COMBINING MULTIPLE INPUT MODALITIES FOR VIRTUAL REALITY NAVIGATION – A USER STUDY

Frank Althoff, Gregor McGlaun, Gunter Spahn, Manfred K. Lang

Institute for Human-Machine-Communication,
Technical University of Munich, Arcisstr. 21, 80290 Munich, Germany

{althoff, mcglaun, spahn, lang}@ei.tum.de

ABSTRACT
This usability study aims at evaluating in which way different groups of users deal with a multimodal interface for navigating in arbitrary virtual worlds. Besides classical input devices like keyboard, mouse and touchscreen the test subjects can control the system by natural speech utterances as well as dynamic hand and head gestures. Experts tend to use haptic devices, whereas normal computer users and beginners prefer combinations of advanced input devices. As an overall result, the multimodal interface was rated very intuitive since users are not forced to use predefined interaction styles, but can freely choose among multiple input devices, instead.

1. INTRODUCTION

Parallel to the rapid development of computer systems the design of the appropriate user interfaces (UIs) has undergone significant changes, too, leading to various generations of man-machine interfaces. Multimodal virtual reality (VR) interfaces currently resemble the highest step in this development. Providing multidimensional input possibilities and employing innovative 3D display strategies these types of interfaces facilitate flexible and intuitive access to the complex functionality of today's computer systems. Moreover, multimodal systems offer an increased level of error robustness since they integrate redundant information shared between the individual input modalities.
With regard to the analysis of human factors, a fundamental task in designing VR systems consists in solving the problem of orientation in three-dimensional space. Our work contributes to multimodal VR research by analyzing the way users deal with a multimodal interface for navigating in arbitrary virtual VRML worlds combining classical input devices like keyboard and mouse with touchscreen (TS) interaction, natural speech utterances and dynamic hand and head gestures. An impression of the working environment and the test setup can be taken from figure 1.

2. EXPERIMENTAL SETUP

The primary goal of our study is to evaluate which modalities, and modality combinations, respectively, are preferred with regard to a given navigation task. In this context, we want to determine to which extent the overall user intention is distributed among complementary, redundant and competing information streams of the individual input modalities. Furthermore, the study is intended to gather multimodal data material, serving as a basis for evaluating various multimodal integration concepts, and developing a multimodal input signal simulator.

2.1. Target applications

Our navigation interface is mainly based on the VRML browser FreeWRL [Ste01]. Navigating is equal to moving a virtual camera through the scene. The probands can exhaust the full spectrum of translational and rotational movements. In the context of a multimodal navigation interface, we extended the original FreeWRL functionality, introducing discrete movements, a step-size mechanism, regulating the amount of increments, as well as a *repeat* and an n-stage *undo* function. The user gets both acoustical and visual feedback, informing him of the task at hand, the current system status, the last recognized command, and potential error states.
For interfacing the browser to additional input devices we are using a technique introduced in [Alt01]. A TCP/IP socket backport enables the browser to react on commands sent over the net. Given in the form of an adapted context-free grammar (CFG), these commands extensively model the browser functionality, and thus provide the representation of domain- and device-independent multimodal information contents. As the individual input devices all share the same formalism, it makes no difference to the browser module by exactly which input device a specific event has been generated. The browser module just operates on the formal model of the CFG, using the socket port as the primary information source and disabling the built-in navigation of the browser.

From: Proc. HCI 2001, Lawrence Erlbaum Ass.

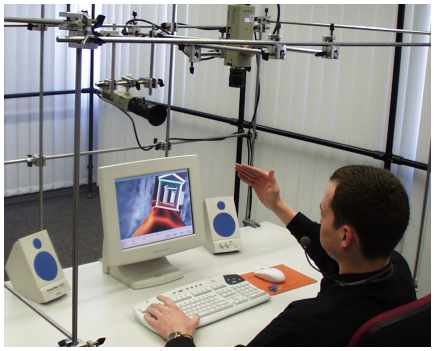

Fig. 1: Working environment

	M	K	H	A	S
M	x	**X**	-	x	**X**
K		x	-	x	x
H			x	x	**X**
A				x	**X**
S					x

Table 1: Modality combinations

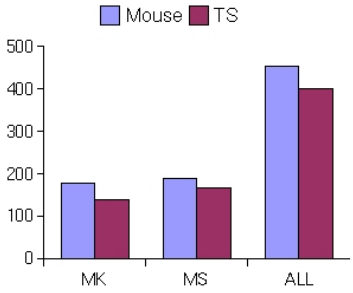

Fig. 2: Transaction times [sec]

2.2 Test methodology

The functionality of the navigation interface is partly realized according to the 'wizard-of-oz' test paradigm. In contrast to haptic interactions that are directly transcribed by the system, the semantic higher-level modalities (hand, head and speech) are simulated by a human person supervising the probands via audio- and video-signals. The so-called 'wizard' interprets the user's intention and generates the appropriate browser commands. To collect a multitude of possible interaction styles the wizard is instructed to be extremely cooperative, accepting both static and dynamic gestures as well as command and natural speech utterances. In the case of ambiguous user actions, the interaction is to be interpreted at best in the current system context.

The user study is carried out at the usability laboratory of our institute, consisting of two rooms. Probands are located in the test room. Separated from this area by a semipermeable mirror, the control room serves for recording and analyzing user interactions. To carry out identical reproducible test runs, we have developed a special software suit called UsaWiz supporting the wizard in managing the various system parameters, semi-automatically announcing the navigation tasks at specified points of time and logging all kind of transactions for a detailed offline analysis. The UsaWiz has already proved its concept in various usability experiments applied in different domains.

2.3 Test plan

The test is mainly designed to evaluate combinations of the following input devices mouse / touschsceen (M), keyboard (K), hand (H) and head (E) gestures, and speech (S). An overview of potential modality combinations is given in table 1. For a detailed validation of the haptical devices two test series are compiled, one with using mouse, one with using touchscreen interaction, instead. Before starting the test, the probands are learning the functionality of the navigation interface in an interactive training period together with the wizard, mainly by employing haptic interaction. At the same time the use of the other modalities and potential modality combinations are explained.

A test run consists of two blocks. In the first block test subjects have to use four prescribed modality combinations (MK, MS, HS, ES), characterized by a capital "X" in table 1, each to solve identical navigation tasks. The second block exposes a much more complex navigation task, but the test subjects are now allowed to freely combine all of the available input devices. After each part the test subject has to evaluate the system by filling out questionnaires.

3. RESULTS

A total of 40 persons participated in the usability tests, 17 in the first test series using mouse interaction and 23 in the second series using touchscreen instead. A control set of three persons participated in both test series. The test subjects were divided into three groups: 11 beginners (*beg*), 20 normal computer users (*norm*) and 9 experts (*exp*) due to the answers given in the first question form. The average age of the probands was 28.8 years. Besides many engineering students, people of different education and profession took part in the tests.

3.1 Preferred modalities

With reference to touchscreen interaction the processing time of persons in the first series using the mouse was significantly higher. As shown in figure 2, this is true for both the pre-given modality combinations and the free combination in the second test block. The biggest difference has been observed when purely combining haptic devices (MK) with touchscreen interaction being 26.43% faster.

The distribution of unimodal system commands is depicted in figure 3. Obviously all users favor keyboard input for single interactions (37,8% - 44,5%). For next best choice experts and normal users clearly prefer mouse / TS (35.1% and 31.7%), whereas beginners tend to employ speech (20.9%), closely followed by hand gestures (15.7%).

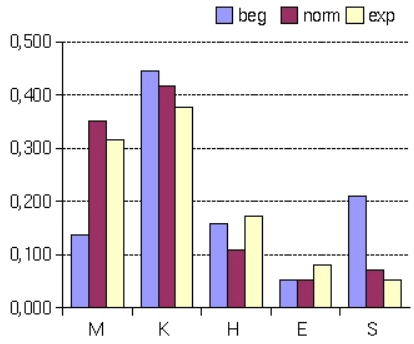

Fig. 3: Distribution of unimodal user interactions

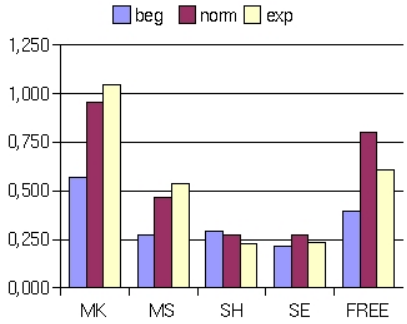

Fig. 4: Efectiveness of various modality combinations

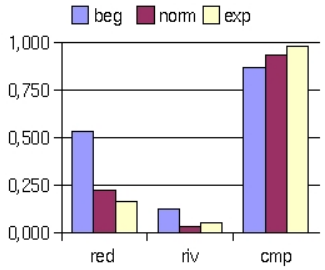

Fig. 5: Distribution of multi-modal interactions

Figure 4 shows the effectiveness of combined user interactions in the test scenarios, measured as the number of performed transactions per time (tpt). Experts and normal users work most efficient when combining mouse with keyboard (1.04 / 0.95 tpt) or speech (0.54 / 0.47 tpt). As an outstanding result we obtained that in part three of the first block (speech, hand) beginners got even higher scores than the other two groups (0.30 tpt to 0.27 and 0.23 tpt, resp.). Purely haptic interaction (MK) is still most time efficient, but concerning the other combination scenarios all groups of test subjects worked noticeably more effective in the second block (free combination), with normal users performing nearly 32% better than experts (0.80 to 0.61 tpt).
Concerning the remarks of the closing questionnaire users asked for advanced navigation features, i.e. they wanted the system to continuously react on head and hand movements. Moreover probands demanded to phrase browser commands applying context knowledge of the current navigation situation, e.g. by simply saying, "*go to that door*".

3.2 Modality combinations

As the navigation tasks were quite simple, unimodal interaction definitely overruled multimodal interaction. Yet, detailed analysis clearly proved that with growing complexity the use of multimodal interaction increases. Although only applied in about one fifth of all interactions, combined multimodal commands symbolize the core interaction style as they were particularly used to change navigation contexts, i.e. from translational to rotational movements.
The distribution of redundant (red), rival (riv) and complementary (cmp) interactions is shown in figure 5. For all groups of users complementary actions occurred most often (86.9% - 98.3%). In 12.5% of all complementary actions more than two modalities were applied. Real multimodal commands (showing no intramodal dependencies) appeared in 67.3% of combined interactions. Especially beginners seem to indicate redundant interactions (53,2%), more than twice as much in comparison normal users (22.4%). These results also emphasize the observation that beginners show complementary behavior coupled with redundancy to a high degree.
As an overall result, analysing the questionnaires well supported the measured values discussed above. Experts and normal users rated mouse / touchscreen in combination with speech best, whereas beginners stated to prefer speech in combination with hand gestures. Interestingly, head gestures were evaluated very bad, which contradicts the measured values, since combined with speech head movements made up at least 20% of all combined interactions.

4. CONCLUSIONS AND FURTHER WORK

In general users highly accepted having the possibility to freely choose among multiple input devices and not to be forced to use predefined interaction styles. The outcome of this user study clearly motivates further research in multimodal interaction systems as they provide the user with greater naturalness, expressive power and flexibility. While a natural speech understanding module has already been realized [Sch01], current research endeavors are to integrate dynamic hand and head gestures modules meeting real-time requirements. Moreover, we are working on the design of integration technologies based on the experience and multimodal data collected in this study.

REFERENCES

[Ste01] Stewart, J. (2001): FreeWRL homepage, http://www.crc.ca/FreeWRL, April 2001

[Alt01] Althoff, F., Volk, T., McGlaun, G., Lang, M. (2001): A Generic User Interface Framework. Poster Proceedings HCI 2001, New Orleans, USA (this conference)

[Sch01] Schuller, B., Althoff, F., McGlaun, G., Lang, M. (2001): Navigation in Virtual Worlds via Natural Speech. Poster Proceedings HCI 2001, New Orleans, USA (this conference)