous criteria are basically related to the specific types of user interfaces (multidevice, multimodal, etc.) considered. Combinations of the issues related to such different user interfaces should also be considered (e.g. evaluate if the appropriate modality has been used for carrying out a certain task on a certain device).

However, as far as the tools for remote evaluation are concerned, a future challenge for automatic tools for performing remote usability evaluation is represented by the assessment of how easy to configure, use and learn are the tools themselves. Indeed, some approaches might have some limitations in terms of hardware and software components (such as video recorders and logging software), while other might have additional requirements on the preparation necessary to be actually ready to use them.

6.4 Conclusions

We have briefly outlined the state of art in remote usability evaluation and some current challenges for this type of approach. In particular, the possibility of remotely evaluating mobile or multi-device applications is important along with the ability of considering many sources of information regarding the user and the surrounding environment. This includes also the possibility of detecting the emotional state of the user, which can heavily affect how tasks are accomplished. Indeed, the advances in technology is more and more allowing evaluators to afford sophisticated hardware and software able to collect information about remote users interacting. This allows evaluators to extend the data collected regarding the user behaviour and state, including the emotional state, in order to have a more complete analysis of what happens during task accomplishment and better identify the potential usability issues

7 Towards a framework for usability evaluation

Sections 2-6 have outlined and evaluated current practice in usability evaluation in important areas of multimodal and natural interactive systems and in tools in support of remote usability evaluation of such systems. This section aims to point to similarities and differences between the areas presented in the previous sections and discusses why one joint framework might not be feasible.

7.1 Similarities and differences

Sections 2-6 all discuss *how* to evaluate (method) and *what* to evaluate (criteria), and Section 6 also mentions tools in support of evaluation.

With respect to how to evaluate, there seems to be general agreement that a range of usability evaluation methods are available and useful independently of which area of multimodal and natural interactive systems is being addressed. The evaluator needs to know which methods are available, how they are used, their advantages and drawbacks, which issues they cover and what they do not help evaluate, etc. On this background, and given information on the development stage of the system to be evaluated, the resources available, the focus of the test, and possibly other issues related to the development project, the evaluator should be able to choose the most appropriate usability evaluation method(s) for his/her purpose.

As regards what to evaluate, there exists a multitude of evaluation criteria. However, there is no such thing as a standard set of evaluation criteria, and the criteria mentioned in the different sections only have some partial overlaps. There is agreement that the criteria which are relevant in a concrete evaluation situation differ. Thus, a framework must list possibilities but should not impose the use of one particular set of criteria. The choice of a relevant subset of criteria should be left to the evaluator in the concrete situation.

The choice of evaluation criteria is tightly related to issues, such as the purpose of the evaluation, the type of application, and the development stage of the system. Basically, there is agreement that performance criteria in a very broad sense are crucial and that information about the context of use is important, not least with respect to applications on mobile devices. Some performance criteria can be measured quantitatively, e.g., the time to carry out a task and the error rate, while others must be measured qualitatively or even subjectively, e.g., naturalness of interaction, ease of use, and perceived cognitive load. In any case, a criterion must be well-defined in some sense for an evaluator to use it and for others to interpret the results unambiguously. Quantitative criteria may be easier to define than qualitative ones. For example, the time it takes to complete a task is easy to measure and understand. However, even quantitative measures are not always that simple. Error rate, e.g., may be defined in several ways with borderline cases. In order to count the number of errors, one needs an exact definition of what kind(s) of error we are looking for and what must be counted as an error. Qualitative criteria may be even more tricky to define, which means that attempts at definitions are often ambiguous, unclear and open for interpretation. For example, ease of learning may be defined as “How easy is it to learn the main system functionality and gain proficiency to complete the job” [http://www.informatik.uni-bremen.de/gdpa/def/def_u/USABILITY.htm], but this definition does not tell us precisely how to measure learnability. A much more strict definition would be, e.g., that ease of learning will be measured in terms of which percentage of users is able to perform a given (number of) basic task(s) within a given time frame. Thus, e.g., the system may be declared as

easy to learn to use if 90% of novice users manage within the set time frame. Subjective criteria typically involve some amount of interpretation. Even Likert scale questions which are often included in questionnaires where users are asked to score a number of statements on a scale between 1 and 5 or 1 and 7, require interpretation. Free-style answers to questions, such as “was it fun to use the system” and “how was the quality of the speech synthesis” are likely to require even more interpretation. This can hardly be avoided but the process may perhaps be supported better than it is today.

7.2 One or more frameworks for usability evaluation?

On the basis of the previous sections and the above discussion, the question to be investigated is if we can create a usability evaluation framework which, at least to some extent, will make life easier for evaluators and give evaluation results that are easier to compare and interpret than is the case today.

Another open question is if it makes sense to go for one joint usability evaluation framework for the whole area of multimodal and natural interactive systems – or at least for the sub-areas covered by the partners.

Regarding usability evaluation methods, a joint framework definitely seems realistic since the methods used are general across application types and sub-areas. In particular, new evaluators would no doubt benefit from a thorough and general description of evaluation methods available, including explanations of advantages and disadvantages, when they are useful, what they help evaluate and what they do not help evaluate, which method combinations may be useful, etc.

Also tools for remote usability evaluation would probably fit into a joint framework with a description of what they can and cannot do, what to be aware of, how to use them, etc.

Concerning evaluation criteria, the question of one or more frameworks is a much more open one which should be further explored. Some criteria are quite closely related to a particular sub-area of multimodal and natural interactive systems, e.g., intelligibility of speech synthesis, while others are much more general, such as ease of use, but may be defined in different ways.

A framework dealing with evaluation criteria should pay attention to differences in definition and use across sub-areas. If this is not possible to do in a proper way, then it may be better to have a criteria framework per sub-area. A framework should include known evaluation criteria along with a proper definition and one or more proposals for how to use them in evaluation.

Future work in the SIMILAR SIG on usability evaluation will further investigate the challenges involved in establishing one or several usability evaluation frameworks and will explore new as well as not yet well-defined evaluation criteria as part of the implementation of the framework(s) decided upon.

References

Albers, M. and Kim, L.: User Web Characteristics Using Palm Handhelds for Information Retrieval. Proceedings of IPCC/SIGDOC Technology and Teamwork. Cambridge, MA , 2000, 125-135

Bernsen, N.O.: Multimodality in Language and Speech Systems - from Theory to Design Support Tool. In Granström, B., House, D., and Karlsson, I. (Eds.): Multimodality in Language and Speech Systems, Kluwer Academic Publishers, Dordrecht, 2002, 93-148.

Bowman D. A.: Interaction Techniques for Common Tasks in Immersive Virtual Environments: Design, Evaluation, and Application. Ph.D. thesis, Georgia Institute of Technology, 1999.

Burdea C. G.: Force and Touch Feedback for Virtual Reality. John Wiley & Sons, Inc, ISBN 0-471-02141-5, 1996.

Card, S., Pirolli, P., Van der Wege, M., Morrison, J., Reeder, R., Schraedley, P., and Boshart, J.: Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability. Proceedings ACM CHI, 2001, 498-504.

Carroll, J. M.: Human-Computer Interaction in the New Millennium. Addison-Wesley Publishing, 2002.

Castillo, C., Hartson, H. R., and Hix, D.: Remote Usability Evaluation: Can Users Report their own Critical Incidents? CHI'98 conference summary on Human factors in computing systems, Los Angeles, California, US, 253-254, 1998, ISBN:1-58113-028-7.

Correa, P, Marqués, F, Marichal, X and Macq, B.: 3D Human Postures Estimation Using Geodesic Distance Maps. SPECOM 2004, St. Petersburg.

Cox, J. C.: A Review of Statistical Data Association Techniques for Motion Correspondence. International Journal of Computer Vision, 10(1):53--66, 1993.

Danielson, D.: Transitional Volatility in Web Navigation. IT&Society, 1(3): 131-158, 2003.

Denis, C., and Karsenty, L.: Inter-Usability of Multi-Device Systems - A Conceptual Framework. In A. Seffah and H. Javahery (Eds.): Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces, ISBN: 0-470-85444-8, November 2003.

Dybkjær, L. and Bernsen, N.O.: Usability Issues in Spoken Language Dialogue Systems. In Natural Language Engineering, Special Issue on Best Practice in Spoken Language Dialogue System Engineering, Volume 6 Parts 3 & 4 September 2000, 243-272.

Dybkjær, H. and Dybkjær, L.: From Acts and Topics to Transactions and Dialogue Smoothness. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004), Vol. V, Lisbon, Portugal, May 2004, 1691-1694.

Elting, C, Strube, S, Möhler, G, Rapp, S. and Williams, J: The Use of Multimodality within the EMBASSI system. Proc. of M&C2002, Usability Engineering Multimodaler Interaktionsformen, Hamburg 2002.

Emery, V. K., Edwards, P. J., Jacko, J.A., Moloney, K. P., Barnard, L. Kongnakorn, T., Sainfort, F., and Scott, I.U.: Toward Achieving Universan Usability for Older Adults Through Multimodal Feedback. Proc. of CUU'03, November 10-11 Vancouver, British Columbia, Canada. 2003, 46-53.

Esch-Bussemakers van M. P. and Cremers, A.H.M.: User Walkthrough of Multimodal Access to Multimodal Databases. Proc. of ICMI 04, Pennsylvania,USA, October 13-15, 2004. 220-226

Feldman, L. S., Sherman, V. and Fried, G. M.: Using Simulators to Assess Laparoscopic Competence: Ready for Widespread Use. Surgery, 135(1):28-42, 2004.

Feygin, D., Keehner, M., Tendrick, F.: Haptic Guidance: Experimental Evaluation of a Haptic Training Method for a Perceptual Motor Skill. Proc. of the 10th Symp. On Haptic Interfaces For Virtual Envir. & Teleoperator Systs. (HAPTICS'02)

Fried, G. M.: Simulators for Laparoscopic Surgery: a Coming of Age. Asian J Surg, 27(1):1-3, 2004.

Grantcharov, T. P., Kristiansen, V. B., Bendix, J., Bardram, L.., Rosenberg, J. and Funch-Jensen, P.: Randomized Clinical Trial of Virtual Reality Simulation for Laparoscopic Skills Training. Br J Surg, 91(2):146-150, 2004.

Hartson, H. R., Castillo, J. C., Kelsa, J., and Neale, W. C.: Remote Evaluation: The Network as an Extension of the Usability Laboratory. In M. J. Tauber, V. Bellotti, R. Jeffries, J. D. Mackinlay, and J. Nielsen (Eds.): Proceedings of the Conference on Human Factors in Computing Systems,Vancouver, Canada, April, New York, NY: ACM Press, 1996, 228– 235.

Ivory, M. Y. and Hearst, M. A.: The State of the Art in Automating Usability Evaluation of User Interfaces. ACM Computing Surveys, 33(4), December 2001, 470-516.

Kaster, T., Pfeiffer, M., and Bauckhage, C.: Combining Speech and Haptics for Intuitive and Efficient Navigation through Image Databases. Proc. ICME'03, November 5-7, Vancouver, British Columbia, Canada. 2003, 180-187.

Kim, H. K., Ratter, D. W. and Srinivasan, M.A., The Role of Simulation Fidelity in Laparoscopic Surgical Training, Proceedings of the 6th International Medical Image Computing & Computer Assisted Intervention (MICCAI) Conference, Berlin, LNCS 2878, pp. 1-8, 2003.

Kneebone, R.: Simulation in Surgical Training: Educational Issues and Practical Implications. Medical Education, 37(3):267-277, 2003.

Krueger, M.: Artificial Reality. Addison-Wesley, 1983.

Lamata, P., Gómez, E.J., Sánchez-Margallo, F.M., Lamata Hernández, F., del Pozo, F. and Usón Gargallo, J., Study of Consistency Perception in Laparoscopy for Defining the Level of Fidelity in Virtual Reality Simulation, Surgical Endoscopy, (in press), 2005a.

Lamata, P., Aggarwal, R., Bello, F., Lamata Hernández, F., Darzi, A. and Gómez Aguilera, E.J., Taxonomy of Didactic Resources in Virtual Reality Simulation, Proceedings of The Society of American Gastrointestinal Endoscopic Surgery (SAGES) Annual meeting, (in press), 2005b.

Lister M.: Streaming Format Software for Usability Testing. Proceedings ACM CHI 2003, Extended Abstracts, 632-633.

Nielsen, J.: Usability Engineering. Boston, MA: Academic Press, 1993.

Paganelli, L. and Paternò, F.: Tools for Remote Usability Evaluation of Web Applications through Browser Logs and Task Models, Behavior Research Methods, Instruments, and Computers. The Psychonomic Society Publications, 35 (3), August 2003, 369-378.

Paternò, F., Piruzza A., and Santoro C.: Remote Usability Analysis of MultiModal Information Regarding User Behaviour. MAUSE Workshop at INTERACT, Roma, September 2005.

Picod, G., Jambon, A. C., Vinatier, D. and Dubois, P.: What can the Operator Actually Feel when Performing a Laparoscopy?, Surg Endosc, 19(1):95-100, 2005.

Schijven, M. and .Jakimowicz, J.: Face-, expert, and referent validity of the Xitact LS500 - Laparoscopy Simulator. Surgical Endoscopy and Other Interventional Techniques, 16(12):1764-1770, 2002.

Sener, B., Wormald, P, and Campbell, I.: Evaluating a Haptic Modelling System with Industrial Designers. Proc. of 2nd Eurohaptics International Conference , Wall, S.A., Riedel, B., Crossan, A. and McGee, M. R. (eds), University of Edinburgh, Edinburgh, Scotland , Edinburgh, Scotland, 2002, pp 165-170 .

Seymour, N. E, Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K. and Satava, R. M.: Virtual Reality Training Improves Operating Room Performance: Results of a Randomized, Double-Blinded Study. Ann Surg, 236(4):458-463, 2002.

Sjostrom C.: Designing Haptic Computer Interfaces for Blind People. Proc. of ISSPA 2001, Kuala Lumpur, Malaysia, August 2001.

Sjostrom C.: Using Haptics in Computer Interfaces for Blind People. Proc. of CHI 2001, Seattle, USA, March 2001.

Spence, R.: Information Visualization. ACM Press: New York, 2001.

Tholey, G., Desai, J. P. and Castellanos, A. E.: Force Feedback Plays a Significant Role in Minimally Invasive Surgery - Results and Analysis, Annals of Surgery, 241(1):102-109, 2005.

Tullis, T, Fleischman, S., McNulty, M, Cianchette, C. and Bergel, M.: An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. Usability Professionals Conference, Pennsylvania, 2002.

Tzovaras, D., Nikolakis, G., Fergadis, G., Malassiotis, S., and Stavrakis, M.: Design and implementation of haptic virtual environments for the training of visually impaired. IEEE Trans. on Neural Systems and Rehabilitation Engineering (June 2004), 266-278.

Umeda, T., Correa, P., Marqués, F., and Marichal, X.: A Real-Time Body Analysis for Mixed Reality Application. Proceedings of the Tenth Korea-Japan Joint Workshop on Frontiers of Computer Vision, FCV-2004, Fukuoka, Japan, February 2004.

Valli, A.: Notes on Natural Interaction, 2004.
http://naturalinteraction.org/NotesOnNaturalInteraction.pdf

Wang, Y., and MacKenzie, C. L.: The Role of Contextual Haptic and Visual Constraints on Object Manipulation in Virtual Environments. CHI Letters volume 2, issue 1, 1-6 April CHI 2000 , 532-539

Waterson, S., Landay, J.A., and Matthews, T.: In the Lab and out in the Wild: Remote Web Usability Testing for Mobile Devices, CHI '2002 extended abstracts on Human factors in computing systems, Minneapolis, Minnesota, USA, 2002, 796-797, ISBN:1-58113-454-1.

Watters, C. and MacKay, B.: Transformation Volatility and the Gateway Model for Web Page Migration to Small Screen Devices. In Proceedings of Hawaii International Conference on System Sciences. Big Island, Hawaii, 2004.

Yu, W. and Brewster, S.: Comparing Two Haptic Interfaces for Multimodal Graph Rendering. Proc. of 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2002. HAPTICS 2002.