Project ref. no.	FP6-507609
Project acronym	SIMILAR
Deliverable status	R
Contractual date of delivery	30 November 2004
Actual date of delivery	6 December 2004
Deliverable number	D16
Deliverable title	Usability evaluation issues in natural interactive and multimodal systems - state of the art and current practice
Nature	Report
Status & version	Draft
Number of pages	60
WP contributing to the deliverable	SIG7
WP / Task responsible	NISLab
Editor	Laila Dybkjær
Author(s) (alphabetic order)	Enrique J. Gómez Aguilera, Niels Ole Bernsen, Samuel Rodríguez Bescós, Laila Dybkjær, François-Xavier Fanard, Pedro Correa Hernandez, Benoit Macq, Oliver Martin, 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, state-of-the-art, current practice.
Abstract (for dissemination)	This deliverable provides an overview of the state-of-the-art in usability evaluation within the broad area of multimodal and natural interactive systems. The overview is delimited by the partners' areas of expertise. It thus covers the areas of spoken dialogue systems, vision-based systems, haptics-based systems, mixed reality systems in surgery, and tools for remote usability evaluation. Moreover, it provides an overview of usability evaluation in the applications surveyed in deliverable D17.

Deliverable D16

Usability Evaluation Issues in Natural Interactive and Multimodal Systems - State of the Art and Current Practice

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

December 2004

SIMILAR Network of Excellence – Research in Human-Machine Multimodal Interfaces
www.similar.cc

Section responsibilities

Section 1: NISLab, Denmark (Laila Dybkjær)

Section 2: NISLab, Denmark (Laila Dybkjær and Niels Ole Bernsen)

Section 3: Tele, Université catholique de Louvain, Alterface SA, Belgium (Pedro Correa, François-Xavier Fanard and Oliver Martin)

Section 4: ITI-CERTH, Greece (Dimitrios Tzovaras and Georgios Nikolakis)

Section 5: GBT, Spain, Tele, Université catholique de Louvain, Belgium (Pablo Lamata de la Orden, Samuel Rodríguez Bescós, Enrique J. Gómez Aguilera, Daniela Trevisan, Jean Vanderdonckt and Benoit Macq)

Section 6: ISTI-CNR, HIIS Laboratory, Pisa, Italy (Fabio Paternò and Carmen Santoro)

Section 7: NISLab, Denmark (Laila Dybkjær)

Contents

1 Introduction

This deliverable provides an overview of the state-of-the-art in usability evaluation within the broad area of multimodal and natural interactive systems. The overview is delimited by the partners' areas of expertise and is necessarily a partial one. Thus we have, e.g., not included any state-of-the-art overview of usability evaluation of systems that can capture smell, are based on acoustics other than speech, or can recognise hand-writing.

More precisely the report includes four sections (Sections 2-5) each of which addresses the state-of-the-art in usability evaluation in a particular area of multimodal and natural interactive systems, while a fifth section (Section 6) deals with state-of-the-art in tools support for a particular kind of usability evaluation of multimodal and natural interactive systems, as explained below.

Section 2 covers the area of speech and hearing-based systems (spoken dialogue systems) usability evaluation. The overview is delimited to systems which include speech in both input and output. It distinguishes unimodal task-oriented speech systems, multimodal task-oriented speech systems, and non-task-oriented speech systems. The state-of-the-art in systems which use speech as input modality only or as output modality only or which use acoustics other than speech is not discussed.

Section 3 discusses the state-of-the-art in vision-based systems usability evaluation. This section distinguishes body and pointing interaction as well as facial interaction. The section only considers vision as input modality. It does not discuss what a user can see, i.e. it does not concern output in terms of e.g. text, images, or other graphics generated by the system.

Section 4 concentrates on usability evaluation of haptics-based systems. Both multimodal and non-task-oriented haptics-based systems are considered. There is quite some focus on systems for the visually impaired and systems for interaction with physical objects.

Section 5 deals with usability evaluation of mixed reality systems specialised to surgery. The section primarily discusses image-guided surgery systems and surgical simulators used for training and testing of surgeons.

Section 6 addresses the important area of tools in support of usability evaluation. The focus is on remote usability evaluation. The section discusses possible techniques and how tools can support remote evaluation.

In addition to these five sections we have found it useful to provide an overview of how concrete applications have been usability evaluated. Therefore, Section 7 provides a survey of the applications described in SIMILAR deliverable D17 and summarises how they were usability evaluated. All, except one, of these applications fall within the areas covered by the state-of-the-art sections mentioned above.

2 State-of-the-art in spoken dialogue systems[1]

We have eventually achieved a rather strong baseline for evaluating the usability of task-oriented unimodal spoken dialogue systems (SDSs) although some important gaps in our knowledge remain. The knowledge we have comes from national and international projects which have contributed in-the-small via usability evaluation of the systems built in these projects, and, not least, from projects which - based on such individual projects and evaluation contributions – have tried to generalise and propose usability evaluation recommendations. EAGLES [Gibbon et al. 1997] and DISC [Dybkjær and Bernsen 2000] are well-known examples of projects that have collected and built on experience and results from many other projects and proposed guidelines for usability evaluation as well as for technical evaluation of SDSs and their components. The PARADISE framework [Walker et al. 1997] is also well-known, focusing on a particular metrics for usability evaluation. The framework views user satisfaction as a measure of system usability and seeks to predict user satisfaction by quantitative metrics. See [Dybkjær et al. 2004a] for a review of EAGLES, DISC, PARADISE and several other projects.

However, since research systems for several years have been moving beyond task-oriented unimodal SDSs towards multimodal task-oriented SDSs and towards non-task-oriented conversational SDSs, there is an increasing need for knowledge of how to evaluate the usability of these systems. In many respects this remains an open research issue. We are not starting from scratch, however, since it would seem obvious to draw on methods and criteria from task-oriented unimodal SDS usability evaluation. But we still need to decide – not least for non-task-oriented SDSs – what exactly is transferable and which new evaluation criteria and metrics are required.

This section discusses current trends and reviews some existing experiences and results in usability evaluation of multimodal as well as non-task-oriented SDSs.

2.1 Challenges in usability evaluation

Usability evaluation of SDSs is to a large extent based on qualitative and subjective methods and criteria and (mostly) concerns the system as a whole, such as the adequacy of its error handling or the spoken interaction naturalness. As mentioned, gaps remain in our knowledge of usability evaluation of unimodal task-oriented systems. A major gap concerns what usability actually is and what exactly makes a user like a system. We know that there are several contributors to user satisfaction but we hardly know them all nor the extent to which each of them contributes. Moreover, the importance of each criterion may differ across users and user groups.

In addition, we are faced with a number of new usability evaluation issues depending on the type of system we are dealing with. For task-oriented multimodal SDSs, a main challenge is to find criteria for evaluating the combinatorial contribution to usability and user satisfaction of the non-speech input and/or output modalities. For non-task-oriented unimodal or multimodal SDSs, usability evaluation must be based on the nature of conversation rather than that of shared-goal information-exchange dialogue, which poses new questions as to which of the criteria typically used in evaluating task-oriented SDSs are relevant at all.

[1] This section is based on [Dybkjær et al. 2004b].

Furthermore, the increasing sophistication of SDSs, whichever their modalities and whether task-oriented or not, continues to demand new evaluation metrics. For example, SDSs may be operated in mobile environments and not only in a static environment. There are now research systems which include on-line user modelling to provide more flexible and adaptive dialogue behaviour. Some systems aim to recognise the user's emotional state to provide more appropriate and natural system reactions. User preferences and priorities raise new issues in such systems. Some implications for usability evaluation are outlined in the following.

Speech may be a good choice in mobile environments due to its modality properties of being hands-free and eyes-free, but speech is not very private in public spaces and speech recognisers are sensitive to noise. Thus, the consideration of complementary modalities becomes highly relevant. Mobile SDSs raise several evaluation issues which have not been fully solved, including how (not) to use, and when (not) to use, (very) small screens in combination with speech, see e.g. [Sturm et al. 2004]; for which purposes (not) to use location awareness and situation awareness; and when and for which purposes it is (not) safe to use displays in, e.g., cars [Bühler et al. 2002, Minker et al. 2004b].

On-line user modelling for SDSs is receiving increasing attention for several reasons [Bernsen 2003]. Users of mobile devices, which are usually personal belongings, may benefit from functionality which builds knowledge of the individual user. Generic user modelling may also be useful. For instance, novice users could receive more extensive interaction guidance and users who repeatedly make particular types of error could be helped by explicit advice or by adaptation of dialogue structure or initiative distribution. General on-line user modelling is an active research area, see, e.g., [Brusilovsky et al. 2003]. Some key evaluation questions regarding on-line user modelling concern: (i) if the user modelling functionality is feasible and (ii) if it will be of benefit rather than a nuisance to the majority of users of the application. For instance, even if the system has enough information on an individual user, adaptation may fail because of too primitive update algorithms or insufficient information about when the user model has been used.

Not only recognition of users' emotional states but also system expression of emotion is an active research area [André et al. 2004]. For spoken input, the main focus is on prosody [Batliner et al. 2000, Hirschberg et al. 2001]. Regarding multimodal interaction, research addresses areas, such as the recognition of facial expressions of emotion [Cohen et al. 2003], or speech-cum-facial emotion, as in the ERMIS project (www.image.ntua.gr/ermis/) on emotionally rich interaction systems. Usability evaluation must consider which impacts (positive and negative) emotion modelling has on users.

User preferences can make life hard for the developer as they may contradict what is empirically the most efficient solution. Some users may, e.g., prefer pen-based input to spoken input or keypad-based input to spoken input, simply because they feel more familiar with GUI-style interfaces [Jameson and Klöckner 2004, Sturm et al. 2004]. Depending on the target user group(s), alternative modalities may be needed because it is likely that each of them will be preferred by some users. This is just one reason why user involvement from early on is recommended and why on-line user modelling appears attractive. Some preferences we can design for, such as modality preferences. Others, however, are hard to cope with. Thus, some users may prioritise speed (no queues on the line) or economical benefit (queues but cheap or free calls), while others prioritise human contact. The question is whether we can build systems with a usability profile that will make these users change their priorities, and exactly which usability issues must be resolved to do so.

There is a growing body of results from very different projects which have built and evaluated various aspects of task-oriented multimodal SDSs. Often, the evaluation is done in much the same way as for unimodal SDSs but with additional focus on what the novel modalities might contribute. For non-task-oriented SDSs, there are still few results. Below, we review some approaches to the usability evaluation of such next-generation systems, being aware that this overview is far from complete. Rather, we try to exemplify different trends today.

2.2 Evaluation of multimodal SDSs

Broadly speaking, we may distinguish between at least the following approaches to usability evaluation of task-oriented multimodal SDSs: (i) Empirical investigations of modality appropriateness, including comparison of SDSs with different modality combinations, and evaluation of user preferences. Focus is on deciding which combination is best suited for a concrete application, user group, environment, etc. (ii) Empirical evaluation of the effects on interaction of animated talking agents. (iii) Theory-based evaluation of SDSs. This is typically done early in the development process and is a relatively cheap method, but it does require an appropriate theory.

2.2.1 Empirical approaches to modality appropriateness

2.2.1.1 System comparisons and frameworks

To get an idea of how well different modalities work in combination and of their effect on users, several comparative studies have been made of users interacting with different systems. Often, the three ISO-recommended usability parameters are used in the evaluation, i.e. effectiveness (measured as dialogue success rate), efficiency (measured as time to task completion), and user satisfaction (measured by a questionnaire) [ISO]. For example, Sturm et al. [2003] compared a user-driven and a mixed initiative multimodal SDS on a train timetable information task. Both interfaces offered spoken and pen-based input and display output. The mixed initiative version used speech to guide the dialogue whereas, in the user-driven version – mainly for expert users - the user communicated via tap-and-talk, i.e., the user indicated on the screen which field to fill in next. The effectiveness was found to be approximately the same for the two interfaces whereas the efficiency was higher for the user-driven interface which was also the interface preferred by most users.

Cohen et al. [2000] compared the use of a standard GUI interface and an interface with pen and voice input and graphics and voice output. The application was a military task in which units and control measures had to be placed on a map. They showed that the pen/voice SDS interface was faster, also regarding error correction, and strongly preferred by users.

The parameters of efficiency, effectiveness and user satisfaction are basically also those we find at the bottom of the PARADISE framework [Walker et al. 1997]. In the German SmartKom project, PARADISE has been extended for the purpose of usability evaluation of task-oriented multimodal SDSs. SmartKom allows input speech and gesture and output via speech and screen graphics. SmartKom operates in three environments, i.e. home, mobile, and public. The questionnaire used was adapted to collect information on the different SmartKom scenarios. It includes and extends the usability survey developed in PARADISE. Also, the measurement of dialogue costs, such as dialogue quality, is modified to take into account that the system includes several modalities which may be used in different combinations [Beringer et al. 2002].

2.2.1.2 User preferences

When speech is the only input/output option, the user is in no doubt about which modality to use, no modality is ignored, and no modality preferences are catered for. The addition of modalities creates the need for usability evaluation of the appropriateness of the offered modalities in relation to application and user group, and of the clarity in presentation to the user of what they can be used for.

den Os et al. [2001] conducted an expert evaluation of a speech and pen input, text and speech output directory assistance service running on an iPAQ. The evaluation showed that it must be unambiguous which modalities are available when during interaction, if this may vary. If, e.g., speech has been available at some point, users will expect speech to remain available unless explicitly told that this is no longer the case. It is a design challenge to clearly convey which modalities are available, and when. The authors subsequently made a user test of the same system. The test showed that users have different modality preferences, which affect the way they interact with an application. Several other studies confirm that users have different modality preferences. Sturm et al. [2004] analysed the behaviour, preferences, and satisfaction of subjects interacting with an SDS using speech input/output, pointing input and graphics output. Jameson and Klöckner [2004] made an experiment showing different modality preferences in a mobile phone task. The task of calling someone while walking around could be carried out using speech and/or keypad input and acoustic and spoken output and/or display.

2.2.2 Animated talking agents

Animated talking agents (face-only or embodied) have become a popular research area. Usability evaluation of these systems often concern issues such as life-likeness, perceived intelligence, credibility, reliability, efficiency, personality, ease of use, and understanding quality [Bickmore and Cassell 2005, Dehn and van Mulken 2000, Heylen et al. 2005]. The effect of this kind of systems is typically measured either in terms of the user's preferences or via the user's performance. Dehn and van Mulken [2000] conclude that, so far, there is no evidence of any general advantage of having an interface with an animated agent over one without. This is supported by [Cole et al. 2005]. It is also in line with the findings in [Bickmore and Cassell 2005] who evaluated the effects on communication of a real-estate talking agent vs. an over-the-phone version of the same system, in which only the apartments and not the agent could be seen on a screen next to the phone. The perception of efficiency seemed to be gender-dependent, but users generally liked the system better in the speech-only condition. Probably, the lack of natural human behaviour of the agent had a negative effect on users. That this may have an effect is to some degree confirmed by the findings in [Heylen et al. 2005] where controlled experiments were made on the effects of different eye gaze behaviours of a cartoon-like talking face on the quality of human-agent dialogues. The most human-like behaviour led to higher appreciation of the agent and more efficient task performance.

Despite the general conclusion in [Dehn and van Mulken 2000] mentioned above, agents do exist which appear to improve, e.g., intelligibility for users with special problems. Granström and House [2005] have used a talking head in several applications, including tourist information, real estate (apartment) search, aid for the hearing impaired, education, and infotainment. Evaluation has shown a significant gain in intelligibility for the hearing impaired when a talking face is added. Eyebrow and head movements enhance perception of emphasis and syllable prominence. Over-articulation may be useful as well when there are special needs for intelligibility. The findings in [Massaro 2005] support these promising conclusions, focusing on applications for the hard-of-hearing, children with autism, and child language learning more generally.

2.2.3 Theory-based approaches

Usability evaluation is often done by some kind of user testing, cf. the descriptions above. However, the approach of [Elting et al. 2002] in the Embassi project is a heuristic one. The Embassi system is meant for interaction with home entertainment systems and allows for speech and gesture input and acoustic and graphical output. Heuristic evaluation is motivated as being less time-consuming and expensive than user testing. Based on the modality properties in [Bernsen 2002], a set of guidelines is derived and used together with GUI design guidelines [Nielsen 1994] to evaluate modality appropriateness.

Given the overwhelming number of modality combinations which could be compared in principle, it may be worthwhile to further explore theory-based approaches. Potentially, much effort could be saved on comparative studies if we can establish a solid set of guidelines based on, e.g., modality theory as suggested by [Elting et al. 2002].

2.3 Evaluation of non-task-oriented SDSs

Despite the frequent use of the term ‘conversational’ by researchers today [van Kuppevelt et al. 2005] only few non-task-oriented SDSs have been developed so far and little has been done regarding their usability evaluation. Some of the usability criteria typically used for task-oriented systems become irrelevant, such as sufficiency of task coverage, and probably also efficiency and informativeness. Instead, other issues arise, such as conversation success and naturalness.

The August system [Gustafson et al. 1999] allowed users to interact with the Swedish author August Strindberg about various topics via spoken input, and speech and facial output. It was developed in the late 1990s but did not lead to novel usability evaluation metrics. The NICE project [Bernsen et al. 2004a] develops a non-task-oriented multimodal SDS enabling “real” conversation with life-like fairytale author Hans Christian Andersen via speech and pointing gesture input and speech and graphics output. The usability evaluation criteria proposed in the project include several known from unimodal task-oriented SDSs, but also include criteria for evaluating modalities other than speech, e.g., quality and adequacy of all input and output modalities. However, new challenges are being considered, including metrics for conversation naturalness, such as conversation success, common ground, interlocutor contribution symmetry, topic shift adequacy, educational value, and entertainment value [Bernsen et al. 2004b].

It is clearly too early to make any firm conclusions regarding usability evaluation of non-task-oriented SDSs but, surely, novel and, in some cases re-defined, metrics will be needed as suggested by NICE.

Cole et al. [Cole et al. 2005] present ongoing work on tutoring systems and envision that it will be possible in the near future to build life-like characters that interact with people much like people interact with each other. Spoken dialogue technology must be combined with computer vision and animated agent technology to achieve this goal. An important evaluation criterion of such tutoring systems will be if there is any learning benefit.

2.4 Conclusion

We have addressed the current usability evaluation baseline for unimodal task-oriented SDSs. We have discussed some remaining gaps in our knowledge of usability evaluation and some new challenges ahead caused by the increasing sophistication of SDSs as well as by research moving into multimodal and to non-task-oriented SDSs. We have reviewed approaches to usability evaluation in several finished and ongoing projects on multimodal task-oriented and non-task-oriented SDSs.

There seems to be a broad need for usability evaluation that can help us find out how users perceive these new kinds of SDSs and how well users perform with them, possibly compared to other types of system. There is a strong wish to find ways in which usability and user satisfaction might be correlated with technical aspects in order for the former to be derived from the latter. We don’t have methods today that enable prediction of how well users will receive a system. We just know that a technically optimal system is not enough to produce user satisfaction. Regarding modality appropriateness which is a central issue in multimodal SDSs, modality theory may be a promising and powerful approach to usability evaluation of modalities at an early stage. However, user tests of the actual design will still be needed, as for unimodal SDSs.

3 State-of-the-art 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.

3.2 Gestural interaction

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.

3.2.1 Whole Body interaction

Since Myron's Krueger visionary work in the mid-1970s [Myron Krueger, 1991], 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].

On another more complex level, individual segmentation is often needed in order to be robust enough, which implies an environment dependence of some kind, even if this aspect is beginning to be less and less restrictive (mere lighting dependent environments are replacing complete blue-key infrastructures).

Whole two-dimensional body gesture interaction is often achieved real-time using bounding-box [Cavazza et al.2004] or convex hulls techniques [Haritaoglu et al. 1998] as their core interactive algorithm, which has to be enriched with an efficient and flexible individual segmentation, collision detection algorithms, a priori context information and anatomical heuristics, to name just a few of the challenges those applications have to be confronted to.

This kind of applications is beginning to perform a commercial breakthrough with very high rates of end-user interest and satisfaction, like those presented by the Alterface company [http://www.alterface.com/], the embodied gaming experience presented by the Helsinki Media Lab [http://www.mlab.uiah.fi/animaatiokone/kungfu/en/], or the Vivid company [http://www.vividgroup.com]. Still, they tend to be very cumbersome infrastructures, if not completely unmovable.

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.

To do so, skin colour detection is becoming a good backup technique, but still too unstable and environment dependent to be used as a feature tracking technique for itself.

Some other three-dimensional techniques using 3D human models [Fua et al. 2004], that were traditionally hard to use real-time but very efficient in human feature extraction are giving good results in this area.

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 et al.].

3.2.2 Pointing based techniques

These techniques mainly aim the navigation area. They can be applied for interface as well as immersive navigation. All the above comments concerning precision and user satisfaction may be applied to this section as well. The visual feedback is in these cases more important in order to counterbalance compensate the lack of precision. Accuracy is nevertheless increasing, reaching values of about 20 pixels [Demirdjian et al. 2002].

Stereo vision often used in those kinds of setups, even though interaction occurs only with two dimensional presentations. The location pointed in a screen is defined by the three dimensional location of the eyes and the pointing hand or finger of the person [Nicekl et al. 2003].

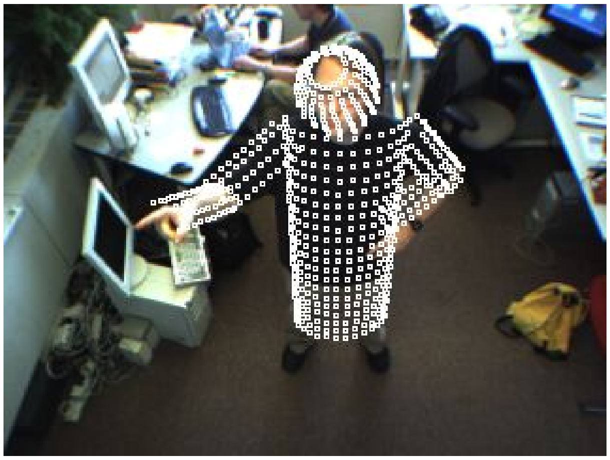

Figure 3.1. Photo from [Dermidian et al. 2002].

3.3 Face analysis techniques

Automatic analysis of facial gestures is rapidly becoming an area of intense interest in signal processing and human-computer interaction communities. Our aim is to explore the issues in design and implementation of a system that could perform automated facial expression analysis. In general, three main steps can be distinguished when tackling the problem. First, the face must be detected. Next is to devise mechanisms for extracting the facial expression information while the final step consists of defining a set of categories, which we want to use for facial expression classification.

3.3.1 Face detection and tracking

Most face detection and face tracking algorithms are built using predefined conditions: illumination and pose are imposed to allow robust detection. However, at least two face-tracking systems dealing with rigid-head motions have already been implemented, leading to satisfying results ([Black et al. 1997], [Edwards et al. 1998]). The method for tracking the face can be of three types : either to detect a face as a whole unit (*holistic* representation), either as a set of features (*analytic* representation) or in a *hybrid* way (a combination of both approaches). The applied face representation and the kind of input images then determines the choice of mechanisms concerning the automatic extraction of the facial expression information.

3.3.2 Face analysis

A huge number of techniques have then been developed to extract some sets of distinctive features, that can be used for automatic emotion classification. It is however very difficult to assess the quality of these specific methods since almost all results are obtained using different databases. In the next section, we discuss some of the techniques that are among the most efficient, based on the literature available so far.

Emotion recognition using facial expression emerged from the work done by Paul Ekman [Ekman et al. 1975, Ekman et al. 2002]. This work shows that facial expressions of six basic emotions are universally recognized and a mapping of facial muscle(described as activation of Actions Units) to these emotions was constructed and named Facial Action Coding System (FACS). Although FACS was designed for emotion recognition by humans, it has been the basis for automatic emotion recognition based on facial expressions. The differences among algorithms are in the feature extraction methods, type of classifiers and whether the recognition is done from still images or video sequences. Essa and Pentland [Essa and Pentland, 1994] used optical flow to extract the features from video sequences and a distance-based classifier. Mase [Mase, 1991] also used optical flow results as features, but the classifiers used were Hidden Markov Model based classification, ruled base classification, Neural Network classifier, and template based K-Nearest Neighbor classifiers. Black et al. [Black et al. 1997] used local parametric models as a means of tracking and recognizing facial expressions using a rule-based classifier. Edwards et al. [Edwards et al. 1998] and Hong et al. [Hong et al. 1998] used template based classifiers with different feature extraction methods for emotion recognition from still images. Recognition rates for still image methods are generally lower than for sequence based methods. Recent work has used both vision and audio information for emotion recognition. Chen et al. [Chen et al. 1998, Chen 2000] used a parametric 3-D model to extract the Action Units from video, and several audio features and a network based classifier. De Silva et al. [De Silva et al.1997] shows that human judgments of emotions improves when using both modalities. A recent paper Donato et al. [Donato et al. 1999] compared various techniques for the automatic recognition.

3.3.3 Usability evaluation

To accomplish the usability evaluation of such systems, one must compare the different systems to the ideal system : the human visual system. It may not be possible to incorporate all features of the human visual system into an automated system, and some features may be even undesirable, but it can certainly serve as a reference point.

A first requirement that should be posed is that all of the stages of the facial expression analysis should be automated. Then, depending on the application, we may want the system to run in real time to avoid unacceptable delays (in interactive application for instance, this is a crucial aspect for the usability of the developed system).

In order to be universal, the system should be able to analyse subjects of both sexes, of any age and ethnicity. Also, no constraints should be set on the outlook of the observed subjects. The system should perform robustly despite changes in lighting conditions and distraction like glasses, changes in hair style, and facial hair like moustache, beard and grown-together eyebrows. Similarly to the human visual system, an ideal system would “fill in” missing parts of the observed face and “perceive” a whole face even if a part of it is occluded. The system should also be able to deal with rigid head motions. Pantic et al. [Pantic et al. 2000] have reviewed most of the most promising approaches, comparing them to this ideal system, leading to one of the most developed usability evaluation known in this domain.

A second approach to assess the usability of the systems is to analyse them in terms of their impact on their users. As the range of applications is huge, we chose to focus on an interactive edutainment application. The goal was to assess the effectiveness of such systems on children.

To accomplish the usability tests of interactive systems, built on top of the emotion recognition systems described above, sessions of children working with the computer were videotaped and segmented. Segments have then been presented to other children and to teachers for their judgments of the appropriateness of the computer's actions, and for suggestions for improvement. In addition, interactional analyses [Jordan et al. 1995] are carried out to examine subtleties in the types of actions being taken by the computer under different user and task state conditions.

Once the system is capable of a reasonable degree of interaction, we can examine the effects that the various computer actions have on the users. This can be accomplished either through analysis of videotapes (including going through a tape with the subject herself to probe her reactions) or through using 'think-aloud' methods in which children verbalize their reactions as they are engaged in the task.

As a system matures we are able to evaluate the extent to which different aspects of proactive interaction facilitates the achievement of learning, motivational, and affective goals. Since the goals are varied, this evaluation must be multidimensional. It is expected to include the length of time children remain at the computer, the proportion of this time that the child is actually engaged in the task, the frequencies of negative and positive reactions, whether children are more successful in accomplishing tasks with proactive assistance and encouragement, and whether principles learned in building one object transfer to building another. A proactive computing system must be constructed in such a way that parts of its interaction capability can be disabled, thus allowing investigations of the effects of the presence or absence of these capabilities.

3.4 Face analysis techniques

3.4.1 Intro

Facial animation control is the mechanism whereby a model can be articulated. Selecting a particular facial control mechanism depends on the purpose of the animation; for example, a lipreadable facial model requires a high level of precise three-dimensional manipulation of the lips, mouth, teeth and tongue. However, facial image-coding requires only a relatively simple model with a focus on pixel manipulation. Obviously, it would be inappropriate to suggest a single strategy for both situations; there are, however, many overlapping techniques that apply to both applications.

3.4.2 Techniques

The facial animation techniques may be split in 6 fundamental categories:

3.4.2.1 Interpolation

Interpolation which stems from the early work of Parke [Parke 1972] is perhaps the most widely used of the techniques for facial animation because it offers an intuitive approach to facial animation. Typically, an interpolation function specifies smooth motion between two key-frames (or more) at extreme positions, over a normalized time interval. It is a fast technique but which presents two major drawbacks: the range of expressions is related to the number of the expression poses available (an expression that falls outside the bounds of the key pose set is unattainable) and each key pose requires an explicit geometric data collection or data generation effort (further combination of independent face motions are difficult to produce).

3.4.2.2 Parametrization

Parametrized animation is more efficient because it only needs to manipulate sets of parameters, to create a sequence of images. Ideal parametrization specifies any possible face and expression by a combination of independent parameter values [Parke et al. 1996]. This method overcomes some limitations of interpolation techniques by providing a large number of facial expressions with relatively low computational cost. However, there are noticeable motion boundaries and the choice of the parameter sets depends on the facial mesh topology. Moreover, manual tuning of parameters is needed and tedious (this may often lead to coding of unrealistic expressions in case of conflict between parameters [Waters et al. 1995]).

There are some well-known models for parametrized facial animation. Among them, the Parke model ([Parke 1974], [Parke 1982], [Parke 1989]) is a procedural model with some parameters for the face construction and some others bounded to the face dimensions to deform. Even with a modest number of parameters, Parke obtained with his model a wide range of faces and expressions.

We should also mention the MPEG-4 model with whom it is possible to animate a face by defining once 84 facial definition parameters (FDP) and animating them with facial animation parameters (FAP). This technique allows the user to code global movements of the visage in the scene and facial expressions, to synchronise lips with natural or synthetic speech and to adapt them to a particular facial morphology [Pandzic et al. 2002].

3.4.2.3 Muscle-based

Skin and muscle-based animation is a technique based on human faces anatomy. Physically based models attempt to model the shape and dynamic changes of the face by modelling the underlying properties of facial tissue and muscle action by means of physically computed forces. Those models fall into three categories. The first one is the mass-spring method, which propagates muscle forces in an elastic spring mesh that models skin deformation [Platt 1985]. The vector representation approach deforms a facial mesh using motion fields in delineated regions of influence [Waters 1987]. The layered spring mesh model extends a mass spring structure into three connected mesh layers to model anatomical facial behaviour more truthfully [Terzopoulos et al. 1990] and [Lee et al. 1995].

A well-known example of muscle-based animation system is the FACS (Facial Action Coding System) created by Ekman and Friesen [Ekman et al. 1978]. It consists of a facial movements catalogue composed of a set of 46 minimal visual actions, the action units (AU). The AU describe the direct effects as well as secondary effects (propagation of wrinkles and folds) caused by a muscle or by a set of muscles movements when their action isn’t differentiable by simple visual observation. That facial movements “alphabet” enables reading and objectively representing the way that any visage is animated in the time.

Muscle-based modelling produces realistic results by approximating human anatomy, but it is daunting to consider the exact modelling and parameters tuning needed to simulate a specific human’s facial structure. The main disadvantage of these models is that they are computationally expensive and difficult to control with force-based functions.

3.4.2.4 Pseudo muscle-based

Pseudo muscle-based animation offers an alternative approach to muscle-based animation by deforming the facial mesh in muscle-like fashion but ignoring the complicated underlying anatomy. We make a simplified model of the visage muscles getting so a geometric and not a physic model. Deformation usually occurs only at the thin-shell facial mesh. Muscle forces are simulated in the form of geometric deformation operators like splines, wires or free-form deformations. The free-form deformations distort volumetric objects by manipulating control points arranged in a three-dimensionnal cubic lattice [Sederberg et al. 1996]. In [Kalra et al. 1992], facial animation is driven by procedures called Abstract Muscle Action [Thalmann et al. 1988] that are similar to AUs of FACS and work on specific regions of the face.

With this technique, we obtain a more simplified parametrization and computation but it does not provide a precise simulation of the actual muscle and skin behaviour so that it fails to model furrows, bulges or wrinkles in the skin.

3.4.2.5 Morphing

Morphing effects a metamorphosis between two target images or models. A 2D image morph consists of a warp (i.e. it maps an image onto a regular shape such as a plane or a cylinder) between corresponding points in the target images and of a simultaneous cross dissolve (i.e. one image is faded in and the other one faded out) [Beier et al. 1992]. The morphing methods can produce realistic facial expressions but they share the same limitations with the interpolation approaches. Selecting corresponding points in target images is manually intensive, dependent on the viewpoint and not generalizable to different faces.

3.4.2.6 Performance-driven animation

Performance-based animation involves capturing real human actions or tracking video images to drive synthetic characters. Motion data can be used to directly generate facial animations or to infer AUs of FACS in generating facial expressions. The tracked 2D or 3D feature motions are often filtered or transformed to generate the motion data needed to drive a specific animation system. A wide range of researchs have concentrated on that subject ([Azarbayejani et al. 1993], [Eisert et al. 1998], [Essa et al. 1994], [Essa et al. 1996], [Kass et al. 1987], [Kato et al. 1992], [Li et al. 1993], [Masse et al. 1990], [Saji et al. 1992], [Saulnier et al. 1995]).

The advantage of this kind of animation is a realistic result and high accuracy for timings. However, tracking the actions requires a lot of work. Actually, accurate tracking of feature points or edges is important to maintain a consistent and life-like quality of animation.

3.4.3 Validation Issue

Obviously, none of the techniques described above leads to an optimal control strategy since each of them has its own specific advantages and disadvantages. That is why validation is an important part of any facial animation control strategy. Typically, the validity of a facial animation sequence is done by inspection (*does it look right ?*). This criteria is sufficient for semi-realistic, cartoon-based faces but as the realism of the synthetic facial models improves, we would like them to mimic reality as closely as possible. In this section, we describe a list of requirements that an ideal facial animation technique should satisfy [Pandzic et al. 2002].

First of all, a first requirement is a wide range of possible faces and expressions. Ideally, it should be possible to express any face with any expression. This is a tremendous requirement even if we just consider realistic human faces and expressions. Moreover, an important feature to take into account is its ease of use, complexity and intuitivity. It is usually inversely proportional with the previous requirement because the more we want to extend the range of faces and expressions, the more complex it becomes.

Subtlety is another non negligible requirement due to human sensivity to extremely subtle changes of facial expression. The slightest twisting of a mouth corner may indicate mild discontent or skepticism. Very small narrowing of the eyelids can produce a threatening look and so on. The technique should be able to express and precisely control even the slightest motion.

The technique should be moreover portable which means that it should deliver the same results in term of high-level expressions when applied to different face models. This enables the easy switching of face models and the reuse of facial animation sequences.

Finally, it should be possible to use the technique as basis for higher-level abstraction and when facial animation is needed to be communicated through a network, it should allow data encoding and streaming.

4 State-of-the-art in haptics-based systems

Haptic interfaces provide tactile and force feedback information to the user through physical interaction, enabling a realistic perception of virtual environments. The devices used for haptic interaction are controlled by the human contact forces and can be programmed to provide the human the sensation of forces or torques [Burdea 1996].

Potential benefits from haptic environments can be found in applications concerning several areas such as education, training, industry, medicine, modelling and many more [Sjostrom 1999]. The technical trade-offs and limitations of the currently developed virtual reality (VR) systems are related to the visual complexity of a virtual environment and its degree of interactivity [Sjostrom2001a, Sjostorm2001b].

There are several methodologies used in order to perform usability evaluation of haptic-based systems (HS). Our experience on the evaluation of HSs comes from the usability evaluation followed in several national and international research projects. A usability evaluation procedure has been followed during the development of haptic application for blind users (ENORASI) as well as for the final prototype software [Tzovaras et. al. 2004]. The usability evaluation during development of the prototype consisted mainly on personal communication with end users and comparing the effectiveness of different haptic loop algorithms. The evaluation of the final prototype was based both on questionnaires and observations of the test leader as well as measured metrics such as time for completion and success rate. Similar evaluation procedures have been followed in other projects presented in [Nikolakis et. al.2003, Nikolakis et. al. 2004]

Generally in order to provide a usability evaluation, activities should begin early in development and continue in frequently throughout. A proposed scheme is presented in [http://www.usabilitynet.org/management/b_overview.htm#Evaluation]

- During the first steps of development, users can be asked to step through their tasks following a sequence of screen sketches or paper prototypes.
- If it is impossible to involve real end users, usability experts may be able to evaluate designs based on user and task goals.
- Working prototypes can be tested more formally by users carrying out typical tasks. Task completion and task completion rates are key factors.
- A usability lab is not always essential but it does have the advantage that the developer may watch and discuss about the tests without disturbing the user.
- When a complete prototype is available, usability requirements for user performance and satisfaction can be tested.

The above procedure is proposed as a general scheme for usability evaluations and applies also to HSs.

This section reviews existing usability evaluation methodologies of unimodal, multimodal as well as non-task-oriented HSs.

4.1 Challenges in usability evaluation

The performance of a haptic interface depends on the mechanical characteristics of the haptic device and the control system architecture. Evaluation of HSs has to take into account these two concepts. Another important issue is the nature of each application in order to select the appropriate methodology and metrics in order to perform a usability evaluation.

There are numerous criteria that have been used in order to evaluate HSs. In most cases, usability evaluation of HSs is based on subjective criteria such as the realism of the provided force feedback within the virtual environment and the ease of use of the whole system.

[KilchenanO'Malley] perform an evaluation task in order to compare a haptic environment to real world. The users are asked to feel the exterior of two ridges and determine which is larger. The test is performed both in real world and using haptic interface in a virtual environment. Results are compared in order to evaluate the effectiveness of using haptic interaction for size discrimination in a virtual environment.

Usability evaluation of two haptic devices concerning the perception of texture, object size and angularity is performed in [Penn et. al. 2000]. The users were asked to distinguish virtual textures with the two examined devices and to find the size and angularity of virtual objects. The results of the two devices were compared in order to select the most appropriate device for haptic representation of the examined properties.

4.2 Evaluation of multimodal HSs

4.2.1 Empirical approaches to modality appropriateness

4.2.1.1 System comparisons and frameworks

[Feygin et al. 2002] perform the evaluation of a haptic training method for a perceptual motor skill. In that case the authors use different training methods in order to compare the additional values between different kinds of interfaces. The three cases are visual, haptic and both visual and haptic. Several metrics such as position accuracy, shape accuracy, timing accuracy, and drift accuracy were measured in order to compare the usability and the usefulness of the three interfaces. Such evaluation procedures can be used to provide useful results in order to verify the usability of haptics and estimate the improvement that haptics provide when used in multimodal interfaces.

[Emery et. al. 2003] evaluate unimodal and multimodal haptic interfaces. The evaluation was performed for a haptic interface, auditory and haptic interface and auditory, haptic and visual interface. Specific metrics were selected to perform the evaluation such as the total trial time, the lack of errors during the task and other task specific measures (times to perform parts of the task). In this way the authors measured the efficiency and the accuracy of the system.

Comparison between two haptic interfaces is performed by [Yu et. al. 2002]. Two audio haptic interfaces designed for visually impaired people are examined. The users are asked to perform a series of experiments and they have to answer to a questionnaire about their findings in the examined environment. After performing the tests, the users had also to answer to a questionnaire regarding the workload of each experimental condition.

A double tracked evaluation was used in [Kaster et. al.2003] in order to evaluate speech and a haptic system. They focused both on the effectiveness of the system by measuring error rates and on the usability by measuring the success rate. The measurements were followed by questionnaires where the users could give information about the functionality of the system, the speed of learning, mental load and satisfaction.

The subjective workload measurement used by [Oakley et. al. 2000] consisted of the mental demand, physical demand, time pressure, effort expended, performance level achieved, frustration expectation and fatigue. Users were asked to fill workload charts for each condition after participating to the experiments.

[WangMacKenzie et. al. 2000] perform the evaluation of their system by measuring the total task completion time, counting the average distance error and angular errors and their variances for each subject in order to find if haptic feedback constrains can improve the human performance during the task.

4.2.1.2 User preferences

There are several cases were evaluation of a haptic system is based on questionnaires and comments of the users on the usability of the interface. For example evaluation of a haptic modelling system is performed in [Sener et. al. 2002]. There are three different sets of user that evaluate the system and they are questioned during and after the test about the ease of use of the evaluated system. The users among others are asked to describe what they were planning to do, what they actually achieved using the proposed interface and what difficulties they had faced.

The evaluation of a speech – visual and haptic interface is examined in [Esch-Bussemakers & Cremers 2004]. The evaluation is performed asking the users to fill a pre-test questionnaire concerning their background information and their past experience in multimodal interfaces and then they were asked to review the system keeping in mind the potential of the different functionalities for the future. After the tests, the users were asked to fill a questionnaire regarding the haptic feedback and their overall opinion about the system.

An evaluation of tele-haptic environments is presented in [Shen et. al. 2004]. The evaluation aims in calculating specific drawbacks of the haptic representation caused by time delays in the collaborative haptic environment. A task-oriented evaluation is performed using the total completion time of the task as a metric.

4.2.2 Theory-based approaches

Several evaluation procedures have been performed in order to evaluate the usability of haptic devices. There are some general characteristics that haptic devices and supporting software need to have in order to be usable. As stated in [Burdea1996] measurements have estimated the maximum exertable forces that can be controlled by a human and the frequencies that can be perceived. Additionally, human tactile sensors can perceive different frequencies of haptic feedback satisfactorily. These metrics need to be taken into account in order to create usable haptic interfaces. It is crucial for haptic devices to provide feedback in the range of the forces that can be controlled and with update frequencies that can be perceived by the users. The software supporting the haptic devices has also to take this in account in order to provide the user with a usable interface. In general theory-based usability evaluation cannot suggest an ideal setup, but only indicate that a specific setup will not be usable or will not be perceived well by the users.

4.3 Evaluation of non-task-oriented HSs

Usually haptic interfaces have been formally evaluated using psychophysical methods [Kirkpatrick 1999], [Agus et. al. 2004]. While these methods provide useful information about haptic rendering of a single stimulus, they provide few insights about how well a haptic interface can render the multiple properties that are sensed when interacting with physical objects.

An evaluation of tele-haptic environments is presented in [Shen et. al. 2004]. The evaluation aims in calculating specific drawbacks in the haptic representation caused by the *sluggish* *response*, which is defined as the interval between the time that the collaborator generates its interaction and the time that the corresponding consequence is “viewed” by the collaborator.

4.4 Conclusion

Usability evaluation of non-task oriented HSs is perfomed using psychophysical methods. This kind of evaluation provides useful results in order to understand the limits of the haptic technology as well the limits of the ability of users to perceive haptic information.

On the other hand, for the evaluation of task-oriented and multimodal HSs there seems to be a great variety of usability evaluation metrics. Although there are some metrics such as, the completion time or percentage of successful (or failed) users, that are commonly used it is not possible to determine specific metrics for all cases since there is a great variety of haptic devices combined to a large number of different applications that use haptics.

Generally, it is a good advise to perform usability evaluation, even in an informal way, with end users during the different faces of development of HSs in order to produce usable interfaces.

5 State-of-the-art in mixed reality systems in surgery

5.1 Introduction

Computers are enabling new approaches to surgery and are offering new interesting interaction capabilities. Surgical procedures are being enhanced with different ways to interact with the patient, with a planning done studying the specific anatomy of the patient, or even with a not too far procedure rehearsal.

Introducing new concepts providing solutions to the very complex problems of man-machine interaction in the operating room is not an easy matter [Cinquin et. al. 1995]. This implies research in domains including Applied Mathematics (3-D, or even 4-D, object representation, inverse problems, e.g.), Computer Vision (design of new sensors, e.g.), Image Processing (multi-modal image segmentation and registration, e.g.), Ergonomics (the "template based surgery" concept stemmed out of a careful analysis of surgical procedures, e.g.), Quality (quantifying the concepts of efficacy and effciency, e.g.), and Robotics (design of specific architecture solving safety issues, e.g.). The resulting concepts must prove their technical feasibility in laboratories, and then lead to development of demonstrators meeting quality requirements compatible with the use in surgical environments. This explains why most of current research efforts in the domain of Computer Aided Surgery systems are focused in developing "functionality centered" solutions for specific pathologies driven by technology, with ad-hoc design and little attention is dedicated to ergonomic support of interaction styles, information interfaces and presentation during the surgery [Bartz, et. al., 2001].

In this section titled “State-of-the-art in mixed reality systems in surgery” we are going to be focused on the computer techniques applied to help a surgeon in different ways. Mixed-reality systems offer interesting possibilities of visualization and interaction which can support surgical procedures and surgical simulators.

Many of the systems to support surgical procedures (e.g. described in section 2) are still in the experimental phase of research and usability aspects have been focused on different kinds of visualization taking into account perceptual cues. On the other hand systems using microscope image-guided technology have already been demonstrated its clinical importance.

In the field of surgical simulation (e.g. described in section 3) there is an increasing importance on validation of methods of training and skills assessment [Gallagher et al. 2003a]. Usability aspects have been focused on clinical validation and in laparoscopic surgery, a technique much more simple to be simulated than an open surgery and more complex than other simulated procedures like endovascular intervention or anastomosis.

5.1.1 Mixed reality systems

Mixed Reality (MR) systems are systems that combine real and computer-based information.

Milgram [Milgran et al. 1994] defines the Reality-Virtuality continuum shown in Figure 5.1. Mixed Reality (AR) is the region between the real world and totally virtual environments. Augmented reality lies near the real world end of the continuum. In AR systems, the perception that predominates is the real world augmented by additional capabilities or information provided by the computer system. Vallino [Vallino, on-line] gives a list of 7 application domains where the use of AR reality systems has been investigated: medical domain, entertainment, military training, engineering design, robotics/telerobotics, manufacturing & maintenance, and consumer design. Augmented virtuality is a term created by Milgram to identify systems which are mostly synthetic with some added real world sources such as texture mapping video onto virtual objects, tracking the user's position, etc. Vallino [Vallino, on-line] expects that this distinction will fade as the technology improves and the virtual elements in the scene become less distinguishable from the real ones. Therefore we can say that AR and AV are parts of the Mixed Reality systems and MR systems are any possible combination of real and virtual information .

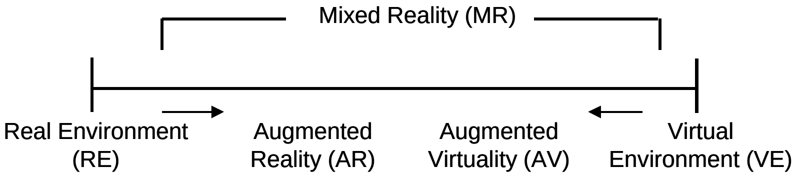

Figure 5.1. Milgram's Reality-Virtuality Continuum [Milgran et al. 1994].

Image-guided surgery (IGS) is a kind of augmented reality system where the purpose is to augment the physical world of the surgeon, i.e. the operating theatre, the patient, the surgical tools, by providing pre-operative information during the surgery. It increases the amount of information available during interventions and then ergonomics issues are becoming more and more crucial.

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 centered on the user's task focus need to be investigated.

5.1.2 Computer Aided Surgery

The major clinical objective of computer-aided surgery (CAS) is to help the physician to use quantitatively multi-modal information, in order to plan optimal therapeutic strategies, to simulate them and finally to execute them very accurately and safely in the context of minimally invasive procedures. The goals of CAS systems are to enhance the dextetry, visual feedback, to complement and enhance the surgeon's skills, while always leaving him in control and never replacing him.

For the purpose of this work, the interest is restricted to navigation or Image Guided Surgery (IGS) systems (e.g. [Edwards et al., 2000]). These are Augmented Reality systems which may be thought of either as "surgical assistants" providing useful information to assist the surgeon's manual execution of a task or instruments supporting the execution of procedures planned from preoperative models (for example through the use of computer graphics overlays on the surgeon's field of view in augmented reality based solutions). Figure 5.2 shows the domain of discourse applied for the microscope image guided surgery [Trevisan et. al., 2003]. In a typical scenario after the patient is positioned for treatment in the operating room, calibration and registration are done to place all of the pre-op images and instruments into a common reference frame. As the surgery proceeds, a navigation workstation, a microscope display and a TV monitor are used to guide the procedure based on the original and processed images. The itra-operative navigation is the most complex interaction design once that the surgeon performs simultaneous tasks (e.g. surgical procedures and interaction with the system).

Augmented Reality display devices such as head mounted displays or microscopes also restrict the field of view for the user. The peripheral vision is known to play a significant role in the sensation of immersion and reality of a scene, and restricting the view can cause misjudgements in distance, object's sizes, the ability to perceive spatial layout and performing simple reaching and grabbing tasks. When the peripheral view is removed we loose context of features in the scene, the strongest contextual cue being the ground plane and horizon. These are two cues we use to judge the relative distance between objects in view and the absolute distance from ourselves to the objects. In surgery the surgeon is looking at a very restricted scene that has neither ground plane nor horizon visible. It often consists mostly of soft deformable tissue and offers no fixed point on which relative depth judgments can be made. The surgeon relies on the disparity information in the stereo overlay images and his knowledge of human anatomy to guide him.

As IGS systems are evolving towards multimodal navigation integrating atlas based features, interfaces must allow more sophisticated presentation strategies enabling the selection of annotated information to display, representation modes, multiresolution exploration of volumetric data, effective layout avoiding occlusion of priority zones (tumours, vital organs, etc.) and update of these decisions all centered on the user's task. Scientific visualization provides techniques for texturing, contour rendering and preattentive multivariate visualisation, while view management of dynamic scenes automatic label placement are investigated in computer graphics.

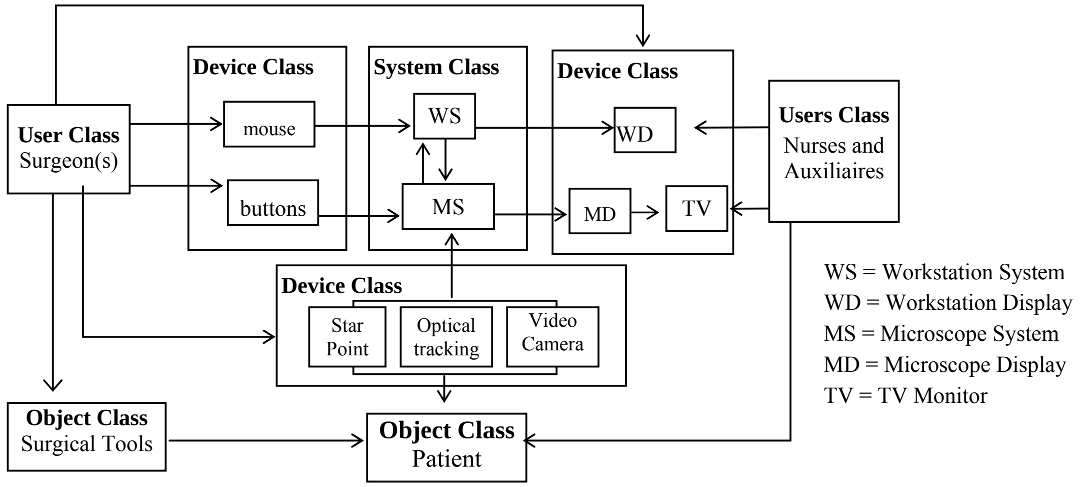

Figure 5.2. Diagram block of a microscope image-guided surgery. The arrows represent the different interactions between entities [Trevisan et. al. 2003].

5.1.3 Surgical simulators

Laparoscopic surgery has very important advantages over open surgery. It minimises tissue trauma and suffering, which leads to short recovery times and cost reduction. However, it requires a long traineeship period in the operating room, which requires continuous personal supervision by an expert. There is also a lack of standards to train and accredit surgeons [Aggarwal et al. 2004].

In recent years, the developments of Virtual reality (VR) surgical simulators have been increased. Virtual environments offer numerous advantages over the traditional learning process:

- They can display a wide variety of different pathologic conditions.
- They allow a trainee to practice at any time and as many times as required.
- They can monitor the skill in real-time, providing constructive feedback to the student as the teacher does.

In short, Virtual reality (VR) simulators are a valuable tool for training and skills assessment [Aggarwal et al. 2004]. In addition, their use permit direct comparison of performance between surgeons since different users can complete exactly the same task, without the effects of patient variability or disease severity.

5.1.3.1 Human Computer Interaction

In minimally invasive surgery the surgeon does the intervention with the laparoscopic tools. Due to the characteristics of these instruments and its limitations related to the restricted movements, only reduced perceptive information will reach the surgeon [Sjoerdsma et al., 1997]. In addition, no 3D-visual information is available; instead only 2D-visual information originating from the laparoscope is feedback. In Figure 5.3, is represented the Block diagram of the minimally invasive surgical process proposed by Stassen et al. [Stassen et al, 2001].

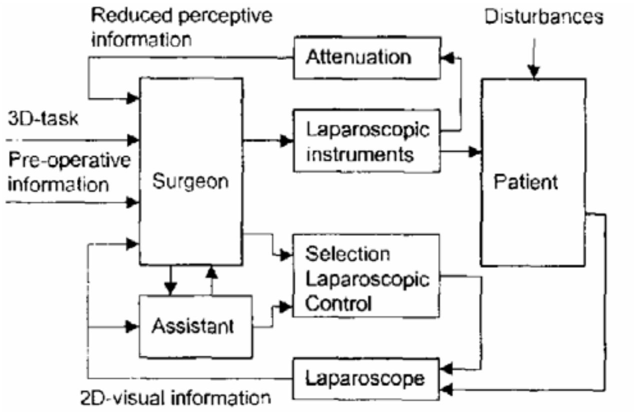

Figure 5.3. Block diagram of the minimally invasive surgical process [Stassen et al, 2001].

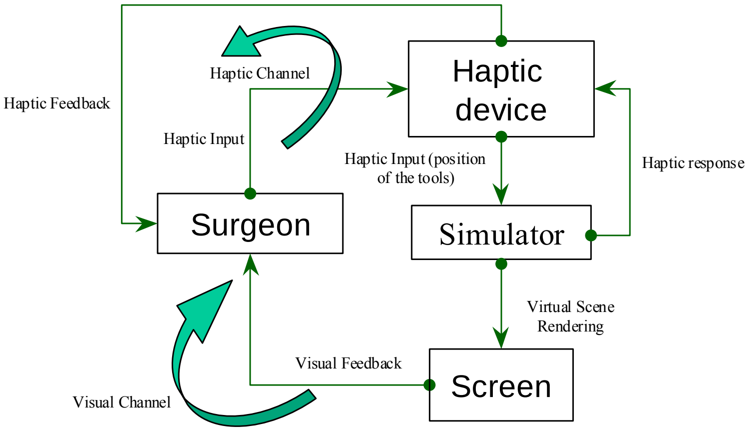

Figure 5.4. Block Diagram of the simulation process.

In the case of VR simulator, the diagram is similar since the interaction interfaces are well defined, the Laparoscopic tools and the screen.

The user interacts through two sensorial channels at the same time: visual and haptic. When he moves a virtual tool with the haptic device, this tool is moved in the monitor and interacts with the virtual organs, and he feels the interaction forces. And all this is performed several times in a second (visual update rate: 15Hz, haptic update rate: 500Hz). The relations of these elements are represented in Figure 5.4.

5.1.3.2 Different degrees of fidelity

One controversial aspect in surgical simulation is the degree of fidelity required in order to provide an effective learning/assessment tool. Several solutions have been provided during last years, ranging from simple box trainers (physical objects manipulated with real objects) to quite advanced virtual reality simulators. Although physical simulators are not virtual/augmented applications (the scope of this chapter), they have been covered in this section as they also need a validation process to be used with confidence.

Three different simulators have been chosen as a representative sample of what's available today in the market:

- MISTELS: McGill Inanimate System for Training and Evaluation of Laparoscopic Skills [Fried et al. 2004]. This is a physical simulator deeply studied and validated.
- MIST-VR: Minimially Invasive Surgery Trainer – Virtual Reality (Mentice AB, Sweden) is a very simple virtual reality simulator, but has been thoroughly validated (see Figure 5.5)
- LapMentor (Symbionix, Israel) is a quite advanced virtual reality simulator which enables a user to practice complete surgical procedures like a colicestectomy (see Figure 5.6)

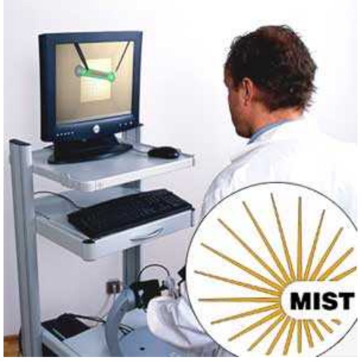

Figure 5.5. MIST-VR.

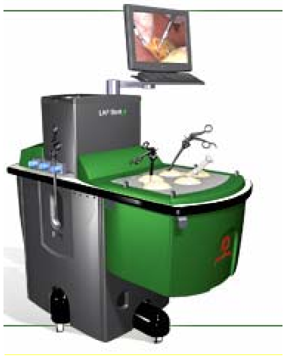

Figure 5.6. LapMentor.

5.2 Evaluation of augmented-reality applications

5.2.1 General guidelines to Augmented Reality Systems

In [Furmanski et al. 2002] they have generated some a priori guidelines based on the major perceptual issues involved with OIV (Obscured Information Visualization) systems but may also generalize to other types of AR and VR displays. Important design guidelines include:

- *Distance conveyance* – In OIV environments, distance and absolute location can be confusing, so AR renderings should disambiguate information about distance or position.
- *Proper motion physics* - For dynamic displays, motion parallax is an important cue to human observers, so it is important that the information in depth move in such a way as to convey its proper position. This can be achieved with properly defined geometries and metrically-accurate models of the environment that rotate and move in realistic ways.
- *Eliminate unneeded AR motion* – Because the human visual system is so sensitive to motion, unneeded motion of rendered material (e.g., the slowly moving self-organization of rendered AR tags) as well as mis-registration should be eliminated or minimized as much as possible.
- *Selective or multiple cues* – Because the accuracy of different perceptual cues vary with distance, specific perceptual cues (e.g., motion parallax or size-constancy/scaling) should be selected if displays are operating in an environment of a limited range of depths/distances. Multiple perceptual cues should be integrated if the ranges of display depth/distances are variable.

There are many other factors that do not deal directly with perceptual influence on OIV, but are important features to cover, including:

- *Define rule space* – Another important factor for OIV system development is to define the conditions under which augmented information should be displayed. For example, if the viewer knows that augmented information will only be presented on hallway surfaces or only within the confines of the building, then these rules can help disambiguate otherwise vague location information.

Carefully defining a series of rules, such as having different color or symbols to designate particular locations, will go a long way towards building an unambiguous OIV system. Even implicit rules or context clues, such as augmented messages only appear in one's office, also reduce positional uncertainty. So, while more detailed rules increase the cognitive complexity for the human user, more sophisticated rule systems can drastically reduce the ambiguity in OIV displays.

- *Effectiveness testing* – While these general guidelines can be used as a starting point, some form of experimental testing (whether pilot testing or more involved experimental procedures) should be incorporated through the design development. Such an empirical approach can improve overall system design as certain specific perceptual cues may be better suited for certain applications. Multiple cues may interact in a way that could be quantified and subsequently compared to other designs. The instructions that participants get in experiments should be created very carefully so as not to deliberately bias subjects' responses.

Finally, the design development of similar systems could be improved using other formalized usability-engineering processes that provide a structured evaluation technique [Gabbard et. al., 2002; Hix et. al., 2004].

5.2.2 Validation criteria in Image Guided Surgery

Validation requires the application of defined criteria to a device or process. Common examples of validation criteria which may be applicable to IGS include [Jannin et. al. 2002]:

Accuracy: Goodman [Goodman, 1998] defines accuracy as the "degree to which a measurement is true or correct". For each sample of experimental data local accuracy is defined as the difference between computed values and theoretical values, i.e., known from a ground truth. This difference is generally referred to as local error. Under specific assumptions, a global accuracy value can be computed for the entire data set from a combination of local accuracy values.

Precision and Reproducibility or Reliability: Precision of a process is the resolution at which its results are repeatable, i.e., the value of the random fluctuation in the measurement made by the process. Precision is intrinsic to this process. This value is generally expressed in the parameter space. Goodman defines reliability as "the extent to which an observation that is repeated in the same, stable population yields the same result".

Robustness: The robustness of a method refers to its performance in the presence of disruptive factors such as intrinsic data variability, pathology, or inter-individual anatomic or physiologic variability.

Consistency or Closed Loops: This criterion is mainly studied in image registration validation [Fitzpatrick, 2001], by studying the effects of the composition of n transformations that forms a circuit: Tn1 ° … ° T23 ° T12. The consistency is a measure of the difference of the composition from the identity. This criterion does not require any ground truth, but the onus is on the user to convince that there is no bias in the error estimates obtained.

Fault Detection: This is the ability of a method to detect by itself when it succeeds (e.g. result is within a given accuracy) or fails.

Functional complexity and computation time: These are characteristics of method implementation. 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).

The requirement of an overall validation of image-guided surgery systems (i.e. including all its components) should also be taken into account by estimating errors at each stage of the IGS process and by modelling how errors propagate through the entire IGS process.

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

As any mixed reality application, IGS systems are characterized by a continuous process of exchange of information at a high resolution and multiple worlds of interaction. When there are multiple sources of information and two worlds of interaction (real and virtual) designers must make choices about what to attend to and when. Users need to focus exclusively on a single item at a time without interference from other items. They occasionally need to time-share or divide their attention between two (or more) items of interest that can be part of the same or different worlds.

Intra-operative IGS annotation support can be characterized as mixed reality focused applications. These systems enhance interaction between the user and her/his environment by providing additional capabilities for perception of the surroundings. They also require the user's hand free for real world tasks and user's focalized attention in limited, well-defined areas of the workplace. The user of our system is immersed within a full-scale physical environment, which is viewed through a stereo-video head mounted display or through a microscope display.

The predetermined metric chosen in the design of the mixed interaction to evaluate the usability is continuity [Dubois, et. al. 2002; Trevisan, et. al., 2004]. This is the capability of the system to promote a smooth interaction scheme during task accomplishment considering perceptual, cognitive and functional properties. The perceptual property 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 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 functional property is related to the effort of the user in experiencing a new interaction mode.

Thus support for Mixed Reality focused applications raises the need to express the notion of a plastic mixed system. The main characteristic of a plastic mixed system is that augmentation of the physical environment, which can enhance interaction, user's action or user's perception, may be dynamically tuned to accommodate a wide range of context of use while preserving predefined usability criteria. In Mixed Reality focused applications links between the physical and digital worlds are statically defined, but the way augmentation is performed depends on context, the user's focus of interest and it is constrained by the predefined properties to be preserved in the change. The adoption of the plasticity framework in the design of Mixed Reality focused applications allows to quickly point out potential critical contexts of use that may involve interactional ruptures or instability in the interaction process.

On going work on the development of targeted scenarios can thus be set up to assess the user experience and early design validation taking account the continuity of the interaction [Trevisan, et. al. 2004].

5.2.3 First experimental results

See-through AR in surgery is still in the experimental phase of research and its benefits speculative. In the small number of systems developed worldwide so far, it is assumed, but not always tested that stereo overlays provide accurate and unambiguous depth information to the surgeon (assuming a good calibration).

In [Johnson et. al. 2004] they have empirically tested this assumption and found that errors in perception still arise despite accurate stereo displays. The most significant findings of the work provided evidence that the presence of a physical surface, placed between the observer and the virtual object, caused a misperception of the virtual object's depth. On average the error induced by the surface was positive causing an underestimation of the depth of the virtual cone. This is consistent with the work of Ellis [Ellis and Menges, 1998]6 who also showed an underestimation of judged depth using a virtual object presented behind a physical surface. Any interaction, in the form of overlap between the virtual cone and physical surface will draw the observer's attention to both surfaces. The observer either has to simultaneously fuse the cone and physical surface, or divide visual attention between them. The observer then has to align the physical pointer with the cone (the most distal of the two surfaces), which results in an overestimation of its depth. With the cone tip coincident with the surface (0mm) exactly half the cone protrudes in front of the surface, and the eyes need only focus at one point. As the cone moves below the surface the percentage of cone surface protruding through the physical surface decreases but the separation between the points of focus increases. The size of the error in locating the cone relative to the surface increases until the cone no longer protrudes or contacts the physical surface (the tip is presented more than 15mm from the surface). It is possible that at this point either the observer's attention is no longer drawn to both surfaces simultaneously as they no longer overlap, or that the separation between the two objects is greater than the fusible range and they choose to focus their attention on the cone. With the cone at distances greater than 40mm below the surface, outside the predicted fusional area, observers underestimate the depth in accordance with the theory of occlusion and proximal vergence effects.

In [Vogt et al. 2004] the augmented reality navigation system prototype was adapted to Magnetic Ressonance-guided needle placement procedures. The table of the MR scanner and the needle are equipped with retroreflective markers, which are tracked by a head-mounted infrared camera. The video-see-through AR system overlays medical images, virtual guidance information, and the needle graphics onto the live stereoscopic video view. During insertion, the needle can be observed virtually at its actual location in real-time, supporting the interventional procedure in a very efficient and intuitive way. The usability criteria investigated in this study was the precision and robustness of the AR needle guidance on Natrosol gel phantoms with embedded target structures of 12mm and 6mm diameter. As results the system can significantly reduce procedure time for AR guided interventions, while at the same time increasing the precision.

5.3 Evaluation of surgical simulators

As explained before, a surgical simulator can be designed with two main aims: for surgical training or for the assessment of technical surgical skills. Therefore the state of the art in the evaluation of surgical simulators has been decomposed in these two features. Three main recent publications can be taken from the literature as a review in the field of laparoscopic surgery [Feldman et al. 2004; Moorthy et al. 2003; Aggarwal et al. 2004]. The reference simulators are the MIST-VR as a virtual one, and the MISTELS [Fried et al. 2004] as a physical one. Both of them have demonstrated to be a valid training and assessment tool. Usability aspects of the user interaction have been suppressed, as they are already well design in commercial available simulators.

5.3.1 Simulator as a skill assessment tool

A simulator can be an examination tool for surgeons. Then it incorporates representative tasks of different technical skills, with different evaluation metrics defined to assess the performance of surgeons. It can even automatically acquire this evaluation metrics, like virtual reality simulators do with time, movements or errors. But other times an expert surgeon is needed to assess them.

When a simulator is designed to assess technical surgical skills it must be reliable, valid and feasible. Then it will be an examination tool to be used with confidence [Aggarwal et al. 2004].

5.3.1.1 Reliability

This is a measure of the consistency of a test, the extent to which the assessment tool yields the same results when used repeatedly under similar conditions. It supposes no learning between the two tests. It is measured by a reliability coefficient, which is a quantitative expression of the reliability of the tests and ranges between 0 and 1. A good reliability coefficient has been approximated as values >0.8. Although lower values (i.e., <0.7) have been reported, they are generally frowned on in the behavioral sciences. Other useful measures of reliability are α, coefficient α, Cronbach's α, or internal consistency [Gallagher et al. 2003a].

Two different aspects are involved:

I) Inter-rater Reliability: it determines the extent to which two different evaluators (raters) give the same score in a test made by a user. Example: two surgeons evaluate a student performing a simulated procedure and both agree on the overall performance score ($p > 0.80$). This feature has little interest in simulators when they automatically acquire the evaluation metrics, like Virtual Reality simulators.

II) Test-retest Reliability: it determines the extent to which two different tests made by the same person in two different times give the same result. Example: students are tested twice on the same test and get equivalent scores each time.

Another common method to establish the reliability is the **split-half method**, for which test items from a single test occasion are split and then the internal consistency of the assessed items is calculated.

One of the main problems in the literature with this kind of studies is that researchers and clinicians are too likely to conclude that a reliable test is ipso facto a good test for any purpose they have in mind [Gallagher et al. 2003a].

5.3.1.2 Validity

This concept relates to the property of "being true, correct, and in conformity with reality". In testing, the fundamental property of any measuring instrument, device, or test is that it ''measures what it purports to measure''. Therefore, validity is not a simple notion; rather, it is comprised of a number of first principles. The result is that within the testing literature, a number of validation benchmarks have been developed to assess the validity of a test or testing instrument. These include face validity, content validity, construct validity, concurrent validity, discriminate validity, and predictive validity [Gallagher et al. 2003a].

I) Face validity: is defined as ''a type of validity that is assessed by having experts review the contents of a test to see if it seems appropriate''. Simply stated, experts review the tests to see if they seem appropriate 'on their face value'. It is a very subjective type of validation and is usually used only during the initial phases of test construction. For example a simulator has face validity when the chosen tasks resemble those that are performed during a surgical task.

II) Content validity: is defined as ''an estimate of the validity of a testing instrument based on a detailed examination of the contents of the test items''. Experts perform a detailed examination of the contents of the tests to determine if they are appropriate and situation specific. Establishing content validity is also a largely subjective operation and relies on the judgments of experts about the relevance of the materials used. For example a simulator has content validity when the tasks for measuring psychomotor skills are actually measuring those skills and not anatomic knowledge.

III) Construct validity: is the determination of the degree to which the test captures the hypothetical quality it was designed to measure. A common example is the ability of an assessment tool to differentiate between experts and novices performing a given task.

IV) Concurrent validity: is defined as ''an evaluation in which the relationship between the test scores and the scores on another instrument purporting to measure the same construct are related''. It would be used when introducing a new assessment tool to replace a pre-existing ''gold standard'' assessment tool.

V) Discriminate validity is defined as ''an evaluation that reflects the extent to which the scores generated by the assessment tool actually correlate with factors with which they should correlate''. This is a much more sophisticated analysis that requires that these factors that should correlate highly actually do correlate highly, and that the factors that should correlate poorly do demonstrate a poor correlation. An example of this would be an assessment tool that could differentiate ability levels within a group with similar experience, such as discriminating abilities of all the residents in postgraduate year 1.

VI) Predictive validity: is defined as ''the extent to which the scores on a test are predictive of actual performance''. An assessment tool used to measure surgical skills will have predictive validity if it predicts who will perform actual surgical tasks well and who will not.

Currently there is no consensus regarding the optimal assessment tool for laparoscopic procedures, and studies have been focused on construct validity [Aggarwal et al. 2004]. All of these validation strategies have merit; however, predictive validity is the one most likely to provide clinically meaningful assessment [Gallagher et al. 2003a].

5.3.2 Simulator as a training tool

A simulator can be used as a training tool. It then incorporates representative tasks of different technical skills in a controlled environment that emulate a real situation. The question is whether this device with its training strategy actually trains or not the skill is supposed to. Several methodologies have been developed to answer this.

5.3.2.1 Face validity

In an interpretation of what is face validity for an assessment tool, validation studies of training tools have been done with this subjective methodology. One example is the study of the Xitact LS500 system [Schijven et al. 2002]. These results can be useful at the early stage of design of the simulator.

5.3.2.2 Concurrent validity

Validity of a tool can be proven when its results are similar to existing validated tools. The training outcome of different simulators has been compared, like the study of MIST-VR and a pelvic trainer that were similar [Kothari et al. 2002]. The limitation of this approximation is that there is no gold-standard training tool to be compared with.

5.3.2.3 Transfer of skills to Operating Room

This is the methodology that can show clinically useful learning results from simulator use. With prospective, randomized and blinded surgical trials novice surgeons are trained in different ways, and results can actually demonstrate how the skills acquired in a simulated environment are transferred to the operating room, to real surgery.

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].

5.3.2.4 Learning curves

A learning curve is a plot of the acquisition of skills along time, measured by different metrics like dexterity, time or errors. If these metrics have been shown to be valid (see former section), a learning curve is a proof of how trainees acquire technical skills.

Simple box trainer (physical simulator) have demonstrated how the learning curve for operator speed is shorter than that for operator accuracy [Smith et al. 2001]. The MIST-VR simulator has shown how novice surgeons improved their performance up to the experts level by practising on it [Gallagher et al. 2002].

5.3.3 Difficulties and future research

A critical issue in the design of simulators for medical training is the relationship between technology and training effectiveness [Kneebone 2003]. A key concern is the level of realism, i.e. fidelity, necessary for proper training [Liu et al. 2003]. A surgeon would ask the maximum level of realism for a simulator to be effective, which might be related with the immersion sensation that he/she expects. However, human beings have perceptual limitations of the sensory, motor and cognitive system. And not always an increment in fidelity leads to an improvement in teaching capability, as it has been shown in the field of endourological skills [Grober et al. 2004]. Serious consideration must be given to the human-factor strengths and limitations of the surgeon [Gallagher et al. 2003b]. Studies are recently being focused on the study of the sensorial interaction of the surgeon to answer these questions [Lamata et al. 2004].

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].

Potential benefits of surgical simulators are very encouraging. In flight aviation pilots are trained with a “zero-time training” in real aircrafts thanks to the development of simulation, and this is an important goal to be reached in surgery. Simulators currently have the ability to teach basic laparoscopic skills, enabling novice surgeons to progress along the early part of the learning curve before entering the operating theatre. With further developments in technology it may be possible to practise complete procedures, such as Nissen fundoplication and colectomy [Aggarwal et al. 2004].

Research is also being focused on the development of a credential process, with the development of an international benchmark for trainee surgeons [Aggarwal et al. 2004]. In essence, relying solely on the currently available simulators would be like judging a surgeon’s competence in the operating room by how he or she ties a knot. As the technology becomes increasingly sophisticated, it may become an integral component the credentialing of new surgeons, or the revalidation of practicing surgeons; however, this point remains several years in the future [Feldman et al. 2004].

5.4 Conclusion

Evaluation studies are being carried out to validate the purpose of brand new technical solutions in medical applications, which are enabling new ways of interaction or training of physicians. First results are being given, but generally there is a lack of accepted methodology.

Even mixed reality systems applied to medicine field have demonstrated their potential usefulness through improved planning, training, execution precision and complication reduction for patients, users interacting with such interfaces have frequent difficulties, which can result in frustration and overall low system usability.

In the effort of developing improved solutions, there is a need not only for novel algorithms such as near-real time methods to extract information from interventional images and adapt it to prior patient models, but to conceive new styles for user interaction and information presentation matching the skills, experience and expectations of the user, especially for interaction in the operating room.

Surgical simulators offer a controlled environment where physicians can be trained or their skills been assessed. Training systems have to demonstrate how skills acquired in simulated environments are transferred to the operating room, but little studies have been done up to now. A simulator designed to evaluate technical skills must be reliable and valid, and systems like MIST-VR or MISTELS have recently demonstrated this in the field of laparoscopic surgery.

Medical applications of virtual/augmented interaction systems are really promising, offering interesting possibilities like accreditation or mission rehearsal. But they are lacking of validation studies and even methodology, what is crucial in every medical application.

6 State-of-the-art in tools for remote usability evaluation

This section is dedicated to tools for remote usability evaluation. The goal is to explain the motivations for such an approach, the possible techniques and the role that automatic tools can play in order to support it.

Remote evaluation means to perform an evaluation where users and evaluators are distant in space and/or time. 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.

With the advent of the Web and the refinement of instrumentation and monitoring tools, user interactions are being captured on a much larger scale than ever before. Multimodal systems (either through the Web or other environments) add a new dimension of complexity, since different input/output channels have to be monitored at the same time. Automated support for the capture, representation, and empirical analysis of user behaviour is leading to new ways to evaluate usability and validate theories of human-computer interaction. It enables remote testing, allows testing with larger numbers of subjects, and motivates the development of tools for in-depth analysis. Data capture can take place in a formal experimental setting or on a deployed system. Particular attention must be paid to remote usability evaluation, where users and evaluators are distant in time and/or space [Hartson *et al.* 1996]. This type of approach can overcome some limitations of usability laboratories: it is often difficult to bring a large number of users to such laboratories, and then they have to interact with the application in an environment different from their daily working environment. In practice, lab-based tests are mostly performed with a small, local sample of the user population, which renders them inadequate for the evaluation of products to be used by people across a wide geographical area and various demographic groups. In single modal applications, remote techniques currently allow evaluation even after the first version of an application has been deployed, and such techniques are needed for multi-modal applications as well. A few additional drawbacks of current approaches and systems are that current systems only measure the user's task performance (actions and their outcomes), but not the user's physical and mental state (e.g. physiology, stress level, attention). In addition, usability test systems, especially automated systems, can generate loads of data that potentially contain valuable user interaction patterns. However, conventional data analysis, data mining and statistical methods are unable to uncover these patterns; thus, better statistical analysis techniques and tools are needed.

A remote support for usability evaluation can be performed in various ways:

- users self-reporting critical incidents encountered in real tasks performed in their normal working environment [Hartson and Castillo 1998];
- use of teleconferencing tools to observe user behaviour in real-time from a remote location;
- instrumented or automated data collection for remote evaluation, where tools are used to collect and return a journal or log of data containing indications of the interactions performed by the user.

These data are analysed later on, for example using pattern recognition techniques; however, usually the results obtained are rather limited for the evaluation of an interactive application.

For example, use of logging tools to store user events such as keystrokes and mouse movements or web pages selected and then this information is analysed. This can be done using various techniques, for example, in [Siochi and Ehrich 1991] pattern recognition techniques are used to identify where usability problems have occurred.

6.1 Automatic evaluation

One important benefit derived from automation testing is incorporating evaluation within the design phase of development; this is important because evaluation with most non-automated methods can typically be done only after the interface or prototype has been built and changes are more costly [Nielsen 1993]. Modelling and simulation tools make it possible to explore designs earlier. It is important to consider automation as a useful complement to standard techniques, and not as a substitute.

Using Balbo's automation taxonomy [Balbo 1995], we can have:

- *None*: no level of automation; all aspects of usability evaluation are performed only by evaluators.
- *Capture*: The first phase of evaluation is supported by a software tool, which automatically records in log files usability data.
- *Analysis*: since captured data, a specific tool analyses them and automatically detects potential usability problems.
- *Critique*: automating this step is somewhat difficult. Software automates analysis and suggests improvements.

Considering assessments of automated capture, analysis and critique techniques, the following criteria can be used:

- *Effectiveness*: how well a method discovers usability problems;
- *Ease of use*: how easy is a method to employ;
- *Ease of learning*: how easy is a method to learn;
- *Applicability*: how widely applicable is a method to WIMP and/or Web Uis other than those originally applied to.

Although automated usability evaluation has great promise as a way to augment existing evaluation techniques, it is greatly under-explored. Ivory [Ivory and Hearst 2001] in her work surveyed 75 UE methods applied to WIMP interfaces, and 57 methods applied to Web UIs. Of these 132 methods, only 29 apply to both Web and WIMP UIs. Methods without automation support represent 67% of the methods surveyed, while methods with automation support collectively represent only 33%. Of this 33%, capture methods represent 13%, analysis methods represent 18% and critique methods represent 2%. All but two of the capture methods require some level of interface usage. To provide the fullest automation support, software would have to critique interfaces without requiring formal or informal use. Ivory's survey found that this level of automation has been developed for only one method type: guideline review.

Now, in the next paragraphs, we would like to describe how testing, inspection and inquiry for Web evaluation are conducted.

6.2 User testing methods

Usability testing with real participants is a fundamental usability evaluation method [Nielsen 1993; Shneiderman 1998]. It provides an evaluator with direct information about how people use computers and their problems with the interface being tested. During usability testing, participants use the system or a prototype to complete a predetermined set of tasks while the tester records the results of the participants' work. The tester then uses these results to determine how well the interface supports users' task completion as well as other measures, such as number of errors and task completion time. Thus, user interface events are elements which can be captured and analysed in order to measure and detect usability issues. User interface events (UI events) are generated as natural products of the normal operation of window- based user interface systems such as those provided by the behaviour with respect to the components that make up an application's user interface (e. g., mouse movements with respect to application windows, keyboard strokes with respect to application input fields, mouse clicks with respect to application buttons, menus, and lists). Because such events can be automatically captured and because they indicate user behaviour with respect to an application's user interface, they have long been regarded as a potentially fruitful source of information regarding application usage and usability. However, because user interface events are typically extremely voluminous and rich in detail, automated support is generally required to extract information at a level of abstraction that is useful to investigators interested in analysing application usage or evaluating usability. Automation has been used predominantly in two ways within user testing: automated capture of use data and automated analysis of this data according to some metrics or a model.

6.3 Automated capture phase

Automated capture methods represent important first steps toward User Interface improvements, because they provide input data for analysis and, in the case of remote testing, enable the evaluator to collect data for a larger number of users than traditional methods. When evaluators assess the interface usability, they have to collect the user actions in order to examine them. This can be done by an evaluator taking notes while the participant uses the system, either live or by repeatedly viewing a videotape of the session. Because both are time-consuming activities, automation is used as it represents a useful instrument. Within the user testing class of UE, automated capture of usage data is supported by two method types: performance measurement and remote testing.

Performance measurement methods record usage data (e.g., a log of events and times when events occurred) during a user test. Video recording and event logging tools are available to automatically and accurately align timing data with user interface events. Such evaluation requires many efforts and much time by evaluators who have to examine video-types and records. Furthermore, if the user actions are recorded in log files, data records produce voluminous log files and make it difficult to map recorded usage into high-level tasks.

Another way to conduct the capture phase is remote testing. Remote testing methods enable testing between a tester and participant who are not co-located. In this case the evaluator is not able to observe the user directly, but can gather data about the process over a computer network. Remote testing methods are distinguished according to whether or not a tester observes the participant during testing or not: Same-time different-place and different-time different-place are two major remote testing approaches used in UE methods. In the first case, the tester observes users exercising the application; software makes it possible for the tester to interact with the participant during the test, which is essential for techniques like the question-asking or thinking aloud protocols that require such interaction. Otherwise, the tester does not observe the user during different-time different-place testing.

A method type of this approach is the journaled session [Nielsen 1993], in which software guides the participant through a testing session and logs the results. Web servers maintain usage logs and automatically generate a log file entry for each request (e.g. the IP address of the requester, request time, name of the requested Web page, etc). Since 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 [Etgen and Cantor 1999], client-side logs are used. Client-side logs capture more accurate, comprehensive usage data than server- side logs because they allow all browser events to be recorded. This approach requires every Web page to be modified to log usage data, or else use of an instrumented browser or special proxy server. Examples of tools which capture client- side usage data, are The NIST WebMetrics tool suite (WebVIP, VISVIP and WebCAT), WebSat, WebRemUsine [Paganelli and Paternò 2003].

6.4 Automated analysis phase

After having captured user data, we have to analyze them. Approaches which we can use in order to analyse log files are:

- *Metric-based approaches*: they generate quantitative performance measurements. In general, performance measurement approaches focus on server and network performance, but provide little insight into the usability of the Web site itself. Service Metrics' tools, for example, can collect performance measures from multiple geographical locations under various access conditions, (e.g. performance bottle-necks, such as slow server response time), that may negatively impact the usability of a Web site.
- *Pattern-matching approaches*: These approaches analyse user behaviour captured in logs, for examples, detecting repeated user actions (e.g., consecutive invocations of the same command and errors), that may indicate usability problems. However, in several evaluation methods, pattern is matched in conjunction with task models.
- *Task-based approaches*: These approaches analyse discrepancies between the designer's anticipation of the user's task model and what a user actually does while using the system. For instance, USINE [Lecerof and Paternò 1998] employs the ConcurTaskTrees notation to express temporal relationships among UI tasks (e.g., enabling, disabling, and synchronization). Using this information, USINE looks for precondition errors (i.e., task sequences that violate temporal relationships) and also reports quantitative metrics (e.g., task completion time) and information about task patterns, missing tasks and user preferences reflected in the usage data. So, USINE processes log files and outputs detailed reports and graphs to highlight usability problems. Remusine [Paternò and Ballardin 1999] is an extension that analyzes multiple log files. WebRemUSINE [Paganelli and Paternò 2003] is a more recent extension of this approach to evaluating Web sites. The tool also provide quantitative measurements regarding task performance.
- *Inferential Analysis*: Inferential analysis of Web log files includes both statistical and visualization techniques. Statistical approaches include trace-based and time-based analysis. However, statistical analysis is largely inconclusive for Web server logs, since they provide only a partial trace of user behaviour and time estimates may be skewed by network latencies. Visualization is also used for inferential analysis. It enables evaluators to filter, manipulate, and render log file data in a way that ideally facilitates analysis. Visualizations provide a high-level view of usage patterns (e.g., usage frequency, correlated references, bandwidth usage, HTTP errors and patterns of repeated visits over time) that the evaluator must explore to identify usability

problems. However, there is no discussion of how effective these approaches are in supporting analysis.

One of the main application areas for remote evaluation is the Web. 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. For these reasons, interest in automatic support for usability evaluation of Web sites is rapidly increasing [Scholtz et al. 1998; Card et al. 2001]. In addition, 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.

6.5 WebRemUSINE

Our approach combines two types of evaluation techniques that usually are applied separately: empirical testing and model-based evaluation. In empirical testing the actual user behaviour is analysed during a work session. This type of evaluation requires the evaluator to observe and record user actions in order to perform usability evaluation. Manual recording of user interactions requires a lot of effort thus automatic tools have been considered for this purpose. Some tools support video registration but also video analysis requires time and effort (in our experience it takes five times the duration of the session recorded) and some aspects of the user interaction can still be missed by the evaluator (such as rapid mouse selections).

In model-based evaluation, evaluators apply user or task models to predict interaction performance and identify possible critical aspects. For example GOMS (Goals, Operators, Methods and Selection rules) [John and Kieras 1996] has been used to describe an ideal error-free performance. Model-based approaches can be useful but the lack of consideration for actual user behaviour can generate results that do not agree with the real user behaviour.

It is important to compare models with empirically observed performance. To this end, the main goals of our work are:

- To support remote usability evaluation in which users and evaluators are separated in time and/or space;
- To analyse possible mismatches between actual user behaviour and the design of the Web site represented by its task model, in order to identify user errors and possible usability problems;
- To provide a set of quantitative measures (such as execution task time or page downloading time), for individuals and group of users, useful for identifying usability problems.

Since we perform remote evaluation without direct observation of the user interactions, it is important to obtain logs with detailed information. We have designed and implemented a logging tool able to record a set of actions wider than those contained in server logs, whose effectiveness is strongly limited because they cannot capture accesses to pages stored in the browser cache. Moreover, server logs completely miss local user interactions with the interface techniques (menus, buttons, fill in text, …).

In order to understand what the user goal is, during navigation, a list of all the possible activities supported by the site is displayed in a separate window. The user is supposed to select the target task from the list, which is derived from the task model corresponding to the actual design of the Web site.

WebRemUSINE compares the logs with the task model and provides results regarding both the tasks and the Web pages supporting an analysis from both viewpoints (Figure 6.1). The method is composed of three phases:

- *Preparation*, it consists in creating the task model of the Web site, collecting the logged data and defining the association between logged actions and basic tasks;
- *Automatic analysi*s, where WebRemUSINE examines the logged data with the support of the task model and provides a number of results concerning the performed tasks, errors, loading time. WebRemUSINE displays all results in various formats both textual and graphical.
- *Evaluation*, the information generated is analysed by the evaluators to identify usability problems and possible improvements in the interface design.

The architecture of our system is mainly composed of three modules: the ConcurTaskTrees editor (publicly available at http://giove.isti.cnr.it/ctte.html) developed in our group; the logging tool that has been implemented by a combination of Javascript and applet Java to record user interactions; WebRemUSINE, a Java tool able to perform an analysis of the files generated by the logging tool using the task model created with the CTTE tool.

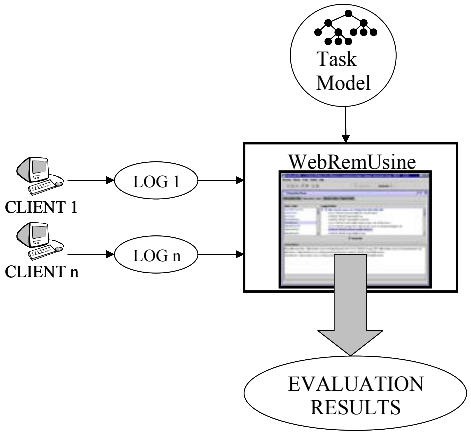

Figure 6.1: The Overall Architecture of WebRemUSINE.

Task models describe the activities to perform in order to reach user's goals. We have used the ConcurTaskTrees (CTT) notation [Paternò 1999] to specify them. CTT is a notation that provides a graphical representation of the hierarchical logical structure of the task model. In addition, the set of temporal and logical relations among tasks and task attributes that can be specified with ConcurTaskTrees is larger than the corresponding set in traditional hierarchical task analysis. In particular, it is possible to specify a number of flexible temporal relations among such tasks (concurrency, choice, enabling, disabling, suspend-resume, order-independence, optionality, …) and for each task it is possible to indicate the objects to be manipulated and a number of other attributes. The notation also allows designers to indicate how the performance of the task should be allocated (to the user, to the system, to their interaction) through different icons.

The logging tool stores various events detected by a browser, using Javascripts that are encapsulated in the HTML pages and executed by the browser. When the browser detects an event, it notifies the script for handling it. By exploiting this communication, the script can capture the events detected by the browser and add a temporal indication. Then, a Java applet stores the log files directly in the application server. Our browser logging tool overcomes the limitations of other approaches to logging: server logs have limited validity since various page accesses are hidden to them because of browser cache memory and they do not detect the local interactions with the user interface elements (check-boxes, type in fields, …) that also in the case of proxy-based approaches [Hong and Landay, 2001] cannot be detected. Our tool works for the two main Web browsers (Microsoft IE and Netscape Communicator).

The WebRemUSINE analysis can detect usability problems such as tasks with long performance or tasks not performed according to the task model corresponding to the Web site design. The task model describes how activities can be performed according to the current design and implementation. Either the designer or the evaluator can develop it. Since ConcurTaskTrees has various temporal and logical operators, it is possible to describe all the various paths the user can follow to accomplish a task. Deviations from the expected paths can be detected in the logs and represent useful information for identifying any pages that create problems to the user. Moreover, the tool analysis provides information concerning both tasks (such as time performance, errors, …) and Web pages (such as download and visit times, …). These results allow the evaluator to analyse the usability of the Web site from both viewpoints, for example comparing the time to perform a task with that for loading the pages involved in such a performance.

WebRemUSINE also identifies the sequences of tasks performed and pages visited and is able to identify patterns of use, to evaluate if the user has performed the correct sequence of tasks according to the current goal and to count the useless actions performed. In addition, it is also able to indicate what tasks have been completed, those started but not completed and those never tried. This information is also useful for Web pages: never accessed Web pages can indicate that either such pages are not interesting or that are difficult to reach. All these results can be provided for both a single user session and a group of sessions. The latter case is useful to understand if a certain problem occurs often or is limited to specific users in particular circumstances.

We are currently working on further extending the tool in such a way to collect multimodal data regarding the user behaviour and thus providing better information for its analysis for usability evaluation purposes. In particular, we are extending it in such a way to collect data through Web cams and remote eye trackers.

6.6 Usability evaluation in European Projects

Some issues related to tools for usability evaluation have been considered in previous European projects, but with limited results and without solving the particular issues raised by multimodal systems (multimodality opens up new directions and affords new challenges for user experience engineering, rather than traditional usability engineering; user experience engineering will cover a wider scope of experience parameters). The first project in this area was MUSIC where only graphical applications were considered. In the U.S., the National Institute of Standards and Technology (NIST) has promoted a number of projects in this area. The objective of the NIST Web Metrics Testbed (http://zing.ncsl.nist.gov/WebTools/) is to explore the feasibility of a range of tools and techniques that support rapid, remote and automated testing and evaluation of website usability. The Web Metrics Testbed includes a number of prototypes that are publicly available. However, such tools are limited to simple metrics for Web site evaluation. A more recent project that has begun to address multimodal issues is NITE (Natural Interactivity Tools Engineering; http://nite.nis.sdu.dk). However, this project is more focused on speech annotation, while it does not consider automatic support and remote evaluation in any depth. The overall objective of the "iEye" project (http://www.cs.uta.fi/hci/ieye/) was to develop new, natural forms of interaction for multilingual applications and to evaluate their usability. Of primary interest is eye-based interaction, where the point of gaze is used as an input channel, either by itself or combined with other forms of input, such as speech. The chosen approach in “iEye” was to build two application prototypes and evaluate them. In SIMILAR we will target a more general solution to the problem of tracking multiple modalities of human behaviour in a robust, accurate and efficient (highly automated) way. Rather than individual pieces, we envisage an integrated environment based on an in-depth understanding of multimodal usability factors and their measurement in an integrated and synergetic fashion. In particular, integration means the strongest link possible between evaluation and design, so that interactive design can become a reality. This is only possible via a coordinated tool approach.

Lastly, it should be noted that no previous project has developed a set of tools able to support remote evaluation in a wide variety of contexts (home, office, on the move, …) or when users can interact through a variety of platforms and modalities.

7 Overview of SIMILAR applications

Sections 2-6 have addressed the state-of-the-art in usability evaluation of important areas of multimodal and natural interactive systems and in tools in support of remote usability evaluation of such systems. This section provides concrete examples of usability evaluation by presenting an overview of how the applications described in SIMILAR deliverable D17 have been or are planned to be usability evaluated. In particular we address which evaluation methods and usability evaluation criteria have been or will be used with these applications.

Seven applications were identified and described by SIMILAR SIG7 members in deliverable D17. Figure 7.1 provides an overview of these systems including a categorisation according to the areas of multimodal and natural interactive systems discussed in the above sections.

Application	Type of application	Main area
Portable Cicero Application	Portable museum guide	May benefit from remote usability evaluation
NICE Hans Christian Andersen	Non-task-oriented multimodal conversational edutainment system	Spoken dialogue system
AlterStationTM System	Gesture-based interactive entertainment system	Vision-based system
Training System for Blind People	Training system for blind people, e.g. cane simulation application	Haptics-based system
SINERGIA laparoscopy virtual simulator	Surgery system	Mixed reality surgery system
Image Guided Surgery	Surgery system	Mixed reality surgery system
Medical Studio	Surgery system	Mixed reality surgery system

Figure 7.1. Applications described in D17.

Six of the seven applications in Figure 7.1 may be viewed as concrete examples of applications which fall within the areas discussed in Sections 2-5.The first application (the portable museum guide) does not fall within any of the areas described in these four sections. However, the museum guide is a good example of a system that may clearly benefit from the application of tools for remote usability evaluation as described in Section 6, because an observer may disturb the user during the museum visit whereas intelligent analysis of automatic logs of user interactions can provide useful insights regarding user behaviour without being intrusive.

All seven applications are very new and most of them are still under development or they are being improved. This is reflected in the evaluation status, cf. Figure 7.2 which reveals that several of the systems have not really been evaluated yet. Figure 7.2 moreover provides an overview of the input and output modalities used by each system and the target users addressed by each system.

Application	Input modalities	Output modalities	Target users	Evaluation status
Portable museum guide	pen, infrared	speech video, map, text	museum visitors; no prior skills required	tested in target museum with real visitors
Non-task-oriented multimodal conversational edutainment system	speech, 2D pointing gesture	speech, 3D graphics (animated agent)	10-18 years old children and teenagers	first prototype tested with target users; second prototype under implementation
Gesture-based interactive entertainment system	image of user filmed by camera, 2D body movements captured by camera	graphics, acoustics	people interested in games – kids as well as adults	already commercialised but also still being improved
Training system for blind people	3D cursor or haptic glove	acoustics, speech, haptic feedback	visually impaired people	initial versions have been tested with target users
Laparoscopic virtual simulator	haptic input	graphics, haptic feedback	surgeons with little laparoscopic experience	under development, no real testing yet
Image-guided surgery system	text and selection data via keyboard and mouse	text, 3D graphics, video, acoustics including speech	surgeons in operating room	under development, no real testing yet
Medical studio	images (scanned) input via keyboard and mouse	graphics; possibility to use stereovision glasses	surgeons	under development, no real testing yet

Figure 7.2. Input and output modalities, target users, and evaluation status for the D17 applications.

7.1 Evaluation methods

Evaluation methods are normally general and do not depend on e.g. the particular type of system to be evaluated or its input and output modalities. They rather depend on the life cycle stage at which the system is evaluated. Thus some methods are well-suited for the early analysis and design phases, e.g. walkthroughs and mock-ups, others are meant to be used only when an implemented system is available, e.g. field tests, while others again may be used at several different stages during the life cycle process, e.g. interviews.

The evaluation methods applied to the systems described in D17 include field test, controlled laboratory test based on scenarios, questionnaire, and interview. These are all well-known methods and are frequently used at a development stage where the entire system or most of the system has been implemented and is ready for user testing. Figure 7.3 shows for each application which evaluation methods have been used or are planned to be used. In addition it provides information about the number of test users (if any and if the information is available), whether they are children or adults, and whether they must have a particular background. The last application (Medical Studio) is at an earlier stage than the other systems and there is very little information available about evaluation plans for or evaluation of this system.

Application	Evaluation method(s)	Number of test users
Portable museum guide	user-system interaction in museum; anonymous questionnaire after use of system	35 adults
Non-task-oriented multimodal conversational edutainment system	partially scenario-based user-system interaction in lab; interview after use of system; analysis of logfiles	18 kids and teenagers
Gesture-based interactive entertainment system	user tests with special focus in lab; analysis of logfiles	not mentioned
Training system for blind people (cane simulation application, interactive presentation application)	monitored scenario-based user-system interaction in lab (for both applications); questionnaire (cane simulation); interview to fill questionnaire (interactive presentation)	26 children and adults (cane simulation) 12 adults (interactive presentation)
Laparoscopic virtual simulator	system's face validity evaluated by expert physicians; two tests planned: 1. test with novice and expert users during surgery course workshop; 2. studies in training transfer from virtual environment to real operating rooms	-
Image-guided surgery system	user test will be carried out followed by questionnaire; task performance to be validated by surgeons	-
Medical studio	no real testing yet; no information on planned testing	3 developers and 2 surgeons have used the system

Figure 7.3. Evaluation method(s) and number of test users for the D17 applications.

7.2 Evaluation criteria

Applying an evaluation method to a system serves to collect data about the system and about the users' interaction with the system. The collected data may subsequently be analysed according to certain usability evaluation criteria. Such an analysis leads to an evaluation of the systems' usability. The choice of usability evaluation criteria determines which aspects of the system's usability will be evaluated. Evaluators may e.g. wish to do a broad usability evaluation covering many issues, or they may want to focus on a few issues which are found to be crucial to user acceptance and user satisfaction, or they may want to focus on new challenging evaluation issues which are not well-understood and which need further exploration and research.

The usability evaluation criteria which have been or will be applied to the applications described in D17 reflect different wishes regarding which and how many issues to evaluate, cf. Figure 7.4. For the laporoscopic virtual simulator the major question seems to be whether surgeons will learn anything from using it. Thus there is only one overall evaluation criterion mentioned, i.e. educational success. For the conversational edutainment system entertainment value and educational value are also listed as evaluation criteria, but they are only two among many other criteria. These other criteria include both commonly used evaluation criteria, such as ease of use which is listed for several of the other applications as well, and new challenging criteria, such as conversation success. So far task success, as mentioned for the training system for blind people, has been an ordinary criterion for task-oriented systems. However, when moving to non-task-oriented systems, the measuring of task success makes no sense. Instead new criteria involving new ways to measure interaction success must be defined. Conversation success is a new criterion and it is not yet entirely clear how to measure conversation success rate which means that the use of this criterion is a challenging exercise which may require new research.

Application	Usability evaluation criteria
Portable museum guide	quantity and quality of information provided, modality presentation, interaction with infrared devices, capacity to help users orient themselves in the museum, ease of use, human voice versus text-to-speech, novice/expert in use of PDA related ratings of system (utility, ease of use, interface, interaction with infrareds)
Non-task-oriented multimodal conversational edutainment system	conversation success, naturalness of user speech and gesture, output behaviour naturalness, sufficiency of the system's reasoning capabilities, ease of use of the game, error handling adequacy, scope of user modelling, entertainment value, educational value, user satisfaction, speech and gesture understanding adequacy, output voice quality, output phrasing adequacy, animation quality, quality of graphics, ease of use of input devices, frequency of interaction problems, sufficiency of domain coverage, number of objects users interacted with through gesture, number of topics addressed in the conversation
Gesture-based interactive entertainment system	reaction time of new users, understanding of tutorial, speed of understanding the system with or without tutorial and in relation to complexity of needed actions (difference between novice and expert digital game players), occupancy, efficiency of use
Training system for blind people (cane simulation application, interactive presentation application)	time to task completion, task success rate, users' opinion on performance and usability (including output modalities); difficulty of tasks (cane simulation) user performance, prior experience in haptic interfaces, ease of use, interest in application (interactive presentation)
Laparoscopic virtual simulator	educational success
Image-guided surgery system	perceptive and cognitive workload, performance
Medical studio	nothing reported

Figure 7.4. Usability evaluation criteria for the D17 applications.

Users' performance with the system is a commonly used criterion. For some of the systems in Figure 7.4 the criterion is listed directly (the training system for blind people and the image-guided surgery system) while for others performance is not mentioned directly as a criterion but may be measured via other criteria, e.g. several of the criteria for the gesture-based system, the conversational edutainment system and the museum guide provide indirect information about user performance, for example reaction time of new users, conversation success, frequency of interaction problems, and capacity to help users orient themselves.

Another frequently used criterion is user satisfaction. It is mentioned directly for the conversational edutainment system and for the training system for the blind. Indirectly the museum guide also includes criteria which may point to user satisfaction, e.g. utility and interface. For good reason the surgery systems have their focus on evaluating how useful the systems are. However, the degree of user satisfaction may probably influence how much surgeons learn and how well they perform and therefore might also be interesting to measure for these systems.

For some of the evaluation criteria for the museum system there is a distinction between whether users are novices or experts in using a PDA. By asking this question one may find out whether expert PDA users are more or less happy about the system than novice users. In other words one may learn whether certain prior skills may affect the users' evaluation of the system. Also the gesture-based entertainment system distinguishes between novice and expert game players when measuring the speed of understanding the system.

We have pointed out a number of commonly used evaluation criteria, such as ease of use, user performance, and user satisfaction, and illustrated them via the criteria listed for the systems in Figure 7.4. Others could be mentioned but we shall stop the illustration here.

In addition to the more general usability evaluation criteria there are also, of course, several criteria in Figure 7.4 which are special to the particular system evaluated and hence are not generally applicable across systems. For example, the criterion concerning interaction with infrared devices mentioned for the museum system is of course only useful for systems which include infrared communication. Similarly, in order to include an evaluation criterion concerning the scope of user modelling, the system under evaluation must have a user model.

Between the generally applicable criteria and the more special ones we may perhaps define a third group which includes general criteria which have to be specialised depending on the system. Success is an example of such a criterion. For the training system for blind people, the general criterion of success specialises to task success while for the conversational edutainment system it becomes conversation success, and for the laporoscopic virtual simulator it is educational success.

8 References

Aggarwal, R., Moorthy, K. and Darzi, A. Laparoscopic skills training and assessment. Br J Surg, 91(12):1549-1558, 2004.

Agus, M., Brelstaff, G. J., Giachetti, A., Gobbetti, E., Zanetti, G., Zorcolo, A., Picasso, B., Sellari Franceschini, S.: Physics-based burr haptic simulation: tuning and evaluation. 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS'04) March 27 - 28, 2004 Chicago, Illinois, USA,128-135.

André, E., Dybkjær, L., Minker, W. and Heisterkamp, P. (Eds.): Affective Dialogue Systems. LNAI 3068, Springer 2004.

Azar, F. S., Perrin, N., Khamene, A., Vogt, S. and Sauer, F.: User performance analysis of different image-based navigation systems for needle placement procedures. Medical Imaging 2004: Visualization, Image-Guided procedures, and Display, edited by Robert L. Galloway, Jr., Proceedings of SPIE Vol. 5367, 2004, 110-121.

Azarbayejani, A., Starner, T., Horowitz, B., and Pentland, A.: Visually Controlled Graphics. IEEE Transaction on Pattern Analysis and Machine Intelligence, June 1993, vol. 15, No 6, pp. 602-605.

Balbo, S.: Automatic evaluation of user interface usability: Dream or reality. In S. Balbo Ed., Proceedings of the Queensland Computer- Human Interaction Symposium (Queensland, Australia, August 1995). Bond University, 1995.

Bartz, D., Strasser, W., Guervit, O. et al.: Interactive and multi-modal visualization for neuroendoscopic interventions. Proc. of Eurographics/ IEEE TCVG Symposium on Visualization, Ascona, 2001.

Batliner, A., Fischer, K., Huber, R., Spilker, J. and Nöth, E.: Desperately Seeking Emotions: Actors, Wizards, and Human Beings. Proc. of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast, 2000, 195-200.

Beier, T. and Neely, S.: Feature-based image metamorphosis. Computer Graphics (Siggraph proceedings 1992), vol. 26, pp. 35-42.

Beringer, N., Kartal, U., Louka, K., Schiel, F. and Türk, U.: PROMISE - a procedure for multimodal interactive system evaluation. Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation, Las Palmas, 2002, 77-80.

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.

Bernsen, N.O.: User Modelling in the Car. In [Brusilovsky et al. 2003], 2003, 378-382.

Bernsen, N.O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D. and Mehta, M.: First Prototype of Conversational H. C. Andersen. Proc. of the International Working Conference on Advanced Visual Interfaces (AVI 2004), Gallipoli, Italy, 2004a, 458-461.

Bernsen, N.O., Dybkjær, L. and Kiilerich, S.: Evaluating Conversation with Hans Christian Andersen. Proc. of LREC, Lisbon, Portugal, 2004b, 1011-1014.

Bickmore, T. and Cassell, J.: Social Dialogue with Embodied Conversational Agents. In [van Kuppevelt et al. 2005], 2005.

Black, M.J. and Yacoob, Y.: Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion, International Journal of Computer Vision 25(1), 23-48, 1997.

Brusilovsky, P., Corbett, A. and de Rosis, F. (Eds.): User Modeling 2003. Proc. of the 9th International Conference, UM 2003, Johnstown, PA, USA, Springer Lecture Notes in Artificial Intelligence, Vol. 2702, 2003.

Bühler, D., Minker, W., Häussler, J., and Krüger, S.: Flexible Multimodal Human-Machine Interaction in Mobile Environments. Proc. of the ECAI Workshop on Artificial Intelligence in Mobile System (AIMS), Lyon, 2002, 66-70.

Burdea, G. C.: 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., 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.

Cavazza, M, Charles, F, Mead, S, Martin, O, Marichal, X, and Nandi, A.: Multimodal Acting in Mixed Reality Interactive Storytelling. IEEE MultiMedia, July/September 2004 (Vol. 11, No. 3) -p. 30-39

Chen L.S.: Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction. Ph.D. Thesis, University of Illinois at Urbana-Champaign, 2000.

Chen, L.S., Tao H., Huang, T.S., Miyasato, T., and Nakatsu, R.: Emotion Recognition from Audiovisual Information. Proc. IEEE Workshop on Multimedia Signal Processing, Los-Angeles, CA, USA, pp. 83-88, 1998.

Cinquin, P., Troccaz, J.: Model Driven Therapy. The instance of Computer Assisted Medical Interventions Methods Inf Med. 2003 ;42(2):169-76 Journal of Image Guided Surgery, 1995.

Cohen, I., Sebe, N., Chen, L., Garg, A. and Huang, T.: Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. In Computer Vision and Image Understanding, Special Issue on Face Recognition, Vol. 91, Issues 1-2, 2003, 160-187.

Cohen, P., McGee, D., and Clow, J.: The Efficiency of Multimodal Interaction for a Map-based Task. Proc. of the Applied Natural Language Processing Conference, Morgan Kaufmann, 2000, 331-338.

Cole, R., van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W. and Yan, J.: Perceptive Animated Interfaces: First Steps towards a New Paradigm for Human-Computer Interaction. In [van Kuppevelt et al. 2005], 2005.

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. Int. Journal of Computer Vision, 10(1):53--66, 1993.

De Silva, L.C., Miyasato, T. and Natatsu, R.: Facial Emotion Recognition Using Multimodal Information. Proc. IEEE Int'l Conf. on Information, Communications and Signal Processing, Singapore, pp. 397-401, 1997.

Dehn, D. and van Mulken, S.: The Impact of Animated Interface Agents: A Review of Empirical Research. Int. Journal of Human-Computer Studies 52, 2000, 1-22.

Demirdian, D and Darrell, T.: 3-D Articulated Pose Tracking for Untethered Diectic Reference. ICMI 2002, Pittsburgh, USA, October 2002.

den Os, E., de Koning, N., Jongebloed, H. and Boves. L.: Usability of a Speech-Centric Multimodal Directory Assistance Service. Proc. of the CLASS Workshop on Information Presentation and Natural Multimodal Dialogs, Verona, Italy, 2001, 65-69.

Donato, G., Bartlett, M., Hager, J., Ekman, P., and Sejnowski, T.: Classifying Facial Actions. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 974-989, Oct. 1999.

Dubois E., Nigay L., and Troccaz J.: Assessing Continuity and Compatibility in Augmented Reality Systems. International Journal on Universal Access in the Information Society, special issue on Continuous Interaction in Future Computing Systems, Vol.1, n°4, 2002, Constantine Stephanidis (ed), 2002, pp.263-273.

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

Dybkjær, L., Bernsen, N.O. and Minker, W.: Evaluation and Usability of Multimodal Spoken Language Dialogue Systems. Speech Communication, Vol. 43/1-2, Elsevier, 2004a, 33-54.

Dybkjær, L., Bernsen, N.O. and Minker, W.: New Challenges in Usability Evaluation -Beyond Task-Oriented Spoken Dialogue Systems. Proceedings of ICSLP'2004, Vol.3, Jeju Island, Korea, 2004b, 2261-2264.

Edwards G. F., Cootes T. F. and Taylor C. J.: Face Recognition Using Active Appearance Models, Proc. European Conf. Computer Vision, vol. 2, pp. 581-695, 1998.

Edwards, P.J., King, A.P., Maurer, Jr.C., de Cunha, D.A., Hawkes, D.J., Hill, D.L.G., Gaston, R.P., Fenlon, M.R., Jusczyzck, A.S., Strong, A.J., Chandler, C.L., and Gleeson, M.J.: Design and Evaluation of a system for microscope-assisted guided interventions (MAGI). IEEE Trans. Med. Imaging 19, 1082-1093, 2000.Eisert, P. and Girod, B.: Analyzing Facial Expressions for Virtual Conferencing. IEEE, Computer Graphics and Applications, 1998, vol. 18, no. 5, pp. 70-78.

Ekman, P. and Friesen, W.: Unmasking the Face: A guide to recognizing emotions from facial expressions. New Jersey: Prentice Hall, 1975.

Ekman, P. and Friesen, W.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto, CA, 1978.

Ekman, P., Friesen, W.V., and Hager, J.C.: The Facial Action Coding System. Second edition, Salt Lake City: Research Nexus eBook. London: Weidenfeld & Nicolson (world), 2002.

Ellis, R. E., Ismaeil, O. M., and Lipsett, M. G.: Design and Evaluation of a High-Performance Haptic Interface. In Robotica 14, 321-327.

Ellis,S.R. and Menges,B.M.: Localization of virtual objects in the near visual field. Human Factors 40, 415- 431, 1998.

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

Essa, I.A., Basu, S., Darrell, T. and Pentland, A.: Modeling, Tracking and Interactive Animation of Faces and Heads using Input from Video. Proceedings of Computer Animation June 1996 Conference, Geneva, Switzerland, IEEE Computer Society Press.

Essa, I.A., Darrell, T. and Pentland, A.: Tracking Facial Motion, Proceedings of the IEEE Workshop on Non-rigid and Articulate Motion, Austin, Texas, November, 1994.

Essa, I.A. and Pentland, A.: A vision system for observing and extracting facial action parameters. Proceedings of Computer Vision and Pattern Recognition (CVPR 94), pages 76-83, 1994.

Etgen, M. and Cantor, J.: What does getting WET (Web Event-logging Tool) mean for Web usability. In Proceedings of the 5th Conference on Human Factors & the Web, Gaithersburg, Maryland, June 1999. Available at http://www.nist.gov/itl/div894/vvrg/hfWeb/ proceedings/etgen-cantor/index.html

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., and 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)

Fitzpatrick, J.M.: Detecting failure, assessing success. Medical Image Registration, Hajnal JV, Hill DLG, and Hawkes DJ (ed.), CRC Press, June 2001.

Fried, G.M.: Simulators for laparoscopic surgery: a coming of age. Asian J Surg, 27(1):1-3, 2004.

Fried, G.M., Feldman, L.S., Vassiliou, M.C., Fraser, S.A., Stanbridge, D., Ghitulescu, G. and Andrew, C.G.: Proving the value of simulation in laparoscopic surgery. Ann Surg, 240(3):518-525, 2004.

Furmanski, C., Azuma, R., Daily M.: Augmented-Reality Visualizations Guided by Cognition: Perceptual Heuristics for Combining Visible and Obscured Information. IEEE International Symposium on Mixed and Augmented Reality (ISMAR'02), Darmstadt, Germany, 2002, 215-224.

Gabbard, J.L., Swan II, J.E.., Hix, D., Lanzagorta, M., Livingston, M., Brown, D. and Julier, S.: Usability Engineering: Domain Analysis Activities for Augmented Reality Systems. Proceedings of the Conference on The Engineering Reality of Virtual Reality 2002, SPIE (International Society for Optical Engineering) and IS&T (Society for imaging Science and Technology) Electronic Imaging 2002, January 24, 2002.

Gallagher, A.G., Satava, R.M.: Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. Learning curves and reliability measures. Surg Endosc, 16(12):1746-1752, 2002.

Gallagher, A.G, Ritter E.M.and.Satava, R.M.: Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg Endosc., 17:1525-1529, 2003a.

Gallagher, A.G. and Smith, C.D.: From the operating room of the present to the operating room of the future. Human-factors lessons learned from the minimally invasive surgery revolution. Semin Laparosc Surg, 10(3):127-139, 2003b.

Gibbon, D., Moore, R. and Winski, R. (Eds.): Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, New York, 1997.

Goodman C. S.: Introduction to Health Care Technology Assessment. Nat. Library of Medicine/NICHSR, 1998, available on-line at http://www.nlm.nih.gov/nichsr/ta101/ta101.pdf

Granström, B. and House, D.: Effective Interaction with Talking Animated Agents in Dialogue Systems. In [van Kuppevelt et al. 2005], 2005.

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.

Grober, E.D., Hamstra, S.J., Wanzel, K.R., Reznick, R.K., Matsumoto, E.D., Sidhu, R.S. and Jarvi, K.A.: The educational impact of bench model fidelity on the acquisition of technical skill: the use of clinically relevant outcome measures. Ann Surg, 240(2):374-381, 2004.

Gustafson, J., Lindberg, N. and Lundeberg, M.: The August Spoken Dialogue System. Proc. of Eurospeech, 1999, 1151-1154.

Haritaoglu, I, Harwood, D, and Davis, L.: Ghost: a human body part labeling system using silhouettes. Proceedings of Fourteenth International Conference of Pattern Recognition, 1998.

Hartson, R., Castillo, J., Kelso, J., Kamler, J., Neale, W.: The Network as an extension of the Usability Laboratory, Proceedings CHI'96, ACM Press, 1996, 228-235.

Hartson, R., Castillo, J.: Remote Evaluation for Post-Deployment Usability Improvement, Proceedings AVI'98, ACM Press, 1998, 22-29.

Heylen, D., van Es, I., Nijholt, A. and van Dijk, B.: Controlling the Gaze of Conversational Agents. In [van Kuppevelt et al. 2005], 2005.

Hilbert, David M., Redmiles, David F., 2000. *Extracting Usability Information from User Interface Events*. - University of California at Irvine - ACM Computing Surveys, Vol. 32, No. 4, 384- 421, December 2000.

Hirschberg, J., Swerts, M. and Litman, D.: Labeling Corrections and Aware Sites in Spoken Dialogue Systems. Proc. of the 2nd SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark, 72-79, 2001.

Hix, D., Gabbard, J. L., Swan , E.II, Livingston, M., A., Höllerer , T.H., Julier, S.J., Baillot , Y., Brown, D.: A Cost-Effective Usability Evaluation Progression for Novel Interactive Systems. Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04), Page: 90276.3, 2004.

Hong, H., Neven, H., and von der Malsburg, C.: Online Facial Expression Recognition based on Personalized Galleries. Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 354-359, 1998.

Hong, J. I., Landay, J. A.: WebQuilt: a framework for capturing and visualizing the web experience. WWW 2001 conference, 717-724, 2001.

ISO (International Standardisation Organisation): ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on usability. http://www.iso.org

Ivory, M. Y. and Hearst, M. A.: Comparing performance and usability evaluation: New methods for automated usability assessment, 1999. Unpublished manuscript. Available at http://www.cs.berkeley.edu/_ivory/research/Web/papers/pe-ue.pdf.

Ivory, M. Y., Hearst, M. A.: The state of the art in automating usability evaluation of user interfaces. ACM Computing Surveys, 33(4), pp. 470-516, December 2001.

Jameson, A. and Klöckner, K.: User Multitasking with Mobile Multimodal Systems. In [Minker et al. 2004a], 2004.

Jannin, P., Fitzpatrick, J.M., Hawkes, D. J., Pennec, X., Shahidi, R., Vannier, M.W.: Validation of Medical Image Processing in Image-Guided Therapy. IEEE Trans. Med. Imaging 21(11): 1445-1449 (2002).

John, B., Kieras, D.: The GOMS family of user interface analysis techniques: comparison and contrast, ACM Transactions on Computer-Human Interaction, 3, 320-351, 1996.

Johnson , L., Edwards, P., Griffin, L. and Hawkes, D.: Depth perception of stereo overlays in image-guided surgery. Medical Imaging 2004: Visualization, Image-Guided procedures, and Display, edited by Robert L. Galloway, Jr., Proceedings of SPIE Vol. 5372, 2004, 263-272, 2004.

Jordan, B. and Henderson, A.: Interaction analysis: Foundations and practice. The Journal of the Learning Sciences, 4, pp. 39-103,1995.

Kalra, P., Mangili, A., Thalmann, N. M. and Thalmann, D.: Simulation of Facial Muscle Actions Based on Rational Free From Deformations. Eurographics 1992, vol. 11(3), pp. 59–69.

Kass, M., Witkin, A. and Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision, 1987, vol. 1(4), pp. 321–331.

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.

Kato, M., So, I. Hishinuma, Y., Nakamura, O. and Minami, T.: Description and Synthesis of Facial Expressions based on Isodensity Maps. In L. Tosiyasu (Ed.), Visual Computing, Springer-Verlag, Tokyo, 1992, pp. 39-56.

Kilchernman O' Malley M.: Comparison of Human Haptic Size Discrimination Performance in Real and Simulated Environments. Proceedings of 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2002. HAPTICS 2002, 10-17

Kirkpatrick, A. E.: Interactive Touch: Haptic Interfaces Based Upon Hand Movement Patterns. CHI'99, Doctoral Consortium, 15-20 May, 1999,59-60

Kneebone, R.: Simulation in surgical training: educational issues and practical implications. Medical Education, 37(3):267-277, 2003.

Kothari, S.N., Kaplan, B.J., DeMaria, E.J., Broderick, T.J. and Merrell, R.C.: Training in laparoscopic suturing skills using a new computer-based virtual reality simulator (MIST-VR) provides results comparable to those with an established pelvic trainer system: J Laparoendosc. Adv. Surg Tech. A, 12(3):167-173, 2002.

Lamata, P., Gomez, E.J., Sanchez-Margallo, F.M., Lamata, F., Gaya, F., Pagador, J.B., Uson, J. and del Pozo, F.: A new methodology to characterize sensory interaction for use in laparoscopic surgery simulation. Medical Simulation, Proceedings, 3078:177-184, 2004.

Laudatrs, D, Campos, M. F. M., and Kumar V.: Design, Implementation and Evaluation of Haptic Interfaces. VI Semana de Pós-Graduação em Ciência da Computação, 4-6 September 2002.

Lecerof, A., Paternò, F.: Automatic Support for Usability Evaluation - IEEE Transactions on Software Engineering, Vol. 24, N° 10, October 1998, 863 - 888.

Lee, Y. C., Terzopoulos, D. and Waters, K.: Realistic face modeling for animation. Siggraph proceedings, 1995, pp. 55-62.

Li, H., Roivainen, P. and Forchheimer, R.: 3-D Motion Estimation in Model Based Facial Image Coding. IEEE Transaction on Pattern Analysis and Machine Intelligence, June 1993, vol. 15, No 6, pp. 545-555.

Lindgaard, G. Testing the Usability of Interactive Computer Systems, In: Lindgaard, G. and Millar, J.: Testing the Usability of Interactive Computer Systems. Proceedings of Workshop at HCI Australia'89. Ergonomics Society of Australia, Computer-Human Interaction Special Interest Group, 1-13, 1989.

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

Liu, A., Tendick, F., Cleary, K. and Kaufmann, C.: A survey of surgical simulation: applications, technology, and education. Presence, 12(6):599-614, 2003.

Magnenat-Thalmann, N., Primeau, N. E. and Thalmann, D.: Abstract muscle actions procedures for human face animation. Visual Computer, 1988, vol. 3(5), pp. 290–297.

Marichal, X., Macq, B., Douxchamps, D., and Umeda, T.: art.live consoritum: Real-time segmentation of video objects for mixed-reality interactive applications. VCIP 2003 - SPIE Visual Communication and Image Processing Intl Conference, Lugano, Switzerland, July 2003, Vol. 5150, 2003, pp. 41-50.

Mase, K.: Recognition of facial expression from optical flow. Institute of electronics information and communication engineers Transactions, vol. E74, pp. 3474--3483, October 1991.

Massaro, D. W.: The Psychology and Technology of Talking Heads: Applications in Language Learning. In [van Kuppevelt et al. 2005], 2005.

Masse, K. and Pentland, A.: Automatic Lip reading by Computer. Trans. Inst. Elec., Info. And Comm. Eng. 1990. Vol. J73-D-II, No.6. pp.796-803.

Michelitsch, G, Williams, J., Osen, M., Jimenez, B. and Rapp, S.: Haptic Chameleon: A New Concept of Shape-Changing User Interface Controls with Force Feedback. Proc. of CHI'04 , Vienna Austia, April 24-29,2004, 1305-1308

Milgram P. and Kishino F.: A Taxonomy of Mixed Reality Visual Displays. IEICE Transactions on Information Systems E77-D (12), pp 1321-1329, 1994.

Minker, W., Bühler, D. and Dybkjær, L. (Eds.): Spoken Multimodal Human-Computer Dialogue in Mobile Environments. Kluwer Academic Publishers, to appear, 2004a.

Minker, W., Haiber, U., Heisterkamp P. and Scheible, S.: The Seneca Spoken Language Dialogue System. Speech Communication, Elsevier, Amsterdam, 2004b.

Moorthy, K. Munz, Y., Sarker, S.K and Darzi, A.: Objective assessment of technical skills in surgery. BMJ, 327(7422):1032-1037, 2003.

Nickel, K and Stiefelhagen, R.: Recognition of 3D-Pointing Gestures for Human-Robot Interaction. Proceedings of Humanoids 2003, Karlsruhe, Germany, 2003

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

Nielsen, J.: Heuristic Evaluation. In Nielsen, J. and Mack, R.L. (Eds.): Usability Inspection Methods. John Wiley & Sons, New York, 1994.

Nielsen, J., and Mack, R.,(eds): Usability Inspection Methods. Wiley, 1994.

Nikolakis, G., Fergadis, G., and Tzovaras, D.: Virtual assembly based on stereo vision and haptic force feedback virtual reality. In Proc. HCI International (June 2003).

Nikolakis, G., Fergadis, G., Tzovaras, D, and Strintzis, M. G.: A mixed reality learning environment for geometry education. In Lecture Notes in Artificial Intelligence (June 2004), Springer Verlag.

Nist Web Metrics: http://zing.ncsl.nist.gov/WebTools/tech.html

Oakley, I., Rose McGee M., Brewster, S. and Gray, P.: Putting the Feel in 'Look and Feel'. CHI Letters volume 2, issue 1, 1-6 April CHI 2000 , 415-422.

Ohya, J., Utsumi, A., and Yamato, J.: Analyzing Video Sequences of Multiple Humans - Tracking, Posture Estimation and Behavior Recognition. The Kluwer International series in video computing. Volume 3. 2002.

Paganelli, L., 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, 2003, 35 (3), 369-378, August 2003.

Pandzic, I. S. and Forchheimer, R.: MPEG-4 Facial Animation. John Wiley & Sons Ltd, West Sussex, England, 2002.

Pantic, M. and Rothkrantz, L.J.M: Automatic analysis of facial expressions: the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 2, pp1424-1445, 2000.

Parke, F. I.: Computer Generated Animation of Faces. Proc. ACM annual conf., 1972.

Parke, F. I.: A Parametric Model for Human Faces. Ph.D. Thesis, University of Utah, Salt Lake City, Utah, 1974, UTEC-CSc-75-047.

Parke, F. I.: Parameterized models for facial animation. IEEE Computer Graphics and Applications, 1982, vol. 2(9) pp. 61-68.

Parke, F. I.: Parameterized models for facial animation revisited. In ACM Siggraph Facial Animation Tutorial Notes, 1989, pp. 53–56.

Parke, F. I. and Waters, K.: Computer Facial Animation. 1996, ISBN 1-56881-014-8.

Paternò, F.: Model-based design and evaluation of interactive applications, Springer Verlag, 1999. ISBN 1-85233-155-0, 1999.

Paternò, F., Ballardin, G.: Model Aided Remote Usability Evaluation. Human Computer Interaction INTERACT'99 (Edinburgh, August 1999), A. Sasse & Ch. Johnson (eds.), IOS Press, 434- 442, 1999.

Paternò, F., Ballardin, G.: RemUsine: a bridge between empirical and model-based evaluation when evaluators and users are distant. Interacting with Computers, Vol.13, N.2, 151-167, 2000.

Penn P., Petrie H., Colwell C., Kornbrot D., Furner S. and Hardwick A.: The Perception of Texture, Object Size and Angularity by Touch in Virtual Environments with two Haptic Devices. Haptic Human-Computer Interaction Workshop, 2000.

Platt, S. M.: A Structural Model of the Human Face. Ph.D. Thesis, University of Pennsylvania, 1985.

Saji, H., Hioki, H., Shinagawa, Y., Yoshida, K. and Junii, T.: Extraction of 3D Shapes from the Moving Human Face using Lighting Switch Photometry. In N. Magnenat-Thalmann, D. Thanlmann (Ed.), Creating and Animating the Virtual World, Springer–Verlag Tokyo 1992, pp. 69–86.

Saulnier, A., Viaud, M. L. and Geldreich, D.: Real-time facial analysis and synthesis chain. In International Workshop on Automatic Face and Gesture Recognition, 1995, pp. 86–91, Zurich, Switzerland, Editor, M. Bichsel.

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.

Scholtz, J., Laskowski, S., Downey L.: Developing usability tools and techniques for designing and testing web sites. Proceedings HFWeb'98 (Basking Ridge, NJ, June 1998), 1998. http://www.research.att.com/conf/hfweb/ proceedings/scholtz/index.html

Scoy, V., Kawai, I., Darrah, S., and Rash, F.: Haptic Display of Mathematical Functions for Teaching Mathematics to Students with Vision Disabilities. Haptic Human-Computer Interaction Workshop, 2000.

Sederberg, T. W., and Parry, S. R.: Free-Form deformation of solid geometry models. Computer Graphics (Siggraph 1996), vol. 20(4), pp. 151 – 160.

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.

Shen, X., Zhou, J., Saddik, A., and Gorganas, D.: Architecture and Evaluation of Thle-Haptic Environments. Proc. of the 8th IEEE International Symposium on Distributed Simulation and Real Time Applications (DS-RT 2004), Budapest, Hungary, October 21-23, 2004

Shneiderman, B.: Designing the user interface: strategies for effective human- computer interaction (Third ed.). Reading, Mass.: Addison-Wesley, 1998.

Siochi, A., Ehrich, R.: Computer Analysis of User Interfaces Based on Repetition in Transcripts of User Sessions, ACM Transactions on Information Systems, 9, 4, October, pp. 309-335, ACM Press, 1991.

Sjoerdsma, W., J.L. Herder, M.J. Horward and A. Jansen (1997). Force transmission of laparoscopic grasping instruments. Minimally Invasive Terapy and Allied Technologies 6, pp 274-278.

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

Sjostrom, C.: Touch Access for People With Disabilities. Licentiate Thesis, in CERTEC Lund University, Sweden, 1999.

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

Smith, C.D., Farrell, T.M., McNatt, S.S. and Metreveli, R.E.: Assessing laparoscopic manipulative skills. The American. Journal of Surgery, 181(6):547-550, 2001.

Stassen, H.G., Dankelman, J., Grimbergen, K.A. and Meijer, D.W.: Man-machine aspects of minimally invasive surgery. Annual Reviews in Control 25 (2001) 111-122.

Sturm, J., Bakx, I., Cranen, B. and Terken, J.: Comparing the Usability of a User Driven and a Mixed Initiative Multimodal Dialogue System for Train Timetable Information. Proc. of Eurospeech, 2003, 2245-2248.

Sturm, J., Cranen, B., Wang, F., Terken, J. and Bakx, I: Effects of Prolonged Use on the Usability of a Multimodal Form-filling Interface. In [Minker et al. 2004a], 2004.

Terzopoulos, D. and Waters, K.: Physically-based facial modeling, analysis, and animation. J. Of Visualization and Computer Animation, March, 1990, vol. 1(4), pp. 73-80.

Trevisan, D., Gemo, M. Vanderdonckt, J. Macq, B.: Focus-Based Design of Mixed Reality Systems. To appear: 3rd International Workshop on Task MOdels and DIAgrams for user interface design, Prague, Czeck Republic, November 15-16, 2004

Trevisan, D., Vanderdonckt, J., Macq, B., Raftopoulos, C.: Modelling Interaction for Image-Guided Procedures. Proc. of Int. Conf. on Medical Imaging SPIE'2003.

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.

Urtasun, R. and Fua, P.: 3d human body tracking using deterministic temporal motion models. Technical Report IC/2004/03, EPFL, January 2004

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

Vallino J.: Augmented Reality Page. Available at http://www.se.rit.edu/~jrv/research/ar /index.html (Department of Software Engineering Rochester Institute of Technology) last access at 22/11/2004.

van Kuppevelt, J., Dybkjær, L. and Bernsen, N.O. (Eds.): Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Kluwer Academic Publishers, to appear, 2005.

Vogt, S., Wacker, F., Khamene, A., Elgort, D. R., Sielhorst, T., Niemann, H., Lewin, J. S. and Sauer, F.: Augmented Reality System for MR-guided Interventions: Phantom Studies and First Animal Test. Medical Imaging 2004: Visualization, Image-Guided procedures, and Display, edited by Robert L. Galloway, Jr., Proceedings of SPIE Vol. 5367, 2004, 100-109.

Walker, M., Litman, D., Kamm, C. and Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. Proc. of the Association of Computational Linguistics (ACL), 1997, 271-280.

Wang, Y., 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.

Waters, K.: A muscle model for animating three-dimensional facial expression. In Maureen C. Stone, editor, Computer Graphics (Siggraph proceedings, 1987) vol. 21 pp. 17-24.

Waters, K. and Frisbie, J.: A Coordinated Muscle Model for Speech Animation. Graphics Interface, 1995 pp. 163 – 170.

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.