GMD Report 138

GMD –
Forschungszentrum
Informationstechnik
GmbH

Wolfgang Broll, Léonie Schäfer (Eds.)

The Future of VR and AR Interfaces

Multi-Modal, Humanoid, Adaptive and Intelligent

Proceedings of the Workshop
at IEEE Virtual Reality 2001
Yokohama, Japan
March 14, 2001

July 2001

GMD –
Forschungszentrum Informationstechnik GmbH
Schloß Birlinghoven
D-53754 Sankt Augustin
Germany
Telefon +49 -2241 -14 -0
Telefax +49 -2241 -14 -2618
http://www.gmd.de

Anschrift der Herausgeber/Address of the editors:
Dr. Wolfgang Broll
Léonie Schäfer
Institut für Angewandte Informationstechnik
GMD – Forschungszentrum Informationstechnik GmbH
D-53754 Sankt Augustin
E-Mail: wolfgang.broll@gmd.de
leonie.schäfer@gmd.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme:
The future of VR and AR interfaces : multi-modal, humanoid, adaptive and intelligent ; proceedings of the workshop at IEEE Virtual Reality 2001, Yokohama, Japan, March 14, 2001 / GMD-Forschungszentrum Informationstechnik GmbH. Wolfgang Broll ; Léonie Schäfer (eds.). - Sankt Augustin : GMD-Forschungszentrum Informationstechnik, 2001
(GMD-Report ; 138)
ISBN 3-88457-975-4

ISSN 1435-2702
ISBN 3-88457-975-4

Abstract

This document contains the contributions to the Workshop on The Future of VR and AR Interfaces held on March 14th at IEEE Virtual Reality 2001 in Yokohama, Japan. The original contributions were submitted as position papers. The papers collected here are revised or extended versions of the submitted papers. There is a total of 16 papers covering a wide range of issues related to VR and AR: from 3D Interaction and Haptics to Augmented Reality, Mobility and Tracking, Mixed Reality and Natural Interaction as well as Conversational User Interfaces and Human Aspects.

Keywords

3D Interaction, Mixed Reality, Augmented Reality, Conversational User Interfaces

Kurzfassung

Dieser Band beinhaltet alle Beiträge des Workshops zu The Future of VR and AR Interfaces. Der Workshop fand statt am 14. März 2001 als Teil der IEEE Virtual Reality 2001 in Yokohama, Japan. Die Beiträge wurden eingereicht als Positionspapiere, welche in diesem Band in überarbeiteter und ausführlicher Form zusammengefaßt sind. Die insgesamt 16 Beiträge behandeln ein weites Spektrum in VR und AR: von 3D Interaction und Haptics zu Augmented Reality, Mobility und Tracking, Mixed Reality und Natural Interaction bis hin zu Conversational User Interfaces und Human Aspects.

Schlagworte

3D Interaction, Mixed Reality, Augmented Reality, Conversational User Interfaces

Preface

Virtual worlds have become more and more visually elaborated and emotive. They aim at an environment nearly indistinguishable from the real world. However, in spite of all technological and artistic advances, simulated worlds are still far from perfect in their realism. It is at the interface between the human and the computer environment, where this lack of realism becomes most apparent. Transferring natural interaction and communication principles from the real world to cyberspace in a seamless fashion is a very challenging task. Can we improve interface technology to the point where people will communicate with a synthetic environment in a natural way, in a style similar to their day-to-day interaction with the real world? Or should we turn away from or go beyond realism to reach the full potential of virtual worlds?

With the goal of deep immersion and perfect integration of real and virtual environments in mind, we are reviewing interface technologies that may be particularly apt to overcome some of the limitations we are still facing today: Multi-modality addresses all human senses and enables a wide variety of human articulation to be part of the interface. Additionally, future interfaces will likely display adaptive and intelligent behavior. Humanoid persona can add an interpersonal touch to the immersive experience.

This collection is a result of the Workshop on the Future of VR and AR Interfaces. The workshop took place on March 14th at IEEE Virtual Reality 2001 in Yokohama, Japan. The purpose of the workshop was to bring together researchers from the area of AR/VR technology, human-computer interaction, AI as well as psychologists, SF authors and other people with a vision of what the interface between humans and computer generated environments should look, sound, feel, and be like. The goal of this workshop was to showcase, develop, and discuss concepts for better AR/VR interfaces and to evolve ideas towards the realization of interfaces enabling deep immersive, elaborated virtual environments.

We would like to take this opportunity to thank all the people who contributed to this workshop. In particular we thank our workshop co-organizers Tobias Höllerer and Doug Bowman for their outstanding commitment, the workshop participants for the fruitful and stimulating discussion, the authors for their submissions, and the IEEE VR organizers for their support.

Wolfgang Broll *Léonie Schäfer*

Contents

3D Interaction and Haptics

Chair: Doug Bowman, Virginia Tech

A Prototype System for Synergistic Data Display

J. Dean Brederson, Milan Ikits, Christopher R. Johnson, Charles D. Hansen

Scientific Computing and Imaging Institute
School of Computing, University of Utah
50 Central Campus Drive Rm 3490
Salt Lake City, UT 84112-9205, USA
{ jdb, ikits, crj, hansen } @cs.utah.edu

Abstract

Multi-modal interfaces have been shown to increase user performance for a variety of tasks. We have been investigating the synergistic benefits of haptic scientific visualization using an integrated, semi-immersive virtual environment. The Visual Haptic Workbench provides multi-modal interaction; immersion is enhanced by head and hand tracking, haptic feedback, and additional audio cues. We present the motivation, design and implementation of the prototype system and describe some challenges ahead in the context of questions to be answered. Preliminary results indicate that visualization combined with haptic rendering intuitively conveys the salient characteristics of scientific data.

Introduction

A primary advantage of haptic interfaces is that they provide bi-directional interaction via position sensing and force feedback, thereby utilizing additional sensory channel band-width of the user. By combining haptic rendering and semi-immersive visualization, we hope to increase intuitive understanding of scientific data. For this purpose, we have designed and implemented a prototype testbed system, the Visual Haptic Workbench (see Figure 1). Using this sys-tem, we are investigating the synergistic benefits of combined visual and haptic scientific data rendering.

We desire an integrated environment capable of *bounded error interaction*, where a unified error tolerance describes the total system error throughout the workspace. Such a goal requires careful consideration of hardware components for performance, integration, and extensibility, a modular and efficient software infrastructure, and robust calibration and coregistration techniques. As a preliminary evaluation of the system, we experimented with synergistic rendering methods for a variety of scientific visualization applications.

Motivation for Design

Research on virtual workbench environments and haptics has produced many interesting results. Several applications of haptics to scientific visualization are relevant to the development of our system, including projects at UNC Chapel Hill, CSIRO, the University of Tsukuba, and the University of Boulder. In addition to these integrated systems, there are several relevant research publications on combined haptic and visual rendering techniques. The Visual Haptic Workbench [1] is a testbed system for investigating the possibilities of synergistic display of scientific data.

Building a multi-modal system for synergistic display of scientific data involves three broad implementation is-sues. *Calibration* increases workspace accuracy to provide faithful data rendering while avoiding conflicting perceptual cues. *Coregistration* methods fuse multiple calibrated workspaces to accommodate their relative location, orientation, and scale. *Compensation* for communication and computational delays maintains interactivity and maximizes user performance. Achieving bounded error interaction requires careful consideration of solutions to these issues.

We also considered specific research applications to pursue with this system. At the SCI Institute, a variety of datasets are routinely investigated. These datasets vary in modality, size, grid type, and may be static or dynamic. Considering these demands, our system infrastructure must be efficient, modular, extensible, and scale well with data size.

State of the Prototype

We have constructed a prototype system consisting of a SensAble PHANToM 3.0 mounted in a T configuration above a Fakespace Immersive Workbench (see Figure 1). The PHANToM is suspended above the workbench with a cross-braced lumber frame. A redundant safety mechanism protects the user during operation. The dominant hand of.the user experiences haptic feedback from the PHANToM, and the subdominant hand navigates through a menu inter-face via Pinch glove contact gestures. A Polhemus Fastrak is used for head and hand tracking, and an audio subsystem provides reinforcing sound cues. Finally, the Immersive Workbench provides a semi-immersive, virtual view for the user based on the tracked head location.

To support our initial investigation, we designed and implemented a software framework for application development on the Visual Haptic Workbench. Our software libraries contain haptic rendering methods, general VR support for immersive environment rendering and interaction, visualization methods, interface widgets, dataset classes, menu functions, and geometry tessellators. These software libraries are realized as runtime daemons and application threads, which communicate via shared memory data models and UDP messages. To maintain interactive update rates, these processes run concurrently on an SGI Onyx2 with 250 MHz R10000 processors and InfiniteReality2 graphics.

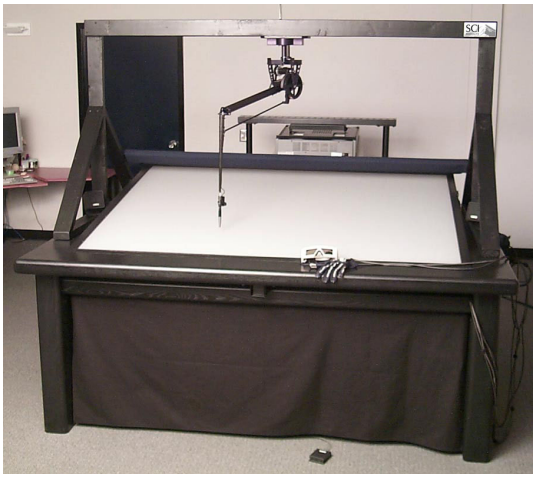


Figure 1: The integrated system prototype.

Challenges to Meet

There are several areas of improvement for our system. We briefly consider the following key issues:

- **Scalability and Interactivity:** Our initial applications have been able to maintain interactive visual framerates and haptic updates of 1kHz. With increasing dataset sizes, we hope to leverage parallel computation and distributed data models to ease the burden of data glut.

- **Quality of Haptics:** Our current system has 6DOF sensing and 3DOF force feedback, which can be a problem for applications where 6DOF is required for natural interaction. The precision of haptic control, bandwidth required, and kinematic calibration are issues to solve to improve the haptic performance of our system.

- **Display Accuracy:** Our current display consists of an analog CRT projector, rear surface mirror, and nonlinear diffusion surface. These characteristics limit the possible calibration and visual performance. Upgrades in progress include a front surface mirror and a new dif-fusion material with superior viewing properties.

- **Tracking Accuracy:** We have developed methods for quantifying and correcting magnetic tracker distortion and incorporated them into our prototype [2]. In addition, we are actively evaluating new tracking technologies as candidates for a system upgrade.

- **Overall System Coregistration:** Although our current coregistration methods are somewhat ad-hoc, we are developing methods to calibrate and coregister the system workspaces more precisely.

Questions to Answer

The Visual Haptic Workbench project was motivated by our desire to answer the following compelling questions:

- What constitutes a truly synergistic interface and how do we quantify user performance for the system?

- What specific factors contribute to the synergistic combination of haptics and visualization?

- What can we achieve with different CHI paradigms?

- How accurately can we engineer our system towards our goal of bounded error interaction?

We believe that our prototype system will provide the basic infrastructure for pursuing this future work. Preliminary results based on informal user evaluation indicate that the Visual Haptic Workbench is an effective tool of discovery for the exploration of scientific datasets.

Acknowledgements

Support for this research was provided by ARO DURIP grant DAAG-559710065, the DOE Advanced Visualization Technology Center (AVTC), and NSF Grant ACI-9978063.

References

[1] J. Brederson, M. Ikits, C. Johnson, and C. Hansen. The visual haptic workbench. In *Proc. Fifth PHANToM Users Group Workshop (PUG)*, 2000.

[2] M. Ikits, J. Brederson, C. Hansen, and J. Hollerbach. An improved calibration framework for electromagnetic tracking de-vices. In *Proc. IEEE Virtual Reality*, 2001 (to appear).

Large Volume Interactions Wearing Haptic Suits

Grigore C. Burdea

Associate Professor of Computer Engineering
Rutgers University,
96 Frelinghuysen Rd., Room 721
Piscataway, NJ 08854, USA.
(732) 445-5309; fax 445-4775
burdea@vr.rutgers.edu

Need

Virtual reality has been expanding in recent years to ever-larger simulation volumes. Natural multi-modal user interaction within these volumes requires ergonomic, light and wireless interfaces, that track user's full body posture and position in the room, process low-level data locally and transmit high-level data to host/hub computers on a wireless line. Such data should include voice (recognition) commands and force feedback to the user.

Technology

The visual feedback component is assured by CAVEs, Domes, Walls, and similar large and surround stereo displays. New flat screen large-surface displays, or large auto-streoscopic displays will reduce the cost and complexity of current systems;

Tracking over large volumes was a challenge that recently was overcome by the wireless tracking suits to be marketed by InterSense Co. Unlike the Motion Star and Star Track magnetic trackers, the InterSense Constellation suit uses inertia/ultrasonic technology immune to metallic interference, and offering a much larger tracking surface.

Wearable computers now exist (such as those made by Xybernaut Co.), albeit with less computation power than needed. It is anticipated that "system on a chip" technology will solve this computational deficit in a few years, allowing robust processing of data, including voice recognition and 3-D graphics on the suit. This in turn will allow face-mounted displays (such as the Olympus Eye Trek) to replace CAVE-type displays.

Challenges

The missing modality at present is haptics. Touch feedback actuators cannot provide sufficient realism. Current force feedback actuators have low power/weight ratio and are energy intensive. Therefore exoskeleton-type suits become too heavy, and cumbersome to be worn without some form of grounding. This in turn negates the concept of full-room interaction and significantly limits simulation natural interaction.

Solution

Several novel actuator concepts have emerged in recent years. While sufficiently compact and light, they have not been integrated in a simulation suit. Competing designs will be analyzed as candidates for integration in suits. A system concept will then be given of such future haptic suits.

VR User Interface: Closed World Interaction

Ching-Rong Lin

Landmark Graphics Corporation,
15150 Memorial Address, Houston, TX 77079, USA
jcrlin@lgc.com

Abstract

We designed and implemented a user interface technique that uses a bounding box as a metaphor to facilitate interaction in a Virtual Reality (VR) environment [1]. Because this technique is based on the observation that some of the VR application fields are contained in a closed world, we call it Closed World Interaction (CWI). After the user defines a closed world, the necessary virtual buttons are shown around the closed world which is presented by a frame. These virtual buttons are then used to interact with models. We also integrate some of the 2D Windows, Icons, Mouse and Pointer (WIMP) metaphors into CWI technique, reflecting our belief that users will be able to adapt to this environment quickly. A series of user studies were conducted to investigate the effectiveness of this technique. The results indicate that users can define a closed world quickly. Experience appears to be an important factor, and users can be trained to become familiar with CWI in the VR environment. The constrained interactions can also enhance the accuracy of selection. Two-handed manipulation somewhat improves the speed. We also applied this technique to a VR application for the geosciences and then invited geoscientists and software developers to evaluate this application. [2]. However, our results also suggest that the major challenge to the successful implementation of VR involves the improvement of VR interaction techniques. In this paper, we will discuss these results.

Introduction

In VR, the interaction devices differ from those used with conventional computers. Six-degrees-of-freedom sensors are provided to interact with objects, enabling the user can manipulate objects using very natural means such as grabbing, holding, or snapping.

Hinckley *et al*. [3] showed that users could finished assigned tasks up to 36% faster without a loss of accuracy using multidimensional input techniques rather than mouse-driven devices. Hinckley *et al*. provided a very comprehensive survey of design issues for developing effective free-space three-dimensional user interfaces [4].

However, why have these kinds of interaction devices seldom gone beyond the research laboratory ? Of course, cost is one important reason. Other problems include a lack of haptic feedback and limited input information. Another factor is that poor interaction may confuse users by providing "too many" degrees of freedom. Lastly, another important reason is that, although a dominant paradigm of interaction using windows and widgets has existed on conventional desktop computers for some time, there is not yet a unified interface in the VR environment.

To enable the users to adapt to the VR environment quickly, many research studies and VR applications incorporated the 2D conventional Windows, Icons, Mouse, and Pointer (WIMP) metaphors into the interaction techniques used in the virtual environment.

However, no systematic evaluation has been performed to quantify the impact of using this kind of approach. Therefore, we designed a user interface technique called Closed World Interaction (CWI) that uses a bounding box as a window metaphor to facilitate interaction in a VR environment. A user study was then conducted to evaluate the usefulness of this technique within the VR environment. The results indicate that users can become familiar with CWI quickly.

Experience appears to be an important factor, and users can be trained to become familiar with CWI in the VR environment. Two-handed manipulation somewhat improves the speed, especially for the novices. The constrained movement can also help fin-grain selection without special training.

The basic concept for this CWI approach arose from the observation that some VR application fields make use of a fixed range of a simulation environment or data area. For examples, the earth can be a closed world, a horizon can be considered as a closed world and a well can be also a closed world. For these kind of closed environments, a bounding box usually exists and serves as the coordinate frame or the

provider of the information metaphor. We believe that this existing frame is a good metaphor upon which to base the interaction. In this paper, we describe the use of the bounding box as a 3-D widget. When a user wants to interact with data, the necessary virtual buttons are shown around the frame. If no interaction is needed, all the virtual buttons on the frame can be made invisible so that this frame simply acts as the reference outside the bounding box. We also use some of the WIMP metaphors to enhance the CWI technique.

User Studies for CWI [1]

A user study was conducted to evaluate the usefulness of the CWI technique within the VR environment. In this study, the virtual world was attached to the non-dominant hand. Two methods which were used to define a closed world within the VR environment were tested:

Method 1: Subjects select one of the corners of the closed world and then drag to the opposite corner to create a rectangular wire frame to specify the closed world. We use "Drag" to name this method for this study.

Method 2: Subjects set one coordinate at one time. We call this method "Slider" here.

The Result and Discussion

1. The slider or drag operations may be considered as trivial interactions in the 2D conventional interface. In this study, the subjects spent nontrivial amounts of time to do the same operations in the VR environment. The experienced subjects were much better than the novices were in the time category for both methods. This result indicates that people can be trained to become familiar with VR interface.
2. The second method, Slider, showed improved accuracy for all subjects. This shows that constrained movement like "Slider" can help accuracy.
3. The experienced subjects and the novices had only about 10 seconds difference (18.429s vs. 27.536s) between them for the "Slider" selection method. On the other hand, there was about 35 seconds difference (23.66s vs. 57.383s) between the two groups for the "Drag" method. Both groups had a similar accuracy for setting the closed world using the "Slider" technique. This shows that constrained movement like "Slider" can help fine-grain selection without special training.

User Studies for Two-handed Manipulation [1]

A user study was also conducted to evaluate whether or not two-handed manipulation enhances the CWI technique. Two primary experimental configurations were tested:

Configuration 1 (one-handed): The menu and the virtual world were fixed at a location specified by the user.

Configuration 2 (two-handed): Both the menu and the virtual world were attached to the non-dominant hand of the user.

The Result and Discussion

1. Two-handed manipulation speeds up the CWI technique and also somewhat improves accuracy.
2. There was a greater time difference between the two configurations for the novice group (32.051 vs. 27.536) than for the experienced group (18.759 vs. 18.429). From this result, the novices appeared to favor two-handed manipulation. In fact, most subjects of the second groups expressed that they felt more comfortable using two-handed manipulation than the fixed location for the menu system.

User Studies for CWI in the Geoscience Application [2]

A group of oil and gas service companies formed the VR in Geosciences (VRGeo) Research Consortium in late 1997 in order to systematically study the use of virtual reality techniques to model hydrocarbon reservoirs. The Consortium is carrying out a portion of its research program in cooperation with the Virtual Environments Research Institute (VERI) at the University of Houston. To evaluate this VR application for geoscience visualization, a number of professional geoscientists and software developers were recruited from the VRGeo Consortium. The evaluations were carried out at the VERI. The evaluators for this user study included managers, geoscience interpreters, and software developers from companies participating in the VRGeo Consortium. A total of seventeen evaluators took part in the user study.

Our results suggest that, though most evaluators felt comfortable interacting with the data, the interaction techniques we implemented in our VR application are still not as effective as the three-dimensional visualization techniques we used. Through the use of familiar interaction techniques such as clipping planes, VR provides

a natural means for manipulating three dimensional data sets.

Conversely, two-dimensional nature of the menu system and virtual icons, coupled to a three-dimensional display and interaction environment. At present, a menu system of some type is an essential component for facilitating interaction with data. Some evaluators suggested that voice recognition might be a workable substitute for the menu system.

References

[1] C. Lin and R. Loftin, VR User Interface: Closed World Interaction, VRST'2000, pp. 153-159.

[2] C. Lin and R. Loftin, Interaction with Geoscience Data in an Immersive Environment, VR'2000, pp. 55-62.

[3] K. Hinckley, J. Tullio, R. Pausch, D. Proffitt, N. Kassel, Usability Analysis of 3D Rotation Techniques, Proceedings of UIST'97, 1997, pp. 1 - 10.

[4] K. Hinckley, R. Pausch, J. Goble, N. Kassel, A Survey of Design Issues in Spatial Input, Proceedings of UIST'94, 1994, pp. 213 – 222

Cooperative and Simultaneous Object Manipulation in Collaborative Virtual Environments

Márcio Serolli Pinho[1,2] ,Carla Maria Dal-Sasso Freitas[1]

[1] UFRGS - Universidade Federal do Rio Grande do Sul
Instituto de Informática, Pós-Graduação em Ciência da Computação
Caixa Postal 15064, 91501-970 Porto Alegre, RS, Brazil
carla@inf.ufrgs.br

[2] PUCRS – Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática , Av. Ipiranga, 6681- Predio 30, Bloco IV - 90.619-900 Porto Alegre, RS, Brazil
pinho@inf.pucrs.br

Introduction

The use of collaborative virtual environments (CVE) has become more and more popular due to the new facilities provided by personal computer systems and network resources, each day cheaper, faster and more reliable.

Most efforts on CVEs are devoted to build support tools and minimize network traffic. Although some works address interaction in these environments, in most of them the actual simultaneous object manipulation is not possible. Usually, when one user captures an object for manipulation, the other cannot participate in the same procedure because some kind of mutual exclusion denies the cooperation.

There are some other works on collaborative interaction but in non-immersive environments. Viullème and Thalmann [10] or example, described a system based on VLNET framework in which the user can select a gesture and a facial expression from a set of options presented on a screen. After the selection the choices are incorporated in an avatar that represents the user inside the virtual environment. Another example of interaction tools for non-immersive environments is Spin [4]. The aim of this system is to create a kind of "conference table". This is built on a computer screen as a set of panels placed side by side on a circular distribution like a round table. The panels can be rotated as if they were around the user's head. To each panel a user can associate another user or an application. To select one application to be executed or other user to talk with, one has simply to rotate the panels until the desired choice is in the middle of the screen.

There are also interfaces that use augmented reality. In these interfaces, the users are located on the same space (usually in the same room) and are able to see each other wearing see-through glasses. These glasses allow presenting virtual objects superimposed to real world objects. This setting can provide the same type of collaborative information that people have in face-to-face interaction such as communication by object manipulation and gesture [1]. Such systems have been used in games like AR^2 Hockey [8], in scientific visualization systems (Studierstube [9]), in discussion support systems (Virtual Round Table [3]) and in object modelers like SeamlessDesign [5].

Simultaneous Interaction

The analysis of existing works shows that there is an important lack to be fulfilled: how to allow more than one user to interact over the same object at the same time in a fully immersive environment.

In most of the existing works, processing of simultaneous manipulation commands is done aiming to destroy this simultaneity. Usually, to do so, they adopt one of the following methods:

- Priority assignment to one of the users;
- Ordering the users commands according to the time they are generated;
- Some lock mechanism, which ensures exclusive access to the object to one user at a time.

- Using these methods, at each moment the object will receive only one action selected among all users actions. In figure 1, for example, as User A has a higher priority than User B, only the action from A will be sent to the object.

In a real simultaneous interaction situation, instead of **choosing** among different actions, the system should try to find ways to **combine** these actions and to produce a new one, which finally can be sent to the object (figure 2).

Broll [2] presented some strategies to solve concurrent interaction over a single object. Margey [6] presents a study concerning about the different aspects of cooperative interaction but does not deal with immersive environments. Noma [7] presents a study about cooperative manipulation using force feedback devices.

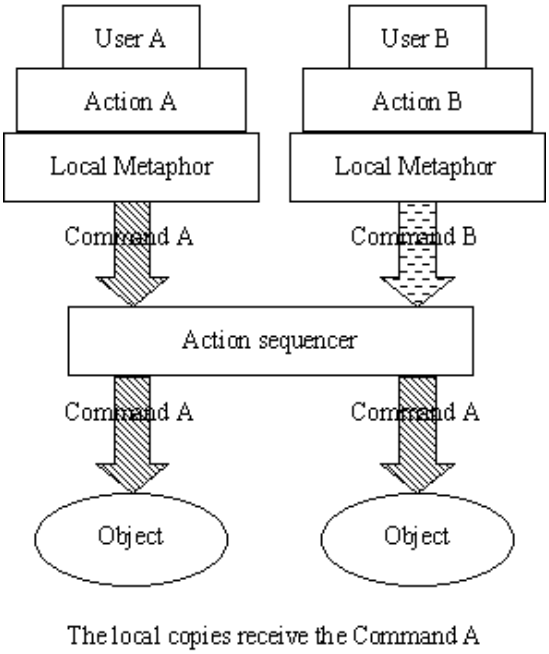


Figure 1: Selection of the highest priority action

The Proposed Architecture

Our goal is to develop a framework for supporting simultaneous interaction between two users (manipulating the same object) inside a totally immersive virtual environment. The user wears a non-transparent HMD, and magnetic position trackers capture his hand and head movements. This proposed architecture intends to solve two fundamental problems in this kind of interaction:

- How to show to one user the action of the other?
- How to combine the commands applied to an object by different users?

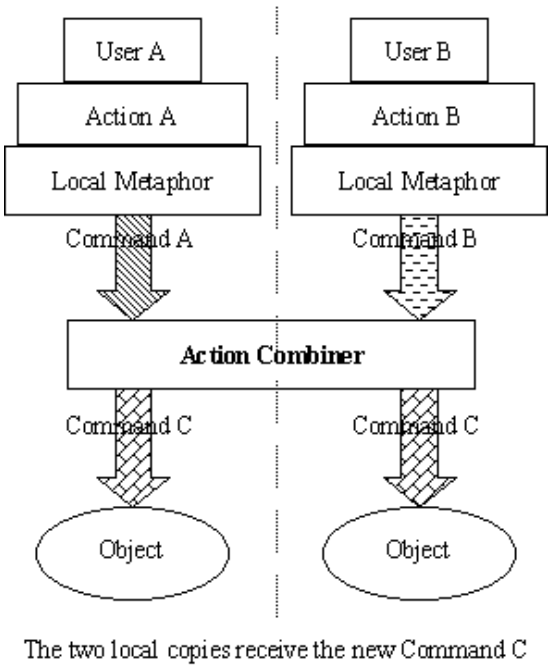


Figure 2 – A new command is generated and sent to the object

A secondary aim is to build a flexible architecture that preserves the interaction metaphor used by each user.

The following modules compose our framework (figure 3):

- Input Interpreter;
- Command Combiner;
- Manipulation Information Generator;
- Movement executor.

The **Input Interpreter** module is responsible for "understanding" the movements received from one user and translating them into **Commands** to the virtual object. This translation is tightly based on the interaction metaphor that is being used in the virtual environment. For example, if one is using a ray-casting metaphor to point to an object, this movement will mean to select it, otherwise, if a direct manipulation metaphor is being used, it could mean a translation. To allow the correct synchronization between users actions, each command generated by the Interpreter is followed by a time-stamp.

The **Command Combiner** is responsible for joining the Commands received from two Input Interpreter and, based on a **Collaborative**

Metaphor, generates a new command to be applied to the object. This **Collaborative Metaphor** should be defined based on the combination of the interaction metaphors that are being used by the users.

The **Manipulation Information Generator** will be responsible for showing to one user the actions accomplished over the object by the other one. In the real world, when we work cooperatively on an object the forces applied to the object are transferred to the other user through the object body. On the other hand, in virtual environments without force feedback devices, this transmission is not feasible. So, we have to find some alternative ways to carry this information from one user to the other. Margery calls this the Activity Metaphor [6]. We intend to test the use of arrows (vectors) exhibited on the object surface to represent the magnitude and the direction of the applied forces. We will try also to use colors to enhance the perception of the same information.

The **Movement Executor** receives from the **Command Combiner** the movement to be applied to the object. The movement execution itself must be done by the graphical system supporting the user application. In addition to these modules there will be a control system that will manage the communication and rendering aspects, but as these are not collaborative subjects they are out the scope of this paper.

Conclusion

This work presented an architecture to allow the simultaneous interaction of two users over a single object. The proposed system is based on the **Collaborative Metaphor** concept that allows joining multiple user interaction into commands directed to a single object.

In the future we intend to study how to combine different kinds of interaction metaphors to create new collaborative metaphors.

References

[1] Billinghurst, M. “Shared Space: An Augmented Reality Approach for Computer Supported Cooperative Work”. Virtual Realty. 1998. Vol. 3, No. Springer.

[2] Broll, W. “Interacting in distributed collaborative environment” IEEE VRAIS’ 95. p.148-155. IEEE Computer Society Press, 1995.

[3] Broll, W. et al “The Virtual Round Table - a Collaborative Augmented Multi-User Environment”. ACM Collaborative Virtual Environments, 2000, pp. 39-46.

[4] Dumas, C. et al “Spin: a 3d Interface for Cooperative Work”. Virtual Reality. 1999. Vol. 4, No.1. Springer. pp. 15-25.

[5] Kiyokawa, K. Takemura, H. Yokoya, N. “Seamless Design for 3D object Creation”. IEEE Multimedia. Jan 2000.

[6] D. Margery, B. Arnaldi, N. Plouzeau. A General Framework for Cooperative Manipulation in Virtual Environments. *Virtual Environments '99*, M. Gervautz, A. Hildebrand, D. Schmalstieg (eds.), Springer, pages 169-178, 1999.

[7] Noma, Miyasato,"Cooperative Object Manipulation in Virtual Space using Virtual Physics", ASME-DSC-Vol.61, pp.101-106,1997

[8] Ohshima, T. et al. “AR2 Hockey: A case study of

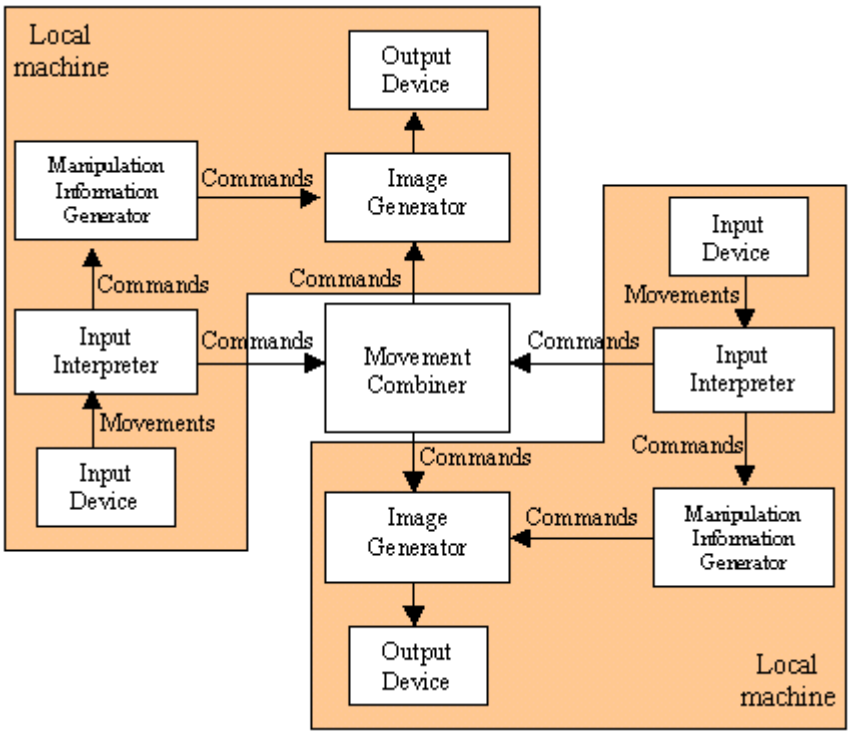


Figure 3: Proposed framework

collaborative augmented reality". IEEE VR' 1998. pp 268-275.

[9] Szalavari, Z., et al. "Studierstube: An environment for collaboration in augmented reality". Virtual Realty. 1998. Vol. 3, No.1. Springer. pp. 37-48.

[10] Vuillème, A. et alii, "Nonverbal Communication Interface for Collaborative Virtual Environments". Virtual Realty. 1999. Vol. 4, No.1. Springer. pp. 49-59.

Augmented Reality, Mobility and Tracking

Chair: Tobias Höllerer, Columbia University

Enhancement of Robotic/Telemanipulator Surgery of the Heart by AR/VR: a Surgeon's Dream

Robert Bauernschmitt, Hormoz Mehmanesh, Rüdiger Lange

German Heart Center Munich, Clinic for Cardiovascular Surgery
Lazarettstr. 36
80636 Munich, Germany
Bauernschmitt@dhm.mhn.de

Until recently, endoscopic methods were not used in heart surgery because the dexterity and the degrees of freedom of movements provided by conventional endoscopic instruments were not sufficient for performing surgery on vessels with a 1-2 mm diameter

To overcome the problem of imprecision, telemanipulator systems have been developed to faciliate endoscopic cardiac surgery. These systems consist of three main components: a surgical console (the man-machine-interface), the computer controller and the robotic arms holding specifically designed endoscopic instruments. At present, surgical telemanipulators are distributed by two companies: Computer Motion Inc. („Aesop" or „Zeus" – system) and Intuitive Surgical (DaVinci" – system). The main difference between the two sysytems are the degrees of freedom (DOF) of the instruments: while the „Zeus" has only 4 DOF, the „DaVinci" offers 6 DOF and is able to mimick closely all movements of the surgeon's hand. The increase of the DOF allows procedures virtually impossible with 4 DOF, like for example, throwing surgical stitches on a line perpendicular to the instrument shafts. In addition, time is saved because of a decreased necessity for instrument changes.

The surgeon manipulates traditionally designed instrument handles at the console, his movents are relayed in real-time to the robotic arms. The vision of the surgical field is displayed by an endoscopic cameras with 3-10 – fold magnification to a screen at the console.

The visualisation system is made up of two 3-chip cameras mounted on a 3-D-endoscope with two separate optical channels. Independently acquired images are transmitted to a high resolution binocular display of the operative field, which is not like a computer screen, but rather allows „visual immersion" of the surgeon into the body cavities.

Three ports, one for the camera and two for instrument arms, are placed into the patient's thorax through 1cm-incisions. The telemanipulated instruments supported by mechanical arms are articulated at their distal extremity so as to reproduce the motion of the surgeon's hands. As soon as the instruments are connected to the arms and introduced into the body, the surgeon unlocks the man-machine interface, inserts his thumbs and index fingers into velcro straps of specifically designed joysticks at the console giving him the impression to hold a familiar surgical instrument (like, for example, a pair of scissors) and starts the operation from the console. At any time during the procedure, it is possible to disconnect the „Master" (Surgeon's handgrips) from the „Slave" (Robotic arms), enabling repositioning of the master handles within the work space while the position of the instruments remains unchanged. This allows to always work in the most favorable ergonomic position and provides optimal hand-eye alignement.

When the 4-DOF-camera has to be manipulated, the surgeon locks the slaves via a footswitch. From the surgeon's perspective, his hands now seem to be attached to the image of the surgical field. Translating the joysticks to the right results in movement of the image to the right by sweeping the endoscope to the left. Likewise, up and down and zooming movements are performed from the console.

The software provides tremor elimination (> 6 Hz), another crucial issue in conventional endoscopic surgery: the longer the instruments shafts are, the more pronounced is tremor transmission to the tip of the tool. Heart surgery handling with structures of a 1-2 mm diameter therefore is almost impossible without tremor filtering. Motion scaling can be chosen between 1:1 and 1:5.

The most common operation with the system is a so-called „single-bypass" procedure: an artery of the inner chest wall is endoscopically

dissected and connected with a diseased coronary artery. In addition, several double-bypass and even a limited number of multiple bypass operations using one or two chest wall arteries was possible. Closure of congenital defects in the septum of the heart and even heart valve surgery are further targets for endoscopic procedures. In March 2000, the first total endoscopic heart valve operation was successfully completed in the German Heart Center Munich.

Despite this remarkable progress, telemanipulator surgery is still at its beginning and only applicable in a small number of patients. Several problems of the system probably can be solved by the introduction of AR/VR methods into the procedure.

Two major shortcomings have to be solved:

Performing aortocoronary bypass surgery, one of the major problems is finding the target vessels of the heart. Usually, coronary vessels are embedded into a layer of fat tissue preventing direct inspection and have to be dissected free. Correct identification of the target vessel is further impeded by the unfamiliar angle of inspection and the magnification provided by the endoscopic camera, sometimes leading to connection of the graft to a wrong, smaller vessel. A possible solution to this problem could be scanning of the heart surface with a ultrasonic Doppler probe; with this method, localisation and size of the arteries and the thickness of the tissue layer above them can be determined. Doppler probes are small and can be introduced through the ports. Ideally, this Doppler picture should be integrated into the surgeon's realistic view provided by the camera.

The majority of cases has to be performed with the help of a heart-lung machine and temporary cardiac arrest. Surgery on a beating heart is extremely difficult and associated with a decrease of precision. If it was possible to share control between the user (surgeon) and the system (compensating the pulsations of the heart), beating heart surgery should be possible with a high level of precision. The dream is performing surgery on a virtually non-moving field, while the heart is still normally beating.

Summarizing, this paper does not add new techniques to VR/AR methods, but rather offers an application. Integration of VR/AR into this emerging new surgical discipline is considered the main goal for the future in order to provide these minimally invasive techniques to an increasing number of patients.

A General Framework for Multi-Modal Interaction in Virtual Reality Systems: PrOSA

Marc Erich Latoschik

AI & VR Lab6, Faculty of Technology
University of Bielefeld, Germany
marcl@techfak.uni-bielefeld.de

Abstract

This article presents a modular approach to incorporate multi-modal – gesture and speech – driven interaction into virtual reality systems. Based on existing techniques for modelling VR-applications, the overall task is separated into different problem categories: from sensor synchronisation to a high-level description of cross-modal temporal and semantic coherences , a set of solution concepts is presented that seamlessly fit into both the static (scenegraph-based) representation and into the dynamic (render-loop and immersion) aspects of a real-time application. The developed framework establishes a connecting layer between raw sensor data and a general functional description of multi-modal and scene-context related evaluation procedures for VR-setups. As an example for the concepts usefulness, their implementation in a system for virtual construction is described.

Introduction

The development of new interaction techniques for virtual environments is a widely recognized goal. Multi-modality is the keyword that suggests one solution for getting rid of the WIMP[1]-style point-and-click metaphors still found in VR-interfaces. The more realistic our artificial worlds become, the more seem our natural modalities gesture and speech to be the input methods of choice, in particular when we think in terms of *communication* and further the possible incorporation of lifelike characters as interaction mediators. Our goal is the utilisation of multi-modal input in VR as an spatial represented environment. Considering the latter, Nespoulou s and Lecour [4] proposed a gesture classification scheme that perfectly describes possible *coverbal* gesture functions when they specify illustrative gestures as:

- *Deictic*: Pointing to references that occur in speech by respective lexical units.
- *Spatiographic*: To sketch the spatial configuration of objects referred to in speech.
- *Kinemimic*: To picture an action associated with a lexical unit.
- *Kinemimic*: Describing the shape of an object referred to in speech.

We are exploiting these gesture types with slight adaptations for enabling basic multi-modal interaction.

Work is done on both sides, on the development of multi-modal interpretation and integration (MMI) and on the enhancement of VR-technology and realism. Despite this fact, there are few approaches that deal with the systematic integration of the MMI-results under general VR-conditions. The latter justifies the foundation for the development of the PrOSA (Patterns On Sequences of Attributes) [3] concepts as fundamental building blocks for multi-modal interaction in VR.

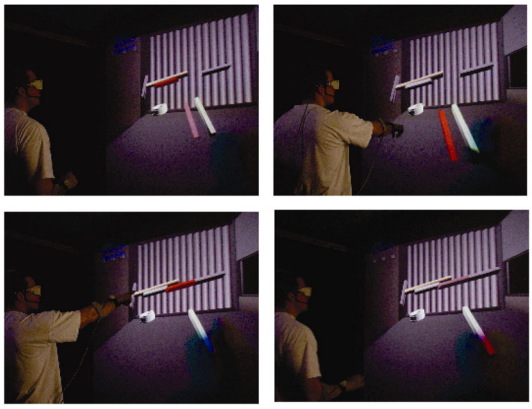


Figure 1: Multi-modal interaction: a user speaks and gestures to achieve a desired interaction, in this case the connection of two virtual objects.

[1]WIMP: Windows, Icons, Menu, Pointing

Gesture Processing

Gesture detection heavily depends on sensor data, which – in general – is neither synchronised nor represented with respect to a common base. There is no agreement about the body data representation, which is suitable in gesture detection tasks: depending on the detection frame work, it is often necessary to abstract from specific numeric sensor-data (e.g. the position of one or several 6DOF sensor or the output of camera-based systems) and to consider relevant quantified movement information:

static and dynamic attributes like finger-stretching, hand-speed, hand-head-distance etc., which PrOSA encapsulates in attribute-sequences, containers that establish a data flow network between a hierarchy of different modular calculation components, which are necessary for gesture analysis and detection. A schematic overview of the concepts and their cooperation in the network is shown in Figure 2:

Actuators

On the basic hierarchy level, the attribute-sequences are anchored in so-called actuators. In our approach actuators are entities that hide the sensor layer and provide reliable movement information even under unreliable frame-rate conditions. This is achieved through asynchronous sensor input and the prohibition of data extrapolation, which is particular important for trajectory interpretation. Actuators perform the following necessary steps to abstract from sensor data:

- Synchronisation
- Representation to a common base
- Processing: transformation, combination, etc.
- Qualitative annotation

Actuators come in different flavours due to their output data format and the number of incoming sensor channels attached to them (s. fig. 2). Important examples are handform-actuators (providing information about finger bending and the angles between adjacent fingers) or single- and multichannel[2] NDOF-movement-actuators for significant body points (fingertips, wrists, head) and associated reference rays[3], line segments that represent deictic or iconic directional and orientational information (pointing direction, palm normal, etc.). Actuators deliver a set of their resulting synchronous movement samples for each frame and feed them into higher level processing units like **detectors** (and their subtypes) or into **motion-modificators**

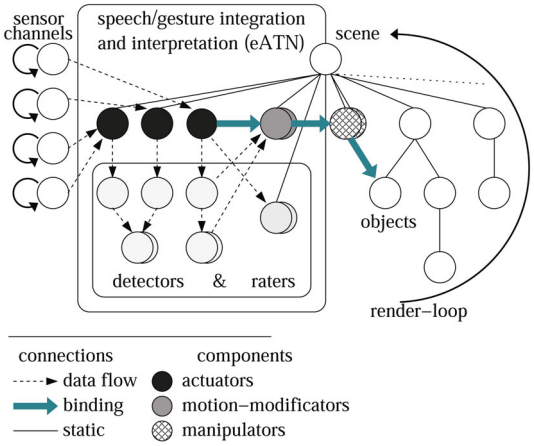


Figure 2: A schematic view of the main PrOSA concepts, their relations and the data flow between the different components

for an ongoing interaction processing.

Detectors

To classify the gesture movements, the incorporated gesture detection relies on template matching of eight spatio-temporal movement features (like shape, dynamics and trajectory changes):

1. Stop-and-Go
2. Leaving an associated rest position
3. Definite shape
4. Primitive movement profile
5. Repetition
6. Internal symmetry
7. External symmetry
8. External reference

The actuators deliver the preprocessed movement data to detector networks. Detectors of different kinds handle basic calculation tasks

[2]Depending on the number of sensor channels the actuator processes.

[3]A ray in the ideal sense. In reality, line segments are used.

and operations for all necessary basic datatypes (real numbers, vectors, quaternions and 4x4 CGM's[4]):

- Addition, subtraction or multiplication etc.
- Threshold tests and comparison operators
- Boolean operators
- Buffering of values over an interval
- etc.

Each detector has just a simple function. But complex calculation networks can be constructed to detect the given spatio-temporal features using multiple detectors. And because the calculation arithmetic is hidden in the data flow network structure, it is easy to modify. Detectors can be added, exchanged, their parameters can be altered or they can be deleted at all. E.g.: To detect definite shape of the hands, handform-actuators feed simple threshold detectors which themselves feed into boolean and/or detectors. A pointing posture can then be defined by the stretching of the index finger in addition of a bending of the other fingers. In combination with a movement stop of the fingers as well as the whole hand this gives a high likeliness that a pointing gesture occurred. A graphical example of a two layer network is again shown in figure 2.

Raters

Raters are a different kind of detectors. They are not concerned with gesture detection but with scene analysis and object rating (hence raters. To handle multi-modal input for deictic utterances, e.g. “Take [pointing] that blue wheel over there...”, it is necessary to process fuzzy input (in this case a pointing gesture) and combine it with the speech analysis. For this purpose the actuators deliver line segments which represent the users view and pointing direction _ an examples of reference ray usage. One type of rate-detectors will handle that input and estimate the difference between the segments and sort the scene objects according to the resulting difference values. This information is one important basis for the multi-modal interpretation.

[4]Computer Graphics Matrix

Motion-Modificators

Interaction is accomplished in two ways: discrete if the utterances form a complete interaction specification or continuously if there is information missing but if there is an ongoing gesture that can be associated with the desired manipulation type. In the latter case motion-modificators abstract from unprecise user movements and map them continuously to precise changes of the virtual scene. The following *binding* will be established: an actuator routes data through a motion-modificator to an appropriate manipulator to map the movement to an attribute change (a *mimetic mapping*). This

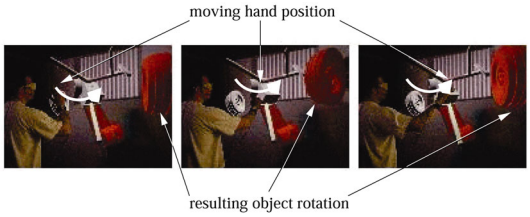


Figure 3: Continuously rotating an object (a wheel). A kinemimic/mimetic gesture is used to achieve the desired object manipulation.

data flow is shown by the binding arrows in figure 2. A sequence of a resulting manipulation is presented in figure 3.
The duration of the binding is defined by the duration of the ongoing movement or the interception by external events. To be more specific: A movement pattern consists of several constraints calculated by a detector network, e.g. a rotation of one hand is defined by:

- static hand form
- continuous movement speed
- movement in one plane
- adjacent strokes have similar angles (not 180°)

These informal descriptions5 are translated into geometric and mathematical constraints based on actuator data to construct the resulting detector network. The motion-modificator receives the calculation results for each frame and keeps on working until the constraints are no longer satisfied. Another method to interrupt an ongoing manipulation is by external signals from the multi-modal interpretation, e.g. when the user utters a “stop” or similar speech commands.

[5]A formal rule-based description can be found in [3]

Motion-modificators map imprecise or coarse movements to precise object changes. To achieve this type of *filtering*, we need to monitor specific movement parameters - e.g. a rotation axis or a direction vector - and to compare them to a set of possible object parameters to modify. Therefore, when the binding is established, each motion-modificator receives a set of parameters that can be seen as *changegrid* members. For every simulation step (for every frame) they are compared to the actual movement parameters and the closest one (e.g. in the case of vectors: the one with the minimal angular divergence) is chosen as the target parameter. This results in the desired filtering. To apply the parameter change to an object, a specific instance of basic **manipulators** receives frametime-adequate manipulation commands from the motion-modificator and changes the object parameter. In addition, by partitioning this operation using two different concepts, it is not only possible to establish a mimetic mapping: You could for example combine a rotation motion-modificator with a color or a sound manipulator. This would result in a kind of *metaphorical mapping*, an ongoing movement results in a color or sound change.

Multi-Modal Interpretation

The interpretation process of gesture/speech-related utterances can handle temporal as well as semantic relations. An enhanced ATN[6] formalism has been developed to achieve the incorporation of temporal cross-modal constraints as well as to evaluate scene-related context information in real-time and to latch the interpretation into the driving render-loop (e.g. by using raters). The latter emphasizes the fact that the actual user's viewing perspective determines the reference semantics of all scene-related utterances dynamically. Their interpretation - the deictic *mapping*[3] - depends on both time-dependent dynamic as well as on static scene- and object-attributes like in: "*Take* [pointing] *that left blue big thing and turn it like* [rotating] *this*". Appropriate verbal (e.g. colors and positions) and gestural (e.g. view- and pointing direction) input is disambiguated during user movements through a tight coupling in the scene representation (see. Raters) and results are stored in so called **spacemaps**. Fig. 4 shows the migration of one specific object (the black oval) in a spacemap during interaction and user movement. Every row represents the result for one simulation step. The first entry in each row holds additional data (time, segment, etc.).

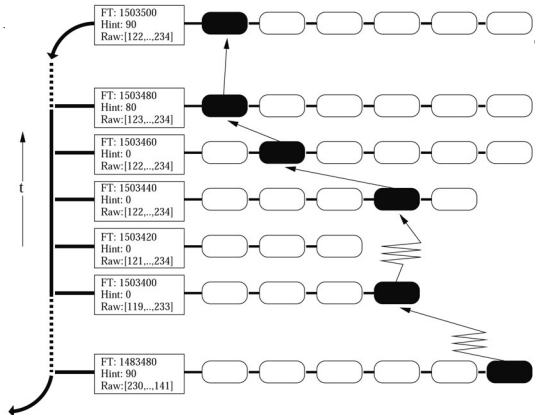


Figure 4: A spacemap as a temporal memory. The relative position of an object to a line segment is represented for every simulation step.

This pre-processing allows the handling of varying multi-modal temporal relationships (e.g. a look-back) without the necessity for buffering all scene descriptions for past frames. In addition, the enhanced ATN allows to express application logic in the same representation as the multi-modal integration scheme. This results in a convenient way to adapt multi-modal interfaces to different applications.

Implementation and application

Figures 1 and 3 show a sequence during user interaction with the virtual construction application [1][7]. In addition to pure speech commands (for triggering actions like opening of doors etc.), basic interactions are enabled using gesture and speech. Objects can be instantiated and connected as well as referenced and moved around with distant *communicative* interaction and with direct manipulation (if desired). There are actuators, detectors and motion-modificators to trigger and evaluate deixis (view, pointing), kinemimic/mimetic gestures (rotating of the hands), grasping and several more symbolic gestures. Work is on the way to add pictomimic/spatiographic (iconic) gestures and to incorporate an articulated figure [2] as well as to test an unification based speech/gesture integration using the existing framework. The goal is to develop a toolkit set of basic PrOSA-concept instances to establish detection networks for often and commonly needed standard interactions in virtual environments. The speech recognition system is a research prototype and works speaker-independent. The current PrOSA-concept implementation makes use of, but is not

[6]Augmented Transition Network

[7]This work is partially supported by the *Virtuelle Wissensfabrik* of the federal state North-Rhine Westphalia and the Collaborative Research Center SFB360 at the University of Bielefeld.

limited to the AVANGO-toolkit [5]. The data flow has been established using field connections (a concept similar to the one found in VRML97). All components can be constructed and all connections can be established by the AVANGO-internal scripting language Scheme, which allows on-the-fly changes and a rapid prototyping approach for new projects. Particular design efforts have been made to achieve portability by an explicit formal definition of all concepts by taking general VR-conditions into account.

References

[1] Bernhard Jung, Stefan Kopp, Marc Latoschik, Timo Sowa, and Ipke Wachsmuth. Virtuelles Konstruieren mit Gestik und Sprache. Künstliche Intelligenz, 2/00:5_11, 2000.

[2] Stefan Kopp and Ipke Wachsmuth. A knowledge-based approach for lifelike gesture animation. In W. Horn, editor, ECAI 2000 - Proceedings of the 14th European Conference on Artificial Intelligence, pages 663_667, Amsterdam,2000.

[3] Marc Erich Latoschik. Multimodale Interaktion in Virtueller Realität am Beispiel der virtuellen Konstruktion. PhD thesis, Technische Fakultät, Universität Bielefeld, 2001.

[4] J.-L. Nespoulous and A.R. Lecours. Gestures: Nature and function. In J.-L. Nespoulous, P. Rerron, and A.R. Lecours, editors, The Biological Foundations of Gestures: Motor and Semiotic Aspects. Lawrence Erlbaum Associates, Hillsday N.J., 1986.

[5] Henrik Tramberend. A distributed virtual reality framework. In Virtual Reality, 1999. [6] http://www.techfak.uni-bielefeld.de/techfak/ags/wbski/wbski_engl.html.

Augmented Reality Interface for Visualizing Information Space

Ulrich Neumann

Computer Science Department
University of Southern California, USA

Introduction

Augmented Reality offers a key technology for solving the problems associated with visualizing information that is spatially and semantically distributed throughout our world. The advent of systems on a chip that include wireless links, image sensors, and general processing suggest that smart systems could become widespread and common place in the near future. However, extrapolating into this future, these same systems will create new interface problems for users trying to navigate the congested "info-spaces". In particular, mobile users will need a method for navigating and interacting with the information that is locally meaningful in an intuitive and customizable fashion. Furthermore, the proliferation of information sources may lead to a confusing myriad of diverse and non-intuitive user interfaces. With the capability for tracking a user's hand or head motions, Augmented Reality Interfaces (ARI) may provide the means for addressing these and other problems through a consistent and intuitive metaphor.

The concept of ubiquitous smart systems that sense and communicate with the environment is not new, and the concept is incrementally becoming a reality. Regardless of the projected uses for such devices, sensing and interacting with people is likely to be essential. As examples, consider smart kitchen appliances; virtual reality home entertainment systems; a distributed office locator system that indicates when your co-worker left a meeting room and which way he or she is heading down the hall; a distributed embassy surveillance system tracking visitors with selected behavior profiles; a 3D immersive teleconference system; or a computer-generated avatar for technical training and instruction. All of these examples require recognizing and tracking objects or motions of some sort. This sensing and interaction is spatially localized around the system and user. As users move through the world, they need a way to gracefully establish and break communication with these systems. Once communication is established, interactions should be intuitive and consistent from system to system. An augmented or mixed reality metaphor offers these capabilities as well as privacy and personal customization.

AR Interface Operation

Consider a hand-held Augmented Reality Interface (ARI) as illustrated in Fig.1. (Think of a future PalmPilot or cell-phone with a video camera and display.) Imagine that as the user moves this device through the world, information services available in the immediate space are shown on the display. For example, in your office, the news-service node advertises the current market quotes and headlines that match your interests. At the same time, the email portal announces five pending messages. This and other information is available in the space you currently occupy via low power RF or IR connections. You select the email portal by tapping the display or pointing the ARI at the physical device. In the latter case your gesture motions and pointing direction are sensed in the ARI by ceramic or MEMS rate gyroscopes and

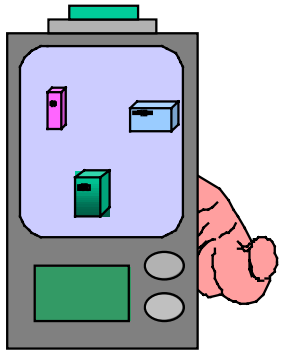


Figure 1: A hand-held ARI incorporates a video camera and display

the video motion field. Your customized email browser appears with the messages and you select the messages of interest by voice input,

screen taps, or ARI gesture. An urgent message requires that you converse with several colleagues about a design problem on product-X so you end the session with a wave of the ARI and walk to the lab where the latest prototype sits on a table. You gesture to the 3D teleconf node in the ceiling and call the three design leaders. Each person appears on the ARI and you place their images in positions around the prototype. By aiming the ARI, around the table you can observe the participants talking about and pointing to features of the real product before you. By the conclusion of the discussion, virtual annotations attached on or about the real product-X prototype summarize the discussion. The annotation data is copied and pasted into the email node with a sweep of the ARI followed by a verbal selection of the message recipients.

The example above illustrates some of the future possible ARI functionality Rather than

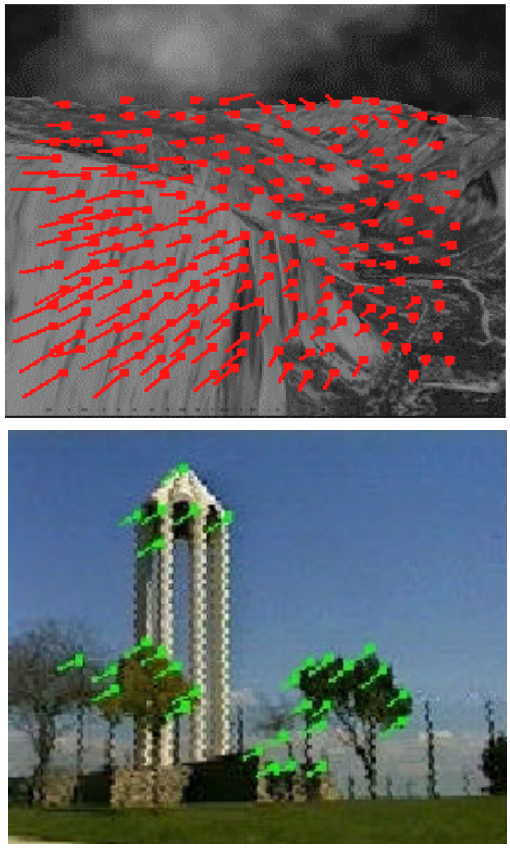


Figure 2: Motion fields computed with hybrid algorithm

consider all the needed ARI technologies, we focus on tracking. Simple short gestures can be tracked with rate gyroscopes. A drift rate of a few degrees per second is easily obtained from low-cost sensors and sufficient for sensing one to two-second duration hand gestures. Over longer periods, tracking ARI orientation requires additional information. The video camera and display is used to observe the real scene as well as the virtual information superimposed on the scene. This same imagery is useful for tracking motion, position and orientation.

Motion Tracking

Camera motion tracking relies on measuring the 2D-image motion field. Feature tracking or optical flow are both suitable approaches. Both require substantial computation and data bandwidths. A nominal color video image has 640x480 color pixels that arrive at 30 frames per second. The aggregate bandwidth assuming 8-bits per RGB color channel is 640x480x3x30 = ~28Mbytes/sec. While this data bandwidth itself is manageable, performing substantial computing on this data stream is not. Assuming each pixel is touched in the computation, an average of ~100ns are allotted for moving and computing each 3-byte RGB pixel. Consequently, all but the most trivial video processing is usually done off-line today.

Real time video processing today depends on dedicated hardware, but this is only available for high-volume standardized-algorithm applications such as JPEG or MPEG image compression. Algorithms for motion tracking require more general and varied processing characterized by substantial decision-making, iterations, and irregular addressing. To date, no dedicated-hardware processing system has succeeded in attracting significant users for motion tracking or other high-complexity vision tasks. When high performance is needed, the available technology has been too restrictive for general algorithms. For example SIMD arrays or fixed data-path processor pipelines can implement portions of tracking algorithms, but general processing and algorithm flexibility must be obtained elsewhere. In the examples in Figure 2 a hybrid approach using optical flow and feature tracking is used to compute the image motion fields [6]. The algorithm performs iterative refinement of its estimates and uses affine (6 parameter) warping of 32x32 image regions to assess tracking accuracy. The result is a high-quality motion field, but the full algorithm runs on a 640x480 image at a rate of < 1 frame per second on a 500 MHz PentiumIII. Since this is only one-step in an ARI motion tracking method, even a 30-50x improvement still leaves no time for other tasks. Realistically, a factor of 300 is more likely needed for a complete ARI vision tracking

system. (Note that Moore's law predicts it will take 7-8 years to get a 30x computer speedup.)

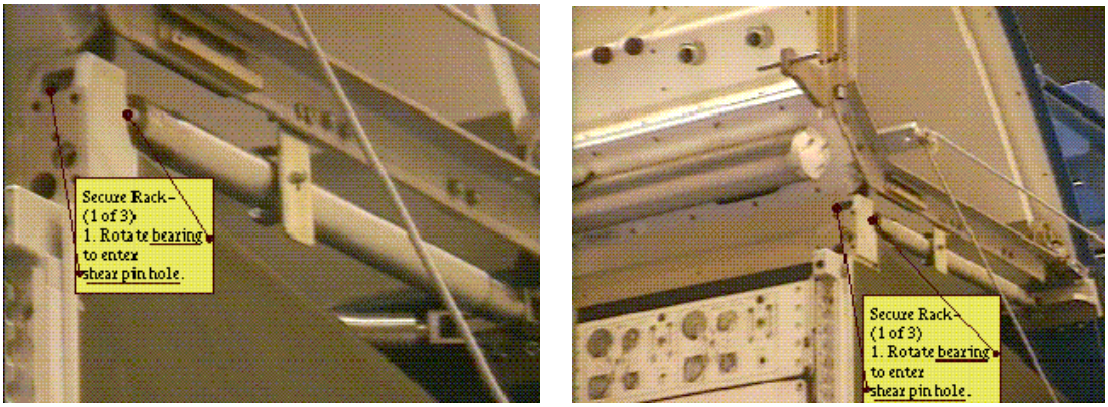


Figure 3: Tracked ARI keeps rack annotations attached to correct features regardless of ARI motion. The raw video sequences were obtained at NASA JSC, Houston

Position and Orientation Tracking

Given a 2D-motion field, the camera's incremental rotation and translation direction can be computed using matrix inversions or recursive filters. Filters are generally faster and more easily implemented in hardware [2]. Integration of the incremental motions can provide an absolute camera pose assuming the translation speed is known. This speed can not be recovered from a 2D-motion field so it must come from other sensors or knowledge of the 3D coordinates of observable points [3]. A few initial calibrated points can be used to calibrate unknown points making, them in turn, suitable for calibrating still others [4]. This recursive autocalibration of the 3D coordinates of observable features requires Kalman Filters and adaptive heuristic decision logic. Unobtrusive features on future smart devices may serve to initialize absolute ARI pose tracking.

Future pose tracking systems are likely to use multiple sensors [1]. Gyroscope data and vision data are complementary, aiding each other in pose tracking. Fusing data from these sensors with Kalman filters has been successful [5]. Integration with accelerometers is more difficult due to their high drift rate, however a closed-loop stabilization of that drift may be possible in the future using real-time vision.

Conclusion

In closing an example of ARI application prototypes are shown in Figures 3. Figure 3 illustrates how tracking is used to annotate a "smart" equipment rack. When viewed with an ARI, the rack reveals its installation and operation information to the user. Note that the instruction text is kept in alignment with the rack features over widely varied viewing poses.

Although pose tracking is only one of the many elements needed to develop ARI systems, it is clearly one of the most compute-intensive. ARI systems are intended for interactive "man-in-the-loop" use, so real-time *and* robust performance is critical [7]. This combination requires high performance and flexible computing solutions simply not available today.

References

[1] R. T. Azuma. A survey of Augmented Reality. Presence: Teleoperators and Virtual Environment, 6(4), pp.355-385, Aug. 1997

[2] G. Welch, G. Bishop, SCAAT: Incremental Tracking with Incomplete Information, Proceedings of Siggraph97, Computer Graphics, pp. 333-344

[3] YoungKwan Cho, Scalable Fiducial-Tracking Augmented Reality, Ph.D. Dissertation, Computer Science Department, University of Southern California, January 1999

[4] J. Park, B. Jiang, and U. Neumann. "Vision-based Pose Computation: Robust and Accurate Augmented Reality Tracking," IEEE International Workshop on Augmented Reality, Oct. 1999, pp. 3-12

[5] Y. Bar-Shalom X. Li, Estimation and Tracking: Principles, Techniques, and Software, Artech House, Norwood MA, 1998

[6] U. Neumann, S. You, "Natural Feature Tracking for Augmented Reality," *IEEE Transactions on Multimedia*, Vol. 1, No. 1, pp. 53-64, March 1999.

[7] U. Neumann and A. Majoros, "Cognitive, Performance, and Systems Issues for Augmented

Reality Applications in Manufacturing and Maintenance," *IEEE VRAIS '98*, pp. 4-11, 1998.

Mobile 3D City Info

Ismo Rakkolainen

Digital Media Institute,
Tampere University of Technology,
P.O.BOX 553, 33101Tampere, Finland
ira@cs.tut.fi

Extended Abstract

We have implemented a mobile 3D City Info. It answers to practical questions like “Where is a certain shop? Where am I now?” It connects a service database (restaurants, hotels, shops, etc.) to a VRML model of a city. The user can query various information of a real city. The system s the query results dynamically with interconnected 3D world and 2D map.

The highly realistic model provides a 3D user interface and an inherent spatial index for the city. It helps in navigation and orientation.

Soon 3D rendering and broadband wireless communications (UMTS etc.) will be embedded into various handheld devices. We customized the system for mobile laptop users by integrating a GPS and a digital compass to it (Figure 1).

Figure 1: The mobile 3D city info laptop and GPS.

For the field tests, the ease of use and simulation of the actual user situation of the future devices was highly important. As the laptop-GPS combination is not very handy in real life situations, and 3D in PDAs is currently very slow, we made a fake "mockup" PDA version of the system.

Figure 2: The usability test with the fake PDA city info.

Only HTML code and images was used (Figure 2). The image-based version was adequate for the purposes of the well-focused and restricted usability test.

Our results show that search and visualization of location-based information of a city becomes more intuitive with life-like 3D. The users prefer 3D over 2D, although both may be needed. The visual similarity with reality helps in finding places virtually and afterwards in real life.

Our system is an early prototype of the future mobile 3D services. They can use the user's location and time-, location-, and context-based information. We see our application as one of the most useful and practical uses of 3D in near-future PDAs and other mobile devices.

We are looking for better and more intuitive user interfaces for the application and for mobile devices in general. What would be a more intuitive way of moving around a virtual city than driving? In search of alternate user interfaces for home users, we have built a steering wheel navigation and an immersive interface with a HMD.

Another example of the future possibilities is shown in Figure 3. It is a small, hand-held navigation or computer stick with an in-scrollable, non-rigid display, and built-in GPS and compass.

Figure 2: A vision of the future 3D/3G device with GPS. The device has a flexible, inscrollable display.

VR and AR are interesting ways to go. Many possibilities become feasible within a few years due to mobile hardware improvements. 3D graphics can even change the current user interface paradigms. We believe that 3D graphics will be important for the future concepts of mobile personal communications

The on-line version is at http://www.uta.fi/hyper/projektit/tred/.

.

Mixed Reality and Natural Interaction

Chair: Wolfgang Broll, GMD-FIT

Tiles: Toward Generic and Consistent MR Interfaces

Ivan Poupyrev,[1] *Desney Tan,*[2] *Mark Billinghurst,*[3]
Hirokazu Kato,[4,6] *Holger Regenbrecht*[5]

[1]Interaction Lab
Sony CSL
3-14-13 Higashi-Gotanda
Tokyo 141-0022
Japan

[2]School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA

[3]University of Washington
[4]Hiroshima City University
[5]DaimlerChrysler AG
[6]ATR MIC Labs

poup@csl.sony.co.jp, desney@cs.cmu.edu, grof@hitl.washington.edu,
kato@sys.im.hiroshima-cu.ac.jp, holger.regenbrecht@daimlerchrysler.com

Abstract

Tiles, is a MR authoring interface for easy and effective spatial composition, layout and arrangement of digital objects in mixed reality environments. In Tiles we attempt to introduce a consistent MR interface model, that provides a set of tools that allow users to dynamically add, remove, copy, duplicate and annotate virtual objects anywhere in the 3D physical workspace. Although our interaction techniques are broadly applicable, we ground them in an application for rapid prototyping and evaluation of aircraft instrument panels.

Keywords

Augmented and mixed reality, 3D interfaces, tangible and physical interfaces, authoring tools

Introduction

Mixed Reality (MR) attempts to create advanced user interfaces and environments where interactive virtual objects are overlaid on the 3D physical environment, naturally blending with it in real time [1, 6]. There are many potential uses for such interfaces, ranging from industrial, to medical and entertainment applications [e.g. 2, 7, see also Azuma, 1997 for comprehensive survey]. However, most current MR interfaces work as information browsers allowing users to see virtual information embedded into the physical world but provide few tools that let the user interact, request or modify this information effectively and in real time [8]. Even the basic and generic interaction techniques, such as manipulation, coping, annotating, dynamically adding and deleting virtual objects to the MR environment have been poorly addressed.

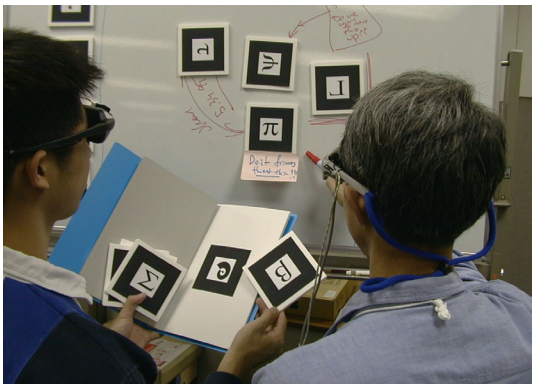


Figure 1: Tiles environment: users arrange data on the whiteboard, using tangible data containers, data tiles, and adding annotations using whiteboard pen.

The current paper presents *Tiles*, a MR authoring interface that investigates interaction techniques for easy spatial composition, layout and arrangement of digital objects in MR environments. In *Tiles* we attempt to design a simple yet effective interface, based on a consistent interface model, providing a set of tools that allow users to add, remove, copy, duplicate and annotate virtual objects. Although our interface techniques are broadly applicable, it has been developed for rapid prototyping and evaluation of aircraft instrument panels, a joint

research initiative carried out with DASA/EADS Airbus and DaimlerChrysler.

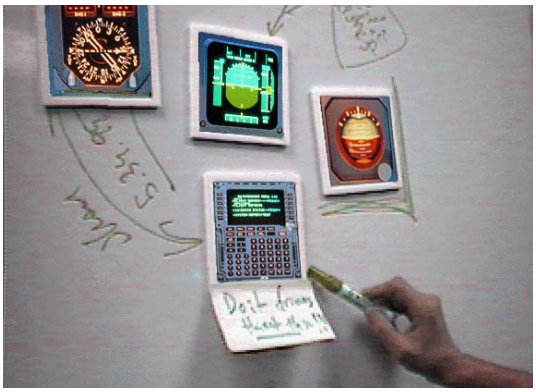


Figure 2: The user, wearing lightweight head-mounted display with mounted camera, can see both virtual images registered on tiles and real objects.

Related Work

The current design of MR interfaces, falls into two orthogonal approaches: tangible interfaces based on tabletop MR offer seamless interaction with physical and virtual objects but results in spatial discontinuities. Indeed, the users are unable to use MR environment beyond the limits of the instrumented area, e.g. a projection table. 3D AR provides spatially seamless MR workspaces, e.g. the user can freely move within the environment, tracked by GPS, magnetic trackers, etc. However, it introduces discontinuities in interaction: different tools has to be used to interact with virtual and physical objects, breaking the natural workflow. In Tiles we attempt to merges the best qualities of interaction styles: true spatial registration of 3D virtual objects anywhere in the space and a tangible interface that allows to interact with

Figure 3: The user cleans data tiles using trash can operator tile. The removed virtual instrument is animated to provide the user with smooth feedback.

virtual objects without using any special purpose

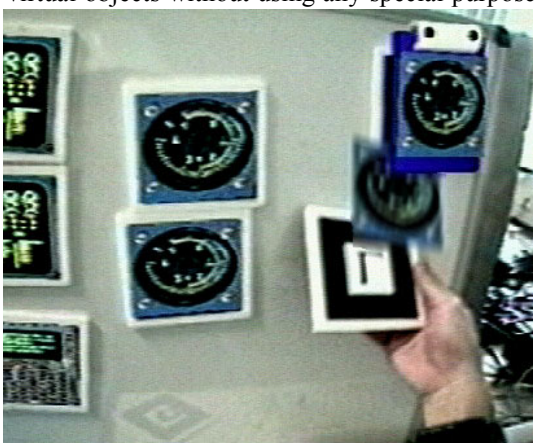


Figure 4: Coping data from clipboard to an empty data tile.

input devices. Tiles also introduces a generic interface model for MR environments and use it

for authoring applications. Although 2D and 3D authoring environments have been intensively explored in desktop and VR interfaces [e.g. 3, 5] there are far fewer attempts to develop authoring interfaces for mixed reality.

Tiles Interface

Tiles is a collaborative Tangible AR interface that allows several participants to dynamically layout and arrange virtual objects in a MR workspace. The user wears a light-weight head-mounted display (HMD) with a small camera attached. Output from the camera is captured by the computer which then overlays virtual images onto the video in real time and presented back to the user on his or her HMD (Figure 1 and Figure 2). The 3D position and orientation of virtual objects is determined using computer vision tracking, from square fiduciary markers that can be attached to any physical object. By manipulating marked physical objects, the user can manipulate virtual objects without need to use any additional input devices.

Interface

Basics: Tiles interface components

The Tiles interface consists of: 1) a metal whiteboard in front of the user; 2) a set of paper cards with tracking patterns attached to them, which we call tiles. Each of has a magnet on the back so it can be placed on the whiteboard; 3) a book, with marked pages, which we call book tiles, and 4) conventional tools used in discussion and collaboration, such as whiteboard pens and PostIt notes (Figure 1 and Figure 2). The whiteboard acts as a shared collaborative workspace, where users can rapidly draw rough layout of virtual instruments using whiteboard markers, and then this layout by placing and arranging tiles with virtual instruments on the board.

The tiles act as generic tangible interface controls, similar to icons in a GUI interface. Instead of interacting with digital data by manipulating icons with a mouse, the user interacts with digital data by physically manipulating the corresponding tiles. Although the tiles are similar to *phicons*, introduced in meta-Desk system [10], there are important differences. In metaDesk, for example, the shape of phicons representing a building had an exact shape of that building, coupling "bits and atoms" [10]. In *Tiles* interface we try to decouple physical properties of tiles from the data- the goal was to design *universal* data containers that

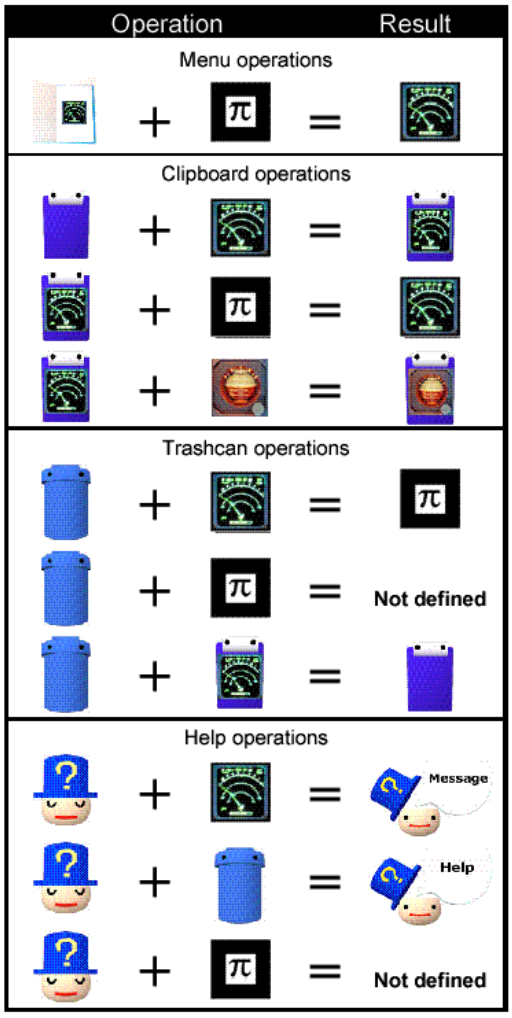


Table 1: Tiles operations: bringing together menu tile and empty data tile moves instrument on the tile (first row).

can hold any digital data or no data at all. Interaction techniques for performing basic operations such as putting data on tiles and removing data from tiles are the same for all tiles, resulting in a consistent and streamlined user interface. This is not unlike GUI interfaces, where operations on icons are the same irrespective of their content. Particularly, all tiles can be manipulated in space and arranged on the whiteboard; and all operations between tiles are invoked by bringing two tiles next to each other (Fig. 3).

Classes of tiles: data, operators and menu

We use three classes of tiles: data tiles, operator tiles and menu tiles. The only difference in their physical appearance is the icons identifying tile types. This allows users without HMD to identify the tiles.

- Data tiles are generic data containers. The user can put and remove virtual objects from the data tiles; if a data tile is empty, nothing is rendered on it. We use Greek symbols to identify the data tiles.

- Operator tiles are used to perform basic operations on data tiles. Implemented operations include deleting data from a tile, copying a virtual object to the clipboard or from clipboard to the data tile, and requesting help or annotations associated with a virtual object on the data tile.

- Menu tiles make up a book with tiles attached to each page (Figure 1). This book works like a catalogue or a menu: as the user flips through the pages, sees virtual objects and chooses the required instrument and then copy it from the book to any empty data tile.

For example, to copy an instrument to the data tile, the user first finds the desired virtual instrument in the menu book and then places any empty data tile next to the instrument. After a one second delay to prevent an accidental copying, a copy of the instrument smoothly slides from the menu page to the tile and is ready to be arranged on the whiteboard. Similarly, if the user wants to "clean" data from tile, the user brings the trashcan tile close to the data tiles, removing the instrument from it (Figure 3). Table 1 summarizes the allowed operations between tiles.

Implementation

The Tiles system is implemented using ARToolKit, a custom video see-through tracking and registering library [4]. We mark 15x15 cm paper cards with simple square fiduciary patterns consisting of thick black border and unique symbols in the middle identifying the pattern. In the current Tiles application the system tracks and recognize 21 cards in total. The software is running on an 800Mhz Pentium III PC with 256Mb RAM and the Linux OS. This produces a tracking and display rate of between 25 and 30 frames per second.

Conclusions

The Tiles system is a prototype tangible augmented reality authoring interface that allows a user to quickly layout virtual objects in a shared workspace and easily manipulate them without need of special purpose input devices. The interface model and interaction techniques introduced in Tiles can be easily expanded and extended to other applications. Object modification, for example, can be quite easily introduced by developing additional operator cards that would let the user dynamically modify objects, e.g. scale them. Although additional interaction techniques would allow Tiles to be used in other applications, in MR environments the user can easily transfer between the MR workspace and a traditional environments such as a desktop computer. Therefore, we believe that the goal of developing MR interfaces is not to bring every possible interaction tool and technique into the MR, but to balance and distribute the features between the MR and other interfaces. Hybrid mixed reality interfaces might be an interesting and important research direction [9] An interesting property of Tiles interface is also its ad-hoc, highly re-configurable nature. Unlike the traditional GUI and 3D VR interfaces, where the interface layout is determined in advance, the Tiles interfaces are in some sense designed by user as they are carrying on with their work.

References

[1] Azuma, R., A Survey of Augmented Reality. Presence: Teleoperatore and Virtual Environments, 1997. 6(4): pp. 355-385.

[2] Bajura, M., Fuchs, H., Ohbuchi, R., Merging Virtual Objects with the Real World: Seeing Ultrasound Imagery Within the Patient. Proceedings of SIGGRAPH ™92. 1992. ACM. pp. 203-210.

[3] Butterworth, J., Davidson, A., Hench, S., Olano, T., 3DM: a three dimensional modeler using a head-mounted display. Proceedings of Symposium on Interactive 3D graphics. 1992. ACM. pp. 135-138.

[4] Kato, H., Billinghurst, M., Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System, Proc. of 2nd Int. Workshop on Augmented Reality, pp.85-94 (1999). Proceedings of 2nd Int. Workshop on Augmented Reality. 1999. pp. 85-94.

[5] Mapes, D., Moshell, J., A Two-Handed Interface for Object Manipulation in Virtual Environments. Presence: Teleoperators and Virtual Environments, 1995. 4(4): pp. 403-416.

[6] Milgram, P., Takemura, H., Utsumi, A., Kishino, F., Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum. Proceedings of Telemanipulator and Telepresence Technologies. 1994. SPIE.

[7] Poupyrev, I., Berry, R., Kurumisawa, J., Nakao, K., Billinghurst, M., et al., Augmented Groove:

Collaborative Jamming in Augmented Reality. Proceedings of SIGGRAPH'2000 CA and A. 2000. ACM. pp. 77.

[8] Rekimoto, J., Ayatsuka, Y., Hayashi, K., Augmentable reality: Situated communication through physical and digital spaces. Proceedings of ISWC'98. 1998. IEEE.

[9] Schmalstieg, D., Fuhrmann, A., Hesina, G., Bridging multiple user interface dimensions with augmented reality systems. Proc. of ISAR'2000. 2000. IEEE. pp. 20-29.

[10] Ullmer, B., Ishii, H., The metaDesk: Models and Prototypes for Tangible User Interfaces. Proceedings of UIST'97. 1997. ACM. pp. 223-232.

A video clip is available at
http://www.cs.cmu.edu/~desney/Tiles/

Printing Virtual Reality Interfaces - The Cognitive Map Probe

Ehud Sharlin,[1] Benjamin Watson,[2] Lili Liu,[1] Steve Sutphen,[1]
Robert Lederer,[1] Pablo Figueroa,[1] and John Frazer[3]

[1] University of Alberta; [2] Northwestern University;
[3] The Hong Kong Polytechnic University

Abstract

The Cognitive Map Probe (CMP) is a novel Virtual Reality interface that attempts to assess the cognitive mapping abilities of its users. The CMP uses a tangible user interface (TUI) in order to support natural acquisition and straightforward assessment of cognitive maps. It is directed at the assessment of early Alzheimer Disease (AD) by measuring the decline in cognitive mapping abilities, a decline associated with early phases of AD. The CMP uses an adaptation of a pioneering "Machine-Readable Model", the Segal Model, which enables the user to interact with a virtual neighborhood environment by manipulating highly realistic, highly detailed, physical 3D models. All the CMP's TUI subparts were designed as realistic 3D small-scale models of physical landmarks and later printed using a 3D printer. This affords a very simple mapping between the virtual and physical elements of the CMP interface. In this short paper we briefly describe the CMP project's fundamentals and the concept of literally printing 3D VR interfaces.

1. Cognitive Mapping and early AD

Cognitive Maps can be defined as: an overall mental image or representation of the space and layout of a ***setting***. Cognitive mapping can be defined as: ***the mental structuring process leading to the creation of a cognitive map*** [2].

The most widely accepted model for cognitive mapping is the Landmark-Route-Survey (LRS) model [4,8]. The highest level of cognitive mapping ability – survey knowledge - is the ability to integrate landmark and route knowledge of an environment into a detailed geometrical representation in a fixed and relatively precise global coordinate system (e.g., the ability to draw a detailed map).

Although different manners of interaction with an environment will lead to different levels of knowledge and might result in different cognitive maps [2], both physical and virtual environments are valid means of acquiring cognitive maps as both are external to the learner [8]. Cognitive mapping using VR is an active research domain [5], with great attention given to the question of knowledge transfer, i.e. *was the cognitive map acquired in the virtual environment useful in the physical world?* Currently there is no clear-cut answer to this question [5,9]. Another open question is the level of immersion actually needed for effective cognitive mapping [11].

Cognitive maps can be probed using several techniques, e.g. verbal, bearing and distance, map-based and functional techniques [5,9]. Related to our efforts is the map placement technique in which the user is asked to point to objects' position on a grid, or to place objects' representation tangibly [3,9,11]. Very few attempts have been made to semi-automate the probing of cognitive maps. Baird et al. [3] displayed a 13x13 grid for computerized map placement. Later, direct computerized bearing input was implemented in various efforts [4,12].

Assessment of the high-levels of cognitive mapping abilities, i.e. survey knowledge, is expected to achieve high discrimination between early AD patients and healthy elderly persons [10]. Early assessment of AD is extremely important since these phases of the disease have major implications on the person's ability to perform everyday activities that were previously well within her capabilities.

2. TUIs and the Segal Model

Tangible user interfaces can be defined as: interface devices that use physical objects as means of inputting *shape, space and structure into the virtual domai*n. Several research groups are active in the field (see [13]). Pioneering work in this field was performed by Frazer and his group [6,7] and by Aish [1] more than 20 years ago. Generally, good TUIs will offer the user good affordances, unification of input and output, and support for "false starts" and "dead ends" in task execution.

The Segal model was built by Frazer and his group to enable users to interact with a floor plan both tangibly and virtually [6,7]. The model is a

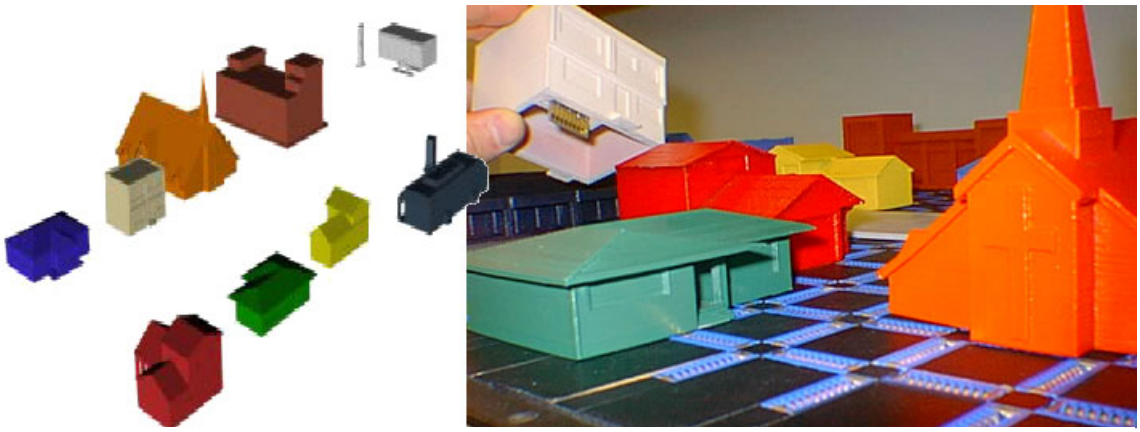


Figure 1: Virtual (left) and physical (right) overviews of the CMP

large board with an array of edge connector slots enabling the connection of numerous objects (each carrying a unique diode-based code) while tracing their location and identification in real-time (see figure 1). Recently, the Segal model was modernized so it can connect to a PC through a standard parallel port, using a Linux driver to scan the board and a Half-life® computer-game-engine to perform the rendering [14].

3. The CMP

The CMP is designed to enable automatic assessment of early AD by attempting to probe the more advanced cognitive mapping abilities, (survey knowledge). The CMP consists of the Segal model as the input device and a large display screen for output. The CMP assessment process begins by familiarizing the subject with a new environment, resembling a typical neighborhood, by enabling exploration of a virtual representation of the environment. The CMP then queries the subject's cognitive map by asking her to reconstruct the virtual environment, or parts of it, using realistic small-scale models of the environment's landmarks as interfaces, placing them on top of the Segal model.

The tangible interaction is supported by a set of realistic small-scale models of unique landmarks, such as residential houses, a church, a grocery store, gasoline station and a fire department (see figure 1). All the models were designed in high detail using 3D-CAD tools and later printed at a consistent scale using a 3D printer. A unique diode ID was manually inserted to a socket printed in each model. While the user manipulates the small-scale physical models, the CMP detects each model's ID and location and renders the model's virtual counterpart accordingly.

We believe that tangible instancing of virtual objects with 3D printers, and use of those instances in VR interfaces, is a worthy topic for future research.

While the CMP hardware is mostly done, the work on the assessment software is ongoing and preliminary user evaluations are expected by mid 2001.

4. Acknowledgments

The authors thank Spencer Tong for the Rhino support and Dr. Jonathan Schaeffer for his ongoing support and advice.

We would like to thank and acknowledge the contribution of John Frazer's team members who created the original Machine Readable Models beginning in 1979. Team members included: Julia Frazer, Peter Frazer, Walter Segal, John Potter, Steven Brown and David McMahon. Funding for the project was from Autographics Software and the University of Ulster.

Our work was made possible through access to the infrastructure and resources generously funded by MACI, Multi-media Advanced Computational Infrastructure for Alberta.

5. Reference

[1] Aish R., "3D input for CAAD systems", Computer-Aided Design, 11 (2):66-70, Mar. 1979.

[2] Arthur P. and Passini R., *Wayfinding: People, Signs, and Architectur*e, Toronto: McGraw-Hill Ryerson 1992.

[3] Baird J. C., "Studies of the Cognitive Representation of Spatial Relations", Journal of Experimental Psychology: General, 108 (1): 90-91, 1979.

[4] Colle H. A. and Reid G. B., "The Room Effect: Metric Spatial Knowledge of Local and Separated Regions", Presence: Teleoperators and

Virtual Environments, 7 (2): 116-129, April 1998.

[5] Darken R. P. and Allard T., "Spatial Orientation and Wayfinding in Large-Scale Virtual Spaces II", Presence: Teleoperators and Virtual Environments, 8 (6): iii-vii, Dec. 1999.

[6] Frazer J. H., "Use of Simplified Three – Dimensional Computer Input Devices to Encourage Public Participation in Design", Computer Aided Design 82, Conference Proceedings, Butterworth Scientific, 143-151, 1982.

[7] Frazer J. H., An Evolutionary Architecture, Architectural Association 1995.

[8] Golledge R. G., "Cognition of Physical and Built Environments." In T. Garling and G. W. Evans (eds.), *Environment, Cognition and Action: An Integrated Approac*h, New York: Oxford UP 1991.

[9] Howard J. H. JR. and Kerst S. M., "Memory and Perception of Cartographic Information for Familiar and Unfamiliar Environments", Human Factors, 23 (4): 495-504, 1981.

[10] Liu L., Gauthier L. and Gauthier S., "Spatial Disorientation in Persons with Early Senile Dementia of the Alzheimer Type", The American Journal of Occupational Therapy, 45 (1): 67-74, Jan. 1991.

[11] Patrick E. and Cosgrove D. et al, "Using a Large Projection Screen as an Alternative to Head-Mounted Displays for Virtual Environments", CHI 2000, 478-485, April 2000.

[12] Ruddle R. A. and Payne S. J. et al, "Navigating Large-Scale 'Desk-Top' Virtual Buildings: Effects of Orientation Aids and Familiarity", Presence: Teleoperators and Virtual Environments, 7 (2): 179-193, April 1998.

[13] Ullmer B. and Ishii H., "Emerging Frameworks for Tangible User Interfaces", IBM System Journal, 39(3-4): 915-931, 2000.

[14] Sutphen S., Sharlin E., Watson B. and Frazer J., "Reviving a Tangible Interface Affording 3D Spatial Interaction", WCGS 2000 (Western Computer Graphics Symposium), Panorama, British Columbia, Canada, March 2000.

Multi-Modal Natural Interface between Human and Virtual World using Gesture and Brain EEG Signals

Adrian David Cheok and Kuntal Sengupta

Department of Electrical and Computer Engineering,
The National University of Singapore, Singapore 117576

Introduction

Recently there has been an increased effort to make the interaction between computers and humans as human-like as possible [1]. Furthermore, in virtual worlds it has become more and more important to introduce multi-modal and natural interaction between the human and the computer, in order that the human can feel totally immersed in the virtual world without having to think about the control or the interface with that world. In similarity to the real world, the virtual world should be able to respond to human gestures. Furthermore, the virtual world allows the human to interact with a world in ways not possible in the real world, or in other words to go beyond realism.

Hence in this paper we address multi-modal interfaces with virtual worlds in two areas. Firstly, natural real world interaction is achieved by recognizing the humans gesture and movements, and secondly a completely new natural interaction is given by recognizing a person's color thoughts with EEG brain wave recognition. This allows the human to communicate and interact in a natural manner and in addition provides addition virtual senses that go beyond realism. Note that the color thought interface is explored in this paper in order to determine if simple thoughts can be deducted from brain signals. The main purpose is to provide an impetus for future research into other possible interactions between the human and virtual world using thoughts.

Multi-Modal Interaction

Recently in the field of virtual reality, much research has focused on investigating different types of human-virtual world interfaces using multiple sources of human information, in order that the virtual world can interact in a more natural and human-like manner. This area of research is termed multi-modal interfacing, and is characterized by combining various human signals and features in order for the computer to understand and interact with humans in a better manner. For example, various research developments have been made in computer based hand gesture recognition, gaze direction and facial expression recognition, and audio-visual speech recognition.

While gesture and gaze features remain very important modalities of virtual world interfaces, in this paper we will also consider human EEG (electroencephalogram) brain signals as an additional modality. The purpose of this research is to show that it is possible to recognize color thoughts solely from EEG signals and use this, together with gesture recognition, to interface, communicate, and modify the virtual world. The great advantage of such recognition is that the both the gesture and EEG signal are completely natural human signals that require no instructions or extra thinking on the user part, leaving the user to be totally immersed in the virtual world.

In this paper an experimental prototype of the research is described. The research prototype has been developed to achieve the following: Using a head mounted display, computer, and an EEG machine, a user may interact with a virtual environment that contains a virtual vase. The vase orientation is naturally controlled by the hand gestures, and the vase's color is controlled by the user thinking of a color. Currently the color that can be chosen is limited to red, blue, and green. The theoretical developments as well as the experimental setup will be described in the sections below.

Gesture Recognition

Before any gesture recognition is done, motion detection is first done to obtain the necessary data. Next, the extraction and processing of useful information is handled using Motion Energy Image (MEI). After which, interpretation of the MEI results are calculated using moments.

One of the simplest approach for motion detection using spatial techniques is to use two images frames $f(x,y,t_i)$ and $f(x,y,t_j)$ taken at times t_i and t_j, respectively, and compare the two

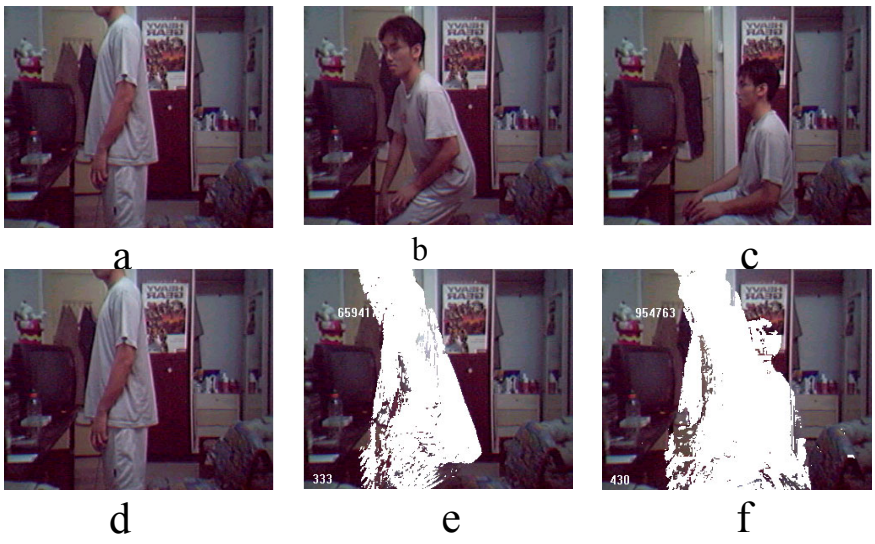


Figure 1 : (a-c) the different positions before the person sits down at frame 0 , 5 and 15 (d-f) the MEIs at frame 0 , 5 and 15. Note that frame 15 is a summation of frames 0 to 14.

images pixel by pixel. One procedure for doing this is to form a difference image. Suppose that we have an image containing a stationary object and it's background. Comparing this image against a subsequent image having the same environment and components but having the object moving results in the difference of the two images canceling out the environment and highlighting the moving object.

A difference image pixel between images taken at times t_i and t_j may be defined as

$$d_{ij}(x,y) = \begin{cases} 1 & if \quad |f(x,y,t_i) - f(x,y,t_j)| > \theta \\ 0 & otherwise \end{cases} \tag{1}$$

where θ is a threshold and x and y are the pixel's coordinates.

MEIs are actually the unification of the difference images (binary motion images) which are formed from the motion detection engine over time. Therefore, using equation (1) with $d_{ij}(x,y)$ at time t being represented as $d(x,y,t)$, the binary MEI $E_\tau(x,y,t)$ is defined as :

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} \mathrm{d}(x,y,t-i) \tag{2}$$

where τ is the time duration used in defining the temporal extent of the action. An example of a MEI is shown below in Figure 1 of a person sitting in a chair.

To make the implementation simpler, the time duration for the MEI was substituted in terms of frames. The gesture recognition algorithm was initially first trained by detecting the 3 gestures over an interval of 15 frames each. The 3 gestures were then randomly shown 10 times each and the results recorded down. Another 2 more sets (B and C) were then included, each with another 3 different random gestures (all in all, a total of 90 instances were tested). To calculate the moments of the MEI, the 3rd and 5th order invariant moments were used. The results using MEI was then shown in Table 1.

	Success	False	Undetected
Set A (amount)	25	2	8
Set A (percentage)	83.3%	6.7%	11.7%
Set B (amount)	24	2	9
Set B (percentage)	80.0%	6.7%	30%
Set C (amount)	23	5	3
Set C (percentage)	76.7%	16.6%	10%

Table 1 : Results using MEI

The results obtained from all 3 sets averaged at 80% success, which is quite acceptable.

EEG signals and color recognition

EEG signals are a recordings of the spatio-temporal averages of the synchronous electrical activity of radially orientated neurons in the cerebral cortex. This activity is of very low amplitude, typically in the range of 10 to 100 μV, and with a low bandwidth of 0.5Hz to 40Hz [2]. The theoretical impetus for using brain EEG signals for color recognition is that studies that have shown that during focused human thoughts on color various effects can be seen in the brain waveforms[3].

Hence, in this research, an objective was to demonstrate that it is possible to recognize color thoughts solely from EEG signals in order to provide another modality to interact with a virtual world in a natural manner simply by thinking of a color.

In order to achieve this, a system was developed to identify and classify human thought patterns of color. The system used wavelet transform feature extraction and a fuzzy inference system that was trained using subtractive clustering and neural network based rule optimization. In the section below, the details of the development of the wavelet feature extraction and fuzzy rule processing system used for color recognition based on EEG signals will be briefly detailed.

A. Wavelet feature extraction and fuzzy rule processing system for EEG based color recognition

The first part of the EEG based speech recognizer system is a wavelet transform algorithm [4]. The reason that the wavelet transform is used is because this transform is highly suitable for the analysis of EEG signals, and has been shown to be effective in the extraction of ERP (Event Related Potentials) [5]. The wavelet transform can be interpreted as a decomposition of a time domain signal into time-scale domain signals where each component is orthogonal (uncorrelated) to each other. Hence, unlike the Fourier transform, which is global and provides a description of the overall regularity of signals, the wavelet transform looks for the spatial distribution of singularities.

As will be detailed in the experimental results section below, the first step in producing an EEG color recognizer is to extract the wavelet features from measured data when the human is thinking of a certain color. Then a classifier can be trained based on this data. In this research a fuzzy classifier was used, and the preprocessed data of EEG signals were subjected to fuzzy subtractive clustering [5] which extracts fuzzy classification rules for pattern recognition. Our motivation for using fuzzy subtractive clustering, as opposed to other methods, is the fact that fuzzy rules are easy to verify due to their heuristic nature.

In order to train the fuzzy system we first separate the training data into groups according to their respective class labels (where the class labels are the colors that a person was thinking at the time of the signal). Subtractive clustering is then applied to each group of data individually to extract the rules for identifying each class of color data. The subtractive clustering algorithm can be summarized as follows:

1. Consider each data point as a potential cluster center and define a measure of the potential of that data point to serve as a cluster center.
2. Select the data point with the highest potential to be the first cluster center.
3. Remove all data points in the vicinity of the first cluster center, in order to determine the next data cluster and its center location.
4. Iterate on this process until all of the data is within radii of a cluster center (select the data point with the highest remaining potential as the second cluster center).

Each cluster found is directly translated into a fuzzy rule and initial rule parameters. The individual sets of rules are then combined to form the rule base of the classifier. When performing classification, the consequent of the rule with the highest degree of fulfillment is selected to be the output class of the classifier.

After the initial rule base of the classifier has been obtained by subtractive clustering, we use an artificial neural network to tune the individual parameters in the membership functions to minimize a classification error measure. This network applies a combination of the least-squares method and the back-propagation gradient descent method for training the fuzzy membership function parameters to emulate the training data set.

EEG Data Training and Recognition

EEG data were recorded from student subjects sitting on a chair with a head-mounted display that projected a single color. During the experiment, the subject was asked to watch the virtual screen, which flashed colored panels of red, green, and blue chronologically for 10

seconds each. The subject was asked to think of the color projected on the display.

The EEG data were acquired from a computer combined with a 16 channel BioSemi ActiveOne system. The data were recorded from 16 locations over the surface of the head, corresponding to the points defined by the international 10-20 electrode system. All data were collected at 2048 Hz at 16 bit resolution.

After data collection the training of the EEG based color recognition algorithm discussed above occurs as follows:

1. Preprocess the signals: this comprises segmentation of data as well as the determination of the wavelet coefficients through decomposition by the discrete wavelet transform. It was found after some experimentation, that the most useful wavelet coefficients were from the 5^{th} scale. These approximately correspond to a band of frequencies from 4 to 10Hz. The detailed wavelet coefficients at this scale consists of 13 points, which would form the reduced feature vector to be input into the classification system.
2. The features thus extracted from the preprocessing operation are subject to subtractive clustering which extracts fuzzy classification rules from them.
3. Using an adaptive neuro-fuzzy inference system (ANFIS), the rule parameters are subsequently optimized.
4. The above happens for every channel and all the outputs are streamlined into an integrator with performs a weighted average of the inputs.

A. Results of EEG classification

The color recognition system was tested with 34 trials. Data from each channel were processed and classified individually. An output integrator, which calculated the weighted average of the inputs, produced the final classification results. The classification results are shown in Table 2.

The results of the classification show that it is possible to classify about 85% red segments, 70% green segments and 92% of the blue segments correctly. Although the results are rather mediocre, it is believed improvements can be made on the recognition results in the future by tuning the algorithm and having more training data and subjects.

Prototype Results

As described above, a gesture and EEG color recognizer was developed in order to provide a natural and seamless interface with the virtual world. To provide a demonstration we developed system which contained a Sony Glasstron head mounted display, computer, and BioSemi EEG machine. The application we developed allowed a user to interact with a virtual environment that contains a virtual vase.

EEG Classification	Red	Green	Blue	Un-known
Signal Red	11/13	1/13	0/13	1/13
Signal Green	1/10	7/10	2/10	0/10
Sinnal Blue	0/12	1/12	11/12	0/12

Table 2: The table of the classification results on the testing data set

The vase orientation is naturally controlled by the gesture, and the vase's color is controlled by the user thinking of a color Note that if a color was not recognized a default texture was displayed on the vase surface.

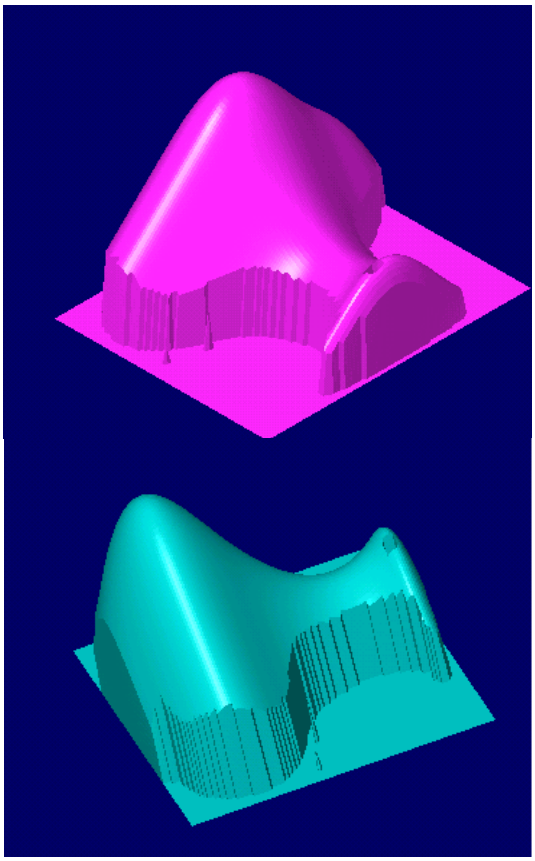


Figure 2: The rendering of the vase with different color.

Conclusion

In this paper we presented a multi-modal interface for virtual worlds in two areas. Firstly, natural real world interaction was achieved by recognizing the humans gesture and movements, and secondly a new natural interaction was given by recognizing a person's color thoughts with EEG brain wave recognition. This allowed the human to communicate and interact in a natural. The color thought interface is explored in this paper in order to show that simple thoughts can be recognised from brain signals. We hope that this paper provide an impetus for future research into other possible interactions between the human and virtual world using thoughts.

References

[1] R. Sharma, V.I. Pavlovic, and T.S.Huang, Toward multimodal human-computer interface, In Proceedings of the IEEE, 86(5): 853-869, May 1998.

[2] Binnie, C.D., Recording the brain's electrical activity, Electrical Engineering and Epilepsy: A Successful Partnership (Ref. No. 1998/444), IEE Colloquium on, 1998 Page(s): 3/1 -3/3

[3] Michael Scherg, FOCUS source imaging: New perspectives on digital EEG review, Division of Biomagnetism, Clinic of Neurology, University of Heidelberg

[4] Zhong Zhang; Kawabata, H.; Zhi-Qiang Liu, EEG analysis using fast wavelet transform, Systems, Man, and Cybernetics, 2000 IEEE International Conference on , Volume: 4 , 2000, Page(s): 2959 –2964

[5] Herrera, R.E.; Sclabassi, R.J.; Mingui Sun; Dahl, R.E.; Ryan, N.D., Single trial visual event-related potential EEG analysis using the wavelet transform, BMES/EMBS Conference, 1999. Proceedings of the First Joint , Volume: 2 , 1999, Page(s): 947 vol.2

[6] Amir B. Geva, Dan H. Kerem, Brain State Identification and Forecasting of Acute Pathology Using Unsupervised Fuzzy Clustering of EEG Temporal Patterns, Fuzzy and Neuro-fuzzy Systems in Medicine, CRC Press LLC

Gesture-Driven Control of Spaces and Objects in Collaborative Augmented Reality

Yaser Yacoob and Larry Davis

Computer Vision Laboratory
University of Maryland
College Park, MD 20742, USA

Abstract

A multi-modal system integrating computer vision and speech recognition to enable collaboration through interaction with virtual spaces/objects by natural gestures and speech is being developed. Our re-search focuses on detection, tracking, recognition and visual feedback of human hand and finger movements in a cooperative user environment, and the integration of gesture and speech recognition for man/machine communication. Computer vision algorithms are employed to measure and interpret hand/finger movement of the users.

Background

Our research focuses on situations where two or more people employ gestures/hand actions to express intentions with respect to a shared virtual environment. The environment is rendered to the user through stereoscopic head-mounted-displays (Sony LDI-D100). Video cameras mounted on the users HMDs are employed to recognize the hand/finger gestures in the context of speech and convey them to the participants through appropriate modifications of the virtual world. Human gestures, in conjunction with speech, can be categorized into

1. Identification gestures, like pointing, that identify locations/objects in space. So, a person might

- place an object,
- identify goals for an object's movement,
- indicate that a collection of objects be treated as a group,
- segment groups into subgroups, or
- change an objects internal state.

So, for example, a person might first use speech to indicate the type of identification action to be performed (e.g., group), and then use a hand gesture to control the application of the action (e.g., outline the set of elements to be grouped).

2. Action gestures that specify the movement of an object or a group so that a person might

- Specify a translation or a rotation of an object, etc.
- Control a virtual tool to “reshape” the virtual world he inhabits, or
- Directly apply forces that deform and alter the shape of an object

Our research builds on our prior work for detection of people and their body parts from color video [2], motion estimation of rigid, articulated and deformable motions [1,3,4] and recognition of facial and body activities [1,3].

Vision

Our vision for future interfaces for VR centers around the following:

- We recognize the centrality of a multi-modal interface (e.g., vision and speech) to normal collaboration between users. Accordingly, integration of speech/visual information is needed for achieving human-computer and human-human interaction in VR.
- Employ vision techniques instead of wearable sensors to avoid imposing contraints on user's use of hands and fingers. Specifically, we design representations for gestures, learning algorithms that can recover these representations from examples and real time computer vision algorithms that can

recognize these gestures from (possibly multi-perspective) video.

- Employ a closed-loop visual feedback framework to enhance the immersion of the user. The is realistically provided by texture mapping the virtual scene with user's hands achieving a blending of real/virtual scenes.

Status

We are developing the necessary infrastructure that supports conducting our research,

We employ speech recognition software (IBM-Via Voice) and Java applications to support text-to-speech and speech-to-text processing. We designed grammars that allow understanding the user's spoken sentences and executing them. We explored the use of a high level software for designing and conducting verbal human computer interactions.

We developed stereo modeling and rendering software that supports rendering graphic scenes on the Sony HMD. It allows the user to tune several parameters (e.g., parallax, eye distance) to control the viewing of the 3D scene. The software employs Open GL to create the scene and texture map image regions (such as user hands) taken from the video camera onto synthetic objects in real-time and render the scene from stereo vantage points.

We developed tracking algorithms to run in real-time on a PC. These tracking algorithms handle rigid, articulated and deformable motions of hands and fingers. This real-time performance is critical since feedback from the system reinforces and corrects gestures that are performed by the user.

We developed skin detection algorithms for hand detection and initialization of hand and finger regions for the tracking algorithm.

We are developing an algorithm for hand depth estimation to support accurate rendering of the user's hand within the geometry of the synthetic scene.

References

[1] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. IJCV, 25(1), 1997, 23-48.

[2] I. Haritaoglu, D. Harwood, and L. Davis, W4S: A real time system for detecting and tracking people in 2.5D. ECCV, 1998, 877-892.

[3] S. X. Ju, M. Black,and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. in Proc. Int. Conference on Face and Gesture, Vermont, 1996, 561-567.

[4] Y. Yacoob and L. Davis, Y. Yacoob and L.S. Davis, Learned Models for Estimation of Rigid and Articulated Human Motion from Stationary or Moving Camera, Int. Journal of Computer Vision, 36(1), 2000, 5-30.

Conversational User Interfaces and Human Aspects

Chair: Léonie Schäfer, GMD-FIT

Immersive Multimodal 3D Interfaces for Music, Stories and Games

Stephen Barrass

CSIRO Mathematics and Information Sciences
GPO Box 664, Canberra ACT, Australia 2602
stephen.barrass@cmis.csiro.au

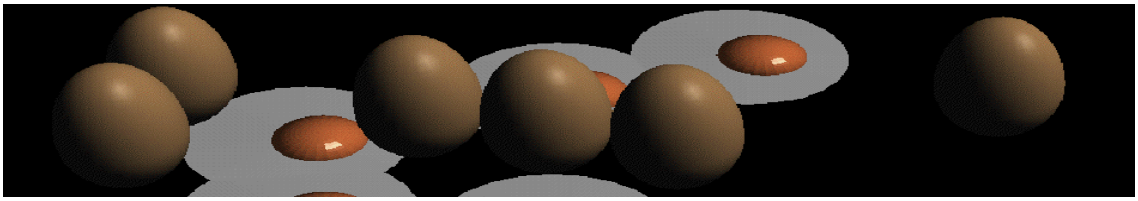
VE interfaces often include conventional 2d desktop widgets such as menus and buttons for selecting files, setting parameters and carrying out other functions. However the large scale, relatively low resolution and six degrees of manual freedom make it hard to use small graphic widgets that are difficult to find and orient, have lots of text and require precise 2d movements or alphanumeric keyboard inputs. We need to develop alternative interfaces based on more physical interactions with the properties, locations and materials of 3d objects. During a recent 18 month post-doc in the Virtual Environments group at the German National Research Institute for Information Technology in Bonn I experimented with VR interfaces with one aim being to forego any use of conventional widgets. I would be pleased to describe these experiments and my observations with them, and to learn from the experiences of others who have been working with VE interfaces.

Weichei.

Weichei is childrens game designed for the Cyberstage Virtual Reality room was shown at the "Virtuelle Welten Erleben" (Discover Virtual Worlds) Art exhibition at the Animax Multimedia Theatre in Bonn <www.animax.de>. When you walk into the room all you see are a dozen football-sized eggs and a giant silver spoon. Each egg hums a little tune and together they produce an a-capella chorus. When you pick up the spoon it whooshes through the air like a sword or light-sabre. You can pick up an egg just like you would with an ordinary spoon, without pressing a button on a stylus. If you tilt the spoon the egg will fall on the floor and break and you can really make a mess ! As the egg falls it whistles like objects in cartoons do, and when it lands the floor shakes. Broken eggs sizzle and fry and some eggs bounce. The minimal graphics is in contrast to most virtual environments which are primarily visual. The sounds give physicality to the virtual spoon, and character to the featureless eggs. Physical lifting and tilting replaces the usual button-based interfaces to computer games. The finite length of the spoon means you can pick up eggs that are behind other eggs, something that is impossible to do with an ray-based stylus that picks the object it first intersects. You can also use the spoon to catch eggs that bounce. However this activity highlighted the difficulties of real-time physical simulation because the egg sometimes falls through the spoon, especially if it is falling fast, due to lack of temporal resolution for determining intersections.

Op-Shop

Op-Shop is a virtual room cluttered from floor to ceiling with glasses, vases, bottles, figurines, carpets, and other bric-a-brac. In the middle is a kitschy art-deco table with a champagne glass. You can shatter this glass by singing a high note, just like opera singers in the movies sometimes do. You will be pleased to know that you can likewise shatter the table - by singing a low note. Op-shop is the perfect place to practice to become an opera star ! Each object responds to acoustic energy based on its size and shape.



Larger objects resonate at lower frequencies and more complicated objects are more fragile. As an object accumulates energy it begins to vibrate and feedback a ringing tone. A large urn requires the concentrated attentions of several divas singing together to rupture it. Because the interaction does not require a special tool, divas can join in or leave the choir at any time, and collaborate on equal footing. OpShop demonstrates the notion of using non-verbal auditory gestures to interact with objects that are out of arms reach in a human-sized virtual environment. Perhaps the most significant observation of people in OpShop is their reluctance to sing in public. I often have to do all the hard work, unless there is someone who can really sing and has an outgoing personality, in which case it can be a lot of fun.

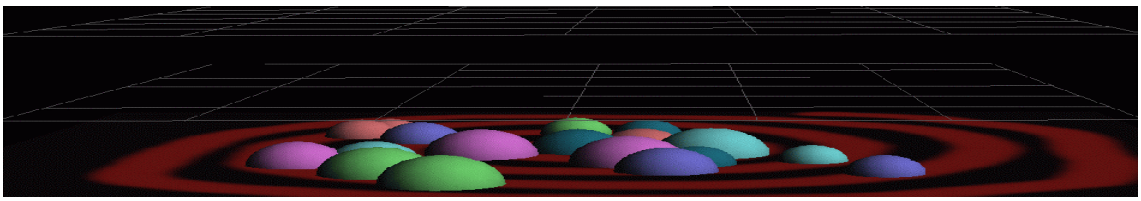
Bounce-Machine

BounceMachine is an immersive drum machine for making rhythms and dancing in the same space. There are 20 coloured balls of different sizes on the floor, and a grid floating in the Cyberstage space. If you drop a ball it triggers a sound when it hits the floor. The colours of the balls indicate different percussive instruments - a blue conga, a green cowbell, a purple scratch, an orange snare etc. The size of the ball indicates its heaviness - heavy balls make louder sounds and cause the floor to vibrate when they hit. You can make rhythms by dropping different balls from different heights and letting them bounce. The grid floating in the space helps you position the balls at exact heights to synchronise rhythms. Although its fun it does not really work very well if you want to make tight rhythms. One problem was that it is really very difficult to drop 2 balls from exactly the same height, especially if one of the balls is bouncing. Secondly you don't hear the ball you drop until it hits the floor which means you have to predict the delay rather than respond to the current beat. Finally the temporal resolution of the physical simulation is still too low so that even if 2 balls are both dropped from exactly the same height they may trigger at noticeably different times.

Beethoven Salon

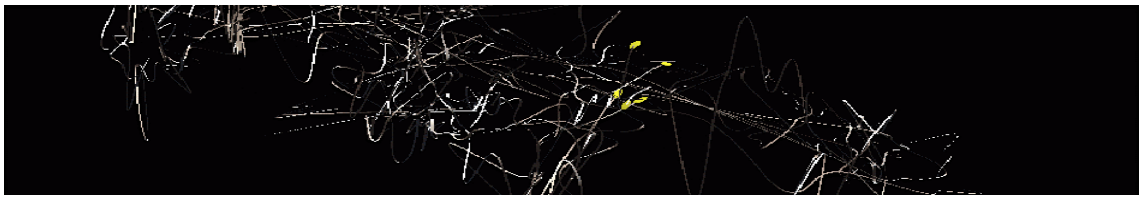
The Beethoven Salon is a room-sized Virtual Environment installation designed for the Beethoven Haus Museum in Bonn. The experience is centred on the Pastoral Symphony which has five movements inspired by Beethoven's feelings for nature. Although the symphony contains musical references to nature Beethoven expressly did not want the Pastoral to be taken literally and inscribed the first manuscript with the words "Mehr Ausdruck der Empfindung als Malerei"-"more an expression of feeling than a painting".

The Beethoven Salon is a model of the room where Beethoven composed the Pastoral symphony. This model acts as a portal to an abstract world of expression generated by the music. Furnishings and artefacts from the Beethoven Haus collection are seamlessly integrated with matching colours, lighting and materials in the virtual extension. Visible on the virtual half of the piano is the manuscript of the symphony inscribed with the famous phrase "more a feeling than a painting" in Beethoven's writing. As visitors enter the room a breeze rustles the curtains and turns the pages to the 4th movement - the thunderstorm. As the music begins the lights dim until only glowing silver music notation is visible on the page of the manuscript. Slowly the notes float up off the page leaving silver staves trailing behind like will-o-wisps. During the tranquil opening passages the notes intertwine slowly like a school of fish, leaving gentle symmetrical wakes. As tension builds the notes become agitated, moving faster, further and more abruptly, until the climactic thunder scatters them to all corners of the room. As the movement progresses the notes dive back and forth, coming together in



moments of calm between the thunderbursts, to be scattered again by the fury of the storm. In the quiet closing moments the notes return to their placid schooling pattern. By the end the room has been re-illuminated in a pattern of light and shadow cast by the music-sculpture left in the wake of the notes.

The notes are artificial-life (a-life) algorithms that model the behavior of flocks, schools, and herds of animals. The a-life are specially designed to react to the dynamic tension of Beethoven's music in an expressive way; These algorithms bring an abstract "nature" into the computer generated virtual environment. They respond to both the music and other sounds in the room. If someone laughs they flitter, and handclaps will startle and scatter them. This responsiveness blurs the line between the real and virtual so the notes seem alive and present, rather than mere computer generated projections in another space. The Beethoven Salon also has acoustic subwoofers under the floor so that the music can be felt through the feet, especially the thunderous sections. This acoustic floor provides the possibility for people who are deaf like Beethoven was to also experience his music.



Interaction for Virtual Environments based on Facial Expression Recognition

Ana Luisa Solís,[1] *Homero Ríos* [2]

[1] Facultad de Ciencias
Universidad Nacional Autónoma de México
Circuito Exterior, Ciudad Universitaria, Delg. Coyacán, México, D.F,
C.P. 04510 Mexico
ana@graf.fciencias.unam.mx

[2] Facultad de Física e Inteligencia Artificial, Universidad Veracruzana
Xalapa, México
hrios@mia.uv.mx

Abstract

This work describes a new interface for virtual environments based on recognition and synthesis of facial expressions. This interface senses the emotional state of the user, or his/her degree of attention, and communicates more naturally through face animation. We believe that such a system can be exploited in many applications such as natural and intelligent human-machine interfaces for vitual reality and virtual collaboration work

Introduction

This work describes a new interaction technique for virtual environments based on the integration of results from computer graphics and computer vision. In the last few years this integration has shown important results and applications [10;17]. For example, Richard Szeliski described the use of image mosaics for virtual environments in 1996 and in the following year for combining multiple images into a single panoramic image. H. Ohzu et al. described hyper-realistic communications for computer supported cooperative work.

Facial expression understanding is a good example of the rich middle ground between graphics and vision. Computer vision provides an excellent input device, particularly for the shapes and motions of complex changing shapes of faces when expressing emotions [16;17].

We have been studying how to analyze efficiently video sequences for capturing gestures and emotions. Relevant expressions and their interpretations may indeed vary depending upon the chosen type of application .

In this work, relevant expressions are assigned with a training system when it is important to know the user's interest on the information that is displayed or when she/he is interacting in a virtual environment.

Facial expression recognition is useful for adapting interactive feedback in a training system based on the user's level of interest. The type of expressions associated with these applications are: degree of interest, degree of doubt on the information presented, boredom, among other expressions, or assessing the time of interest or lack of interest presented by an application.

The work mentioned here strives to capture the high resolution motion and appearance of an individual face.

Analysis and interpretation of facial expressions

The communicative power of faces makes it a focus of attention during social interaction. Facial expressions and the related changes in facial patterns convey us the emotional state of people and help to regulate both social interactions and spoken conversation. To fully understand the subtleness and expressive power of the face, considering the complexity of the movements involved, one must study face perception and related information processing.

For this reason, face perception and face processing have become major topics of research

by cognitive scientists, sociologists and more recently by researchers in computer vision and computer graphics.

The automation of human face processing by computer will be a significant step towards developing an effective human-machine interface for virtual environments. We must consider the ways in which systems with this ability understand facial gestures (analysis), and the means of automating this interpretation and/or production (synthesis) to enhance human-computer interaction.

Facial Displays as a New Modality in Human-Computer Interaction

Facial expressions can be viewed in either of two ways. One regarding facial gestures as expressions of emotional states. The other view facial expressions related to communication. The term "facial displays" is equivalent to "facial expressions" but does not have connotation emotional.

The present paper assumes the second more general view. A face as an independent communication channel.

Figure 1: Facial Communication

Theory of Communicative Facial Displays

First of all, facial displays are primarily communicative. They are used to convey information to other people. The information that is conveyed may be emotional or any other kind of information; indications that the speaker is being understood, listener responses, etc.

Facial displays can function in an interaction as means of communication on their own. That is, they can send a message independently of other communicative behavior. Facial signs such as winks, facial shrugs, and listenner's comments (agreement or disagreement, disbelief or surprise) are typical examples. Facial displays can also work together with other communicative behaviour (both verbal and non verbal) to provide information.

Categorization of Facial displays used in the Virtual Environment

We must consider the facial displays we are interested in recognizing in the Virtual Environment. The kind of expressions are very specific:

- Thinking/Remembering. Eyebrow raising or lowering. Eyes closing. Pulling back one side of the mouth.
- Facial shrug/I don't know. Eyebrow flashes. Mouth corners pulled down. Mouth corners pulled back.
- Backchannel / Indication of attention. Eyebrow raising. Mouth corners turned down.
- Doubt. Eyebrow drawn to center.
- Understanding levels
- Confident. Eyebrows raising. Head nod.
- Moderately confident. Eyebrows raising.
- Not cofident. Eyebrows lowering.
- "Yes". Eyebrows raising.
- Evaluation of utterance.
- Agreement. Eyebrows raising.
- Request for more information. Eyebrow raising.
- Incredulity. Longer eyebrows raising.

Description of the system

The general purpose of the system is to be able to generate an agent understands the expressions of the real person and communicates using both verbal and non-verbal signals. This interface senses the emotional state of the user, or his/her degree of attention, and communicates more naturally through face animation.

Facial expression recognition is useful for adapting interactive feedback in the training system based on the user's level of interest.

The dialog is simulates in a 3D virtual scene which can be viewed from different remote sites over network.

Implementing such system needs solving many independent problems in many fields: image processing, audio processing, networking, artificial intelligence, virtual reality and 3D

animation. In designing such a system we divide it into modules, each module being a logical separate unit of the entire system. These modules are: synthesis and animation of expresions, facial expression recognition, audio

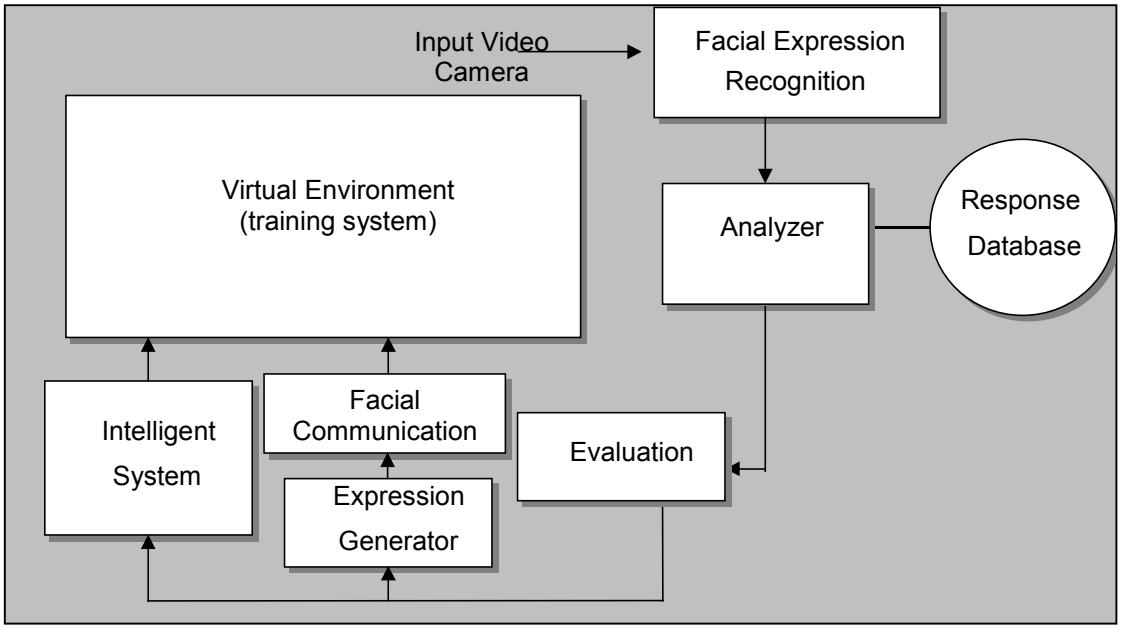


Figure 2: Global architecture of the system

processing, interpretation and response generation, and audiovisual synchronization.

Conclusion

In this paper we described a new interface for virtual environments based on the integration of results from computer graphics and computer vision.. This interface senses the emotional state of the user and help regulate the flow in the virtual environmet.

Acknowledgments

This work has been funded by Mexican National Council of Science and Technology (CONACYT) as project ref. C098-A and C100-A, “Gesture recognition interfaces and intelligent agents for virtual environments”.

References

[1] Bruce, V. and Green, P. (1989), *Visual Perception: Physiology, Psychology and Ecology*. Lawrence Erlbaum Associates, London.

[2] Campbell, L.W., Becker, D.A., Azarbayejani, A., Bobick, A., and Pentland, A. (1996). Invariant features for 3-D gesture recognition. *MIT Media Laboratory Perceptual Computing Section,* Technical Report no. 379.

[3] Cassell, J., Pelachaud, C. Badler, N. et . al (1994) Animated Conversation: Rule-based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. Computer Graphics Proceedings, pp.413-420.

[4] Chellapa, R., Wilson, C.L. and Sirohey, S. (1995). Human and machine recognition of faces: a survey. *Proceedings of the IEEE,* Vol. 83, No.5, pp. 705-740

[5] Cohen, M. and Massaro, D. (1993) Modeling Coarticulation in Synthetic Visual Speech,

[6] Models and Techniques in Computer Animation, Springer Verlag, pp. 139-156.

[7] Cootes, T.F., et al. (1992). Training models of shape from sets of examples. *Proceedings of the British Machine Vision Conference.*

[8] Cootes, T.F., Edwards, G.J. and Taylor, C.J. (1998). *Proceedings of the European Confe rence on Computer Vision,* Burkhardt, H. and Neumann, B. (Eds.)., Vol. 2, pp. 484-498, Springer-Verlag.

[9] Eisert, P. and Girod, B. (1998). Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics and Applications*, Vol. 18, No. 5, pp. 70-78.

[10] Ekman, P. and Friesen, W.V. (1975). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*, Englewood Cliffs, New Jersey; Prentice Hall, Inc.

[11] Ekman, P. and Friesen, W.V. (1978), *Manual for the Facial Action Coding System*. Consulting Psychologists Press, Inc. Palo Alto, CA.

[12] Lam, K.M. and Yang, H. (1996). ``Locating and extracting the eye in human face images''.

Pattern Recognition, Vol. 29, No. 5, pp. 771-779.

[14] Ohzu, H. and Habara, K. (1996). Behind the scenes of virtual reality: vision and motion. *Proceedings of the IEEE*, Vol. 84, No. 5, pp. 782-798.

[15] Oliver, N., Pentland, A. and Berard, F. (1997). Lafter: lips and face real time tracker. *MIT, Media Laboratory Perceptual Computing Section,* Technical report no. 396.

[16] Parke, F.I. and Waters, K. (1996). *Computer Facial Animation*. A K Peters.

[17] Pentland, A.P. (1996). Smart rooms. *Scientific American*. April 1996, pp. 54-62.

[18] Rios, H.V. and Peña, J. (1998). Computer Vision interaction for Virtual Reality, In *Progress in Artificial Intelligence*, Helder Coelho (Ed.), *Lecture Notes in Artificial Intelligence*, *Subseries of Lecture Notes in Computer Science*, No. 1484, pp. 262-273. Springer-Verlag.

[19] Russel, J. (1994). Is There Universal Recognition of Emotion From Facial Expression?. *A Review of Cross-Cultural Studies Psychological Bulletin* 115(1) 102-141.

[20] Schwartz, E.I. (1995). A face of one´s own. *Discover* the world of Science. Vol. 16, No. 2, pp. 78-87.

[21] Starner, T., Weaver, J., Pentland, A. (1998). Real-time American sign language recognition using desk and wearable computer based video. *MIT, Media Laboratory Perceptual Computing Section,* Technical report no. 466.

[22] Sucar, L.E. and Gillies, D.F. (1994). Probabilistic reasoning in high-level vision. *Image and Vision Computing,* Vol. 12, No. 1, pp.42-60.

[23] Terzopoulos, D. And Szeliski, R. (1992). Tracking with Kalman Snakes. In *Active Vision,* Blake, A. and Yuille, A. (Eds.), MIT Press.

[24] Waters, K. (1987). A Muscle Model for Animating Three Dimensional Facial Expression. *Computer Graphics Proceedings* Vol. 21, No 4 , pp.17-23.

[25] Waters, K. and Levergood, T.(1994) An Automatic Lip-Synchronization Algorithm for Synthetic Faces. *Proceedings of Multimedia 94, ACM*, pp149-156.

ReLIVE : Towards a communicative, supportive and 'enjoyable' Virtual Environment

Ben Salem

Polywork
12 Saint Mary's Lane, Sheffield,
S35 9YE, England
electronote@mail.com

Introduction

We have launched the ReLIVE project with the aim of delivering a communicative, supportive and 'enjoyable' virtual environment (VE). This is a multidisciplinary project involving the design of avatars, the development of a virtual environment and the definition of prevailing interaction principles to be involved. The design of the avatars that will be populating our VE focuses on delivering expressive avatars capable of emotion and personality rendering. While for the virtual environment we are looking into visual and audio features as well as the development of choreography to control the behaviour of the environment's avatars. We are at the same time creating a narrative as a social etiquette which will be the 'rule of the land' in the environment.

As part of the ReLIVE project, we are undergoing the development of an interface device to be used as a gateway to the system, a virtual environment rich with artefacts and a population of avatars, and finally a pet robot, the projection into the real world of the VE.

Although the two other parts of the project are important we will focus in this paper on the development of the virtual environment and the population of avatars.

Initial principles

The avatars are the embodiments in the VE of both the participants who can then see each others as well as see the environment agents. Looking at current environments there is two trends in avatars design namely avatars with a humanoid shape at different level of realism and animal, abstract or free shape avatars. Since our objective is to facilitate human to human communication, we have chosen to concentrate on the development of humanoid avatars. Our design of the avatars is inspired from the masks of Commedia dell'Arte which are rich in emotions and will be easily understood by the environment participants. Essentially we are driven in the design of the avatars by two aims. Initially to create an avatar that can be perceived as the representation of other participants, and also to facilitate the communication between participants by means of emotional and expressive avatars. Choosing highly realistic avatars would have required the delivery of sophisticated behaviours and complex choreography (e.g. Set of movements to express some emotions). Further to this initial tests have indicated the high level of expectations users have when presented with a realistic avatar.

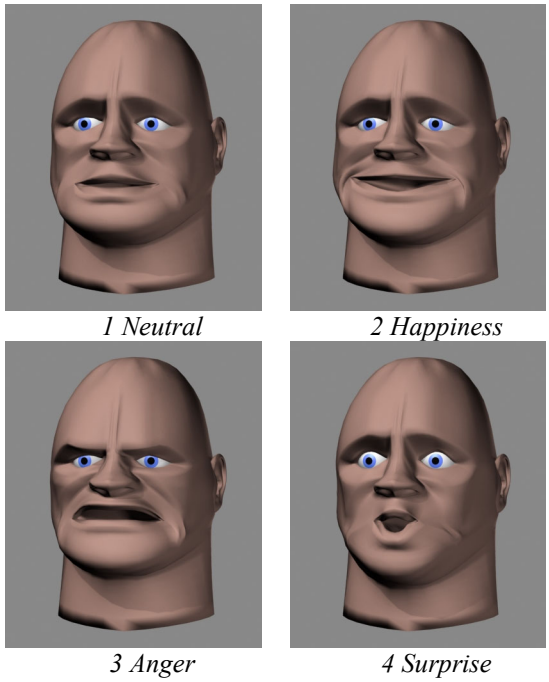


Figure 1: Realistic Avatars (courtesy of Y.Chapriot)

An example a face of a realistic avatar is shown in Figure (1) this face has strong features and can express emotions, however every effort to render

a realistic face have still left the avatar expressions ambiguous in some cases.

One can see that the neutral expression could be mistaken for sadness, and happiness for nonchalance. Our conclusion is that the rendering of emotions in a natural fashion will be very difficult to achieve and indeed impossible in the case of current online VEs (e.g. blaxxun, active world…).

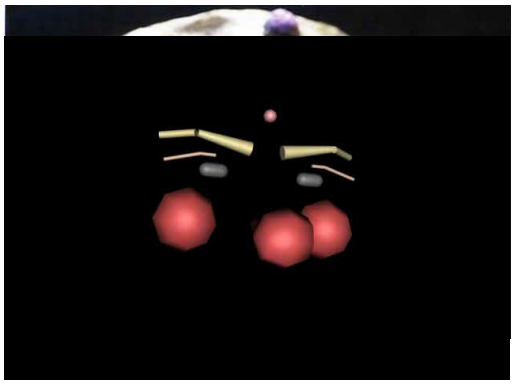
We thereafter explored other ways of expressing emotions, and we find inspiration in the world of theatre.

1 Original Mask (courtesy of J. Dixon)

2 Virtual Mask

Commedia dell'Arte

Commedia dell'Arte was of particular interest for several reasons. Essentially the masks worn by the actors, the comprehensive set of movements to express different emotions and state of mind, and most importantly the fact that Commedia is



essentially a improvisation theatre where the actors continuously add to and adapt the plot.

Figure 2: Commedia Mask: Il Dottore

Commedia's masks are interesting in many ways to our work as they provide caricaturised faces which can be rendered using simple elements as illustrated in Fig 2.1. The masks have also the particularity of reinforcing the most communicative features of the face, the overall shape, the shape and relative size of the nose, the forehead, the cheeks and the eyebrow. The mouth is often left uncovered.

Avatar's Behaviours

The behaviour of Commedia actors is highly structured and defined, the result is what looks like exaggerated movement of the characters on the stage. From our point of view this is very useful as it is a expression and emotion rich set of gestures and postures that can be used by the participants of the VEs. The choreography we are proposing for the avatars of ReLIVE is a set of hand gestures, facial expressions, body postures and also Commedia dell'Arte 'gestures', such a vocabulary will deliver a communication rich visual language.

Depending on the situation, for example casual meetings, work discussion or games, we have also outlined the synopsis of a short piece of theatre which we can translate into social etiquette in the VE. The synopsis is based on traditional Commedia characters, masks and storylines.

Such a narrative or' interactive theatre play' is in fact a scenario of the experience the participants will have in the environment. For example how the greeting of the participant(s) is achieved, how the encounter in the VE is done and what 'atmosphere' is set up for the 'venue'.

Figure 3: A Commedia dell'Arte posture(Attention)

The Virtual Environment Experience

A virtual environment is an interactive online three-dimensional visual space, which should not be about mimicking realism. It should be about delivering a communicative, supportive and 'enjoyable' environment. To do so we are relying on several elements to support the roles of the avatars, namely Dynamic interface objects, artefacts, atmosphere and the design of the environment.

Dynamic interface objects such as morphing 3D icons, widgets and teleports are used to facilitate the interaction within the VE, for example to assist during the navigation through the environment.

Figure 3: 3D Icon

A set of artefacts distributed in the environment, produced by a fine artist involved in this project. We are developing some 'furniture' to deliver specific functions (a kiosk for help) as well as some 'construction' to give special clues to the participants.

Atmosphere, audio and lighting effects controlled within the VE. We hope to have these parameters finely tuned to the narrative of the space.

Finally, the actual design of the environment is done following architecture and landscape principles, in terms of spaces, organisation and dimensions. Particularly to this point, Commedia suggest a set of 'décor' which we can inspire ourselves from when designing the VE. There is however a balance to find between design a theatre set and delivering a VE with a wide variety of applications.

Conclusion

Through this work we have created a multidisciplinary team involving a computing engineer, a 3D animation specialist, a designer, a fine artist and a theatre producer. With such a mixture of expertise we are delivering a comprehensive set of enhancements to current VEs, as well as exploring new adventures in the design of successful environments.

Specifically, to deliver a communicative environment we propose to make the 'meeting' of avatars a straight forward process facilitated by a responsive environment which adapt to the current activities. We are also hoping to deliver highly communicative avatars, which will deliver an interesting and useful environment as well as enrich the 'conversation' and communication between participants. Such a supportive VE will facilitate co-operative work. We aim to deliver an enjoyable VE by providing the participants with an experiences rich environment.

Acknowledgments

My thanks to Yoann Chapriot for his help producing the facial expressions, and to John Dixon for the work on the Commedia dell'Arte (masks and choreography).

A Testing Methodology based on Memory Awareness States for Virtual Environment Spatial Tasks

Katerina Mania

Department of Computer Science
University of Bristol, UK
MVB, Woodland Road, BS8 1UB, Bristol, UK
mania@cs.bris.ac.uk

Abstract

Simulation fidelity is characterized as the extent to which a Virtual Environment (VE) and relevant interactions with it are indistinguishable from a participant's interaction with a real environment. The research community is challenged to determine ways towards assessing simulation fidelity for specific applications through human-centered experimentation. This document describes a methodology behind comparative studies between real-life situations and 3D simulations displayed on typical desktop monitors as well as on Head Mounted Displays (HMDs). These studies are conducted in the University of Bristol, UK and investigate the effects of immersion on memory and spatial perception.

Background and Experimental Design

Comparative studies between different technologies mostly focus on comparisons of task performance between conditions. In a specific study conducted in the University of Bristol, UK, subjective measures (presence assessments and memory awareness measures) are incorporated together with objective measures of spatial memory recall, in a comparative study of a Virtual Environment (VE) against the real world. We investigated how exposure to a computer-generated replica of an environment, displayed on a typical desktop display and a Head Mounted Display (HMD) would compare to exposure to the same environment and spatial memory task in the real world.

In the process of acquiring a new knowledge domain, visual or non-visual, information retained is open to a number of different conscious states. It would be challenging to identify if different levels of technological immersion have an effect on the actual mental processes participants employ in order to achieve a memory-related spatial task, while using relevant VE systems.

Some elements of a visual space may be 'remembered' linked to a specific recollection event and mental image (episodic memory type) or could just pop out, thus, could be just 'known' (semantic memory type). Tulving [1] introduced this distinction providing the first demonstration that people can make these responses in a memory test, item by item out of a set of memory recall questions, to report their conscious states. 'Familiar' and 'Guess' states are also added requiring participants to choose between these four states for each of their recollections. This experimental methodology offers a range of possible statistical correlations other than solely task performance comparisons; these are related to the actual mental processes that participants employ in order to achieve a certain goal

Results

Although results relevant to the spatial memory awareness task were slightly higher for the desktop and real compared with the HMD condition, there was no significant difference across conditions for the memory recall scores. However, it is interesting to observe that the probability that the responses under the 'remember' awareness state were accurate was higher for the HMD condition compared with the real and also compared to the desktop. In contrast, the probability that 'familiar' responses were correct was significantly higher for the real condition compared to both the desktop and the HMD condition. Since the 'remember' awareness condition is linked with recollective memory, we conclude that although there was not significant difference for the task across all conditions, the actual spatial memory recall for the HMD condition was expressed by more

'vivid' or 'realistic' recollection. In addition, it was expressed with less confidence - higher amount of 'familiar' responses - in the real condition. Presence assessments were significantly higher for the 'real' condition compared to the desktop and HMD conditions.

Generally, the incorporation of cognition-related measures as the distinction between memory awareness states offers a valuable input towards unwrapping results related to presence and memory recall that, otherwise, don't show any significant differences. We established, that usability studies involving only task performance measures while considering a possible VE design or VE technology, as the Hewlett Packard HMD prototype used in this study, are not sufficient to actually conclude on the effectiveness of the design or hardware in question.

Acknowledgements

This research is funded by the Hewlett Packard Laboratories External Research Program.

References

[1] Tulving, E. Memory and Consciousness. *Canadian Psychologist, 26*, pp 1-12, (1985).

[2] Mania, K., Chalmers, A., Troscianko, T., Hawkes, R. The Effects of Levels of Immersion on Memory and Presence in Virtual Environments: A Reality Centered Approach, to appear in Cyberpsychology and Behavior journal, issue 4.2.

Summary

The Future of VR and AR Interfaces

Workshop Summary and Outlook

Wolfgang Broll,[1] Léonie Schäfer,[1] Tobias Höllerer,[2] Doug Bowman[3]

[1] GMD-FIT, Germany,
[2] Columbia University, USA,
[3] Virginia Tech, USA

What can Virtual Reality and Augmented Reality technologies do for us? Can they help us make better use of computers? Will they ever improve the quality of our personal lives? What kind of immersive interfaces could re-shape the way we communicate with computers? How far along are we on that road, and what remains to be done?

This is only a small selection of the questions discussed at the workshop on the future of VR and AR interfaces at IEEE Virtual Reality 2001. Main areas of focus were conversational user interfaces, natural interaction, 3D interaction techniques, haptics, augmented and mixed reality, and mobile and wearable interfaces. In this column we would like to mix facts with fancy in taking a look at tomorrow's interface technology. Today's research paves the road for technological possibilities that go beyond the caves and cables, the glasses and gloves, the pixels and polygons that still dominate the appearance of VR – possibilities that can assist you in virtually all situations of life.

The Vision

Watching an old 2D movie on your projection wall one night, you smile about how the main character is stumbling through the day. You remember this all too well yourself: Rushing to the office, always late, forgetting to switch off the coffee machine when leaving the house. Being stuck in traffic and bumping your car. Not knowing where a meeting will take place. Not having the right information available for a presentation. Not remembering names or places. Missing important dates such as your best friend's birthday. And most of all: lacking the time and organizational skills to make all your arrangements.

Luckily, times have changed for you: When you are about to leave the next morning, a person appears near the door: "Good morning. You modified some files on your notebook last evening. Do you want me to transfer them to your office PC or will you take the notebook with you?" The conversational interface of your personal guardian angel appears as a holographic projection, moving around. "And by the way, your keys are over here!" After you leave, the smart house environment switches off the coffee machine – no need to consult you for such an obvious task.

When you enter the elevator in your office building, the angel again appears in front of you. This time her image is painted directly onto your eyeballs by an inconspicuous retinal display embedded in your eyewear – this way the projection is invisible for anyone else. "The meeting is in the conference room on the 11th floor. Your boss is already there, your visitors will arrive in about five minutes." Nobody else hears these words, since they come through your wireless micro earphones. Knowing your preferences for concise visual presentations, the angel software augments your environment with additional information such as the names and affiliations of the other participants. Since this does not require a conversational interface, the angel stays in the background. Interacting with a conversational interface is still a little bit cumbersome in situations where you cannot use natural speech and gestures, although the eye-blinking interface is actually working quite reasonably for many tasks. Still, you are really looking forward to finally test out the new thought interface. The big controversy about the safety of brain-wave-activated interfaces has given this technology a lot of publicity.

After your meeting the angel appears in your office reminding you about tonight's invitation to an old friend's birthday party. She presents you with a selection of true-to-life renderings of possible gifts for direct ordering.

Coming from work you would usually go jogging or stop by the fitness center. Today, however, there is not very much time left before the party. Thus you decide to do only a short exercise program at home. Your guardian angel knows your usual workout program and supports you by demonstrating the exercises and giving you feedback. Later, on your way to the party you do not exactly remember where to go, but

your angel software has already supplied your car navigation system with all relevant information. Unfortunately, fully automatic driving still has not yet been approved.

The personal user interface gives you pleasant and convenient access to your private information and the electronic services surrounding you, wherever you go. Indeed, the personal virtual interface is so successful that all major providers of electronic services go to great lengths to support it, finally working out a standard for the electronic ether. You sometimes really do not know how you managed life without your personal guardian angel.

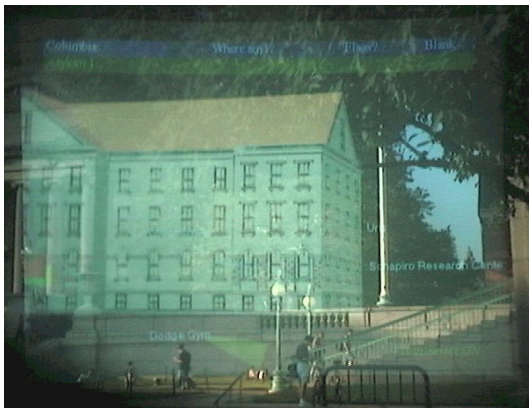


Figure 1: Outdoor mobile augmented reality application, visualizing historic architecture in their original location. © Columbia University

Back to Reality

While real guardian angels are not easy to get hold of, some of the computer technology needed for such a personal assistant is already available; other parts exist in the form of research prototypes, and yet other parts need some technological breakthroughs before they can be realized, let alone be integrated into our daily routines.

Science fiction literature and Hollywood movies and feature films such as *Disclosure*, *The Matrix*, *The Thirteenth Floor* or the *Star Trek* series have already shown us what an unobtrusive, personal, and expansive 3D interface may look like. In these works of imagination the virtual worlds often appear undistinguishable from the real world.

Future VR and AR interfaces will not necessarily try to provide a perfect imitation of reality, but instead will adapt their display mechanisms to the individual requirements of their users. Their emergence will not rely on a single technology but rather depend on the advances in a large number of areas, including computer graphics, display technology, tracking and recognition devices, natural and intuitive interactions, 3D interaction techniques, mobile and ubiquitous computing, intelligent agents, conversational user interfaces, to name but a few (see Figure 2).

Rosenblum [1] and MacIntyre and Feiner [3] take a look into the future of VR technology and multimedia interfaces, respectively. Brooks in his IEEE VR '99 keynote and follow-up survey article [2] gives a personal assessment of the state of the art in VR at the turn of the millennium. What we would like to do for the rest of this paper is to review research agendas in different areas of AR/VR to shed some light on the feasibility of user interfaces of the kind introduced in our guardian angel scenario.

Display Technology

Today's personal head-worn displays are already much smaller and provide a better resolution than only a few years ago, making them suitable for real-world applications and mobile usage. They would already allow us to project a guardian angel within the user's environment, although the user's perception of the image would be limited to a rather narrow field of view. New technologies, such as retinal displays and laser emitting diodes are currently under development. They will further enhance the image quality and the field of view, while facilitating lightweight models. This miniaturization process will continue, providing affordable, high-resolution personal displays, indistinguishable from regular sunglasses, within a few years. Additionally, more sophisticated auto-stereoscopic 3D displays that do not require any wearable equipment at all will become available. This includes multi-viewpoint displays as well as true holographic projections. The technology to be chosen for a particular application or scenario can then be based on the observer's needs or preferences for a personal or a public display. In our scenario, for example, the angel was d by a holographic projection system within the private environment of the user, while personal displays were used in public areas.

Sensors and Devices

There currently exist a large number of input devices for VR and AR, ranging from position and orientation trackers of different kinds to computer vision based systems and haptic interfaces. Tracking and other sensing devices have become smaller and more accurate over the years, but robustness, general applicability and

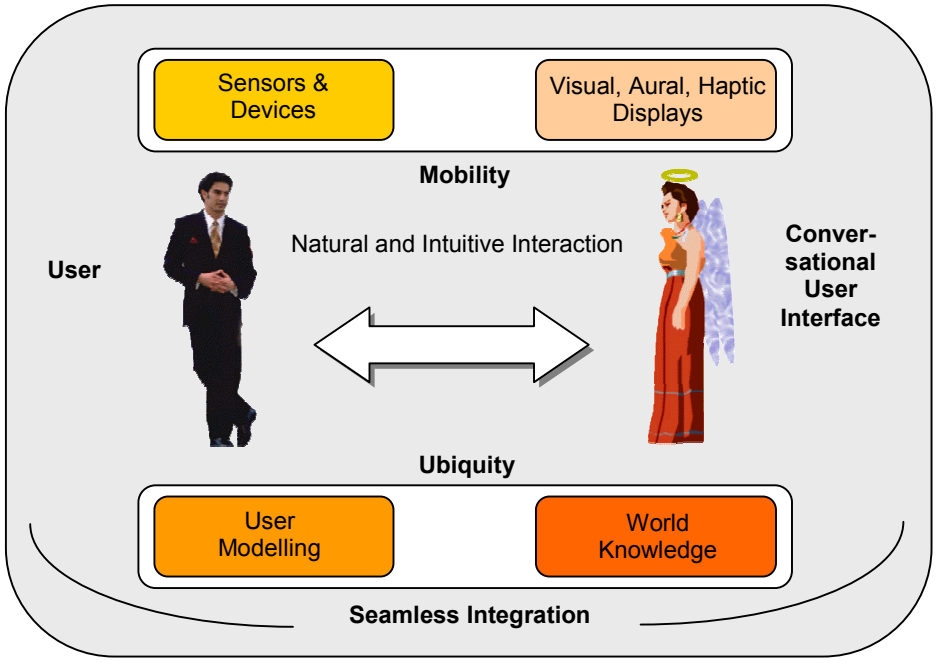


Figure 2: Seamless integration of future AR and VR interfaces, bringing about the *guardian angel* interface of our example

the absence of tether-free solutions remain problem areas. The range of most indoor tracking devices is still limited to a rather small area. Computer vision based systems are currently being explored in prototypes and may emerge as the most flexible and universal tracking device in the future. This, however, will require significant enhancements in performance and robustness. Sensors will be further miniaturized to be less obtrusive. Sensor fusion, e.g. among a network of connected camera systems, as well as between different types of sensors (e.g. ultrasonic, gyroscopic, pedometer- or odometer-based, or vital stats) will give an integrated overall account of the user's movements, behavior and mood. In our example, this provides the angel with the information about the user's whereabouts in the house or office environment, or allows her to react to items and persons viewed by the user and augment them with additional information (e.g. during the meeting).

Mobility

We already use a very large number of mobile devices today: notebooks, PDAs, cellular phones, etc., using a wide range of wireless services. In the future the combination of sensors, smaller and smaller devices and ubiquitous electronic services will not only allow you to access any data from anywhere, but also provide you with precise sensor data about yourself. Similar to GPS-based navigation systems today, future services will be available to announce your exact position within fractions of an inch, everywhere! While presently high-resolution GPS based tracking is used in combination with dead-reckoning for outdoor AR applications (see figure), indoor tracking mechanisms are limited to rather small areas. Hybrid systems will overcome these limitations and provide the accuracy and speed required for AR visualizations.

Natural and Intuitive Interaction

Natural and intuitive interfaces play an important role in the dissemination of VR and AR technology. Today's interfaces are often cumbersome, requiring rather long training of the participants and are often based on heavy or obtrusive equipment. Applications intended for inexperienced users demand a low learning curve and will benefit most from natural and unobtrusive interfaces. Other (more complex) interfaces on the other hand are often more efficient for experts. Some natural input modalities, such as gesture and speech input, as well as recognition of facial expressions are already used in some experimental set-ups today. More robust mechanisms will become available in the future, providing better support for groups

of users (see figure 3). For example, communication by voice and gestures will be feasible, even when talking to other people or residing in crowded places.

Promising research in the area of novel input mechanisms such as brain activity scanning (the thought interface) has already started, although it will take quite a few more years until such interfaces will reach a level to be used in real applications. Support for disabled people will probably be the first type of applications where we will find these types of interfaces. Nevertheless, such interfaces will never completely remove the need for multi-modality – at least from the user's point of view.

Figure 3: Augmented round table: intuitive interaction in collaborative working environments

Conversational User Interfaces

Computer graphics and processing power have advanced dramatically throughout the last 20 years. Assuming these rates of improvements continue in the future, image realism will take great strides towards perfection. Animated characters, complete with synthetic gestures, voice and facial expressions will become commonplace. Virtual humans represent a natural, familiar and convenient user interface. Today we are still far away from providing a realistic conversational interface, but at some point in the future, after 3D video recording has been perfected, there will not be any perceivable difference anymore between a 3D recording of a human and an artificial character like the guardian angel of our example. The latter, of course, does not only rely on highly sophisticated graphics and animations, but also depends on the quality of its behavior. Finally, it has to perfectly adapt to the individual user (based on ubiquitous access to user modeling information) as well as to the local environment, and combine this information with the general world knowledge in order to create a smart conversational interface resembling a real person. It is an open issue if we should always be aware of communicating with a computer as opposed to a human being.

Seamless Integration

Another important aspect comes along with mobility: the continued integration of all electronic data with the network (the web) and therewith the ability to access and update all types of information just in time, just in place. Access to a bank account from a WAP capable mobile phone or email communication using your PDA are just the first signs of the dramatic change we will see within the next years. In our example the smart house environment receives the information that you left, so it could switch off the coffee machine. The angel knows about the people in the meeting room and can inform you about them ahead of time. This integration of digital information with an advanced 3D user interface will make future use of VR and AR interfaces much more attractive to the average user.

Universal User Interfaces

The large success of the WIMP interfaces was based on the fact that users familiar with one application immediately knew how to use the interface elements of another application. While several approaches have been made to provide a similar universal interface for 3D environments, this is still an open issue. Different applications provide very individual interfaces leading to a long learning curve. Flexible, adaptable and more universal interfaces are still a big open issue. A personal conversational interface such as the guardian angel would be one possible approach to cover a large area of applications.

Conclusion

The presented scenario of a guardian angel exemplifies how future developments in AR/VR user interfaces might change the way we interact with computers. While this example is just one of several plausible scenarios, it demonstrates that beside all risks, a completely networked, sensor equipped and visually enhanced environment provides a lot of advantages and opportunities.

Acknowledgements

We wish to thank all participants of IEEE VR 2001's workshop on the future of VR and AR interfaces for the vivid discussions and their inspiring contributions.

References

[1] L. Rosenblum, "Virtual and Augmented Reality 2020", *IEEE Computer Graphics and Applications,* 20(1), 2000, pp. 38-39.

[2] F. P. Brooks, Jr. "What's Real About Virtual Reality?", *IEEE Computer Graphics and Applications, 19(6), 1999*, pp. 16-27.

[3] B. MacIntyre and S. Feiner, "Future Multimedia User Interfaces", *Multimedia Systems*, 4(5), 1996, pp.150-168.

Organizer Biographies

Doug A. Bowman

Doug A. Bowman is an Assistant Professor of Computer Science at the Virginia Polytechnic Institute and State University (Virginia Tech). Doug's research interests are in the intersection between human-computer interaction and computer graphics. Specifically, he is studying interaction and user interfaces in immersive virtual environments, and focusing on the design, evaluation, and application of 3D interaction techniques for complex VEs. His work has appeared in journals such as Presence and IEEE Computer Graphics & Applications, and he has presented research at conferences including ACM SIGGRAPH, the ACM Symposium on Interactive 3D Graphics, and IEEE VRAIS.

Wolfgang Broll

Wolfgang Broll is a lecturer at the University of Technology in Aachen and head of the Collaborative Virtual and Augmented Environments Research Group at GMD - the German National Research Center for Information Technology located in Sankt Augustin, Germany. He received his Diploma in Computer Science from Darmstadt University of Technology in 1993 and a PhD from the University of Tübingen in 1998. He has been doing research in the area of shared virtual environments, multi-user VR and 3D interfaces since 1993. He is now concerned with the intuitive collaboration of multiple users using augmented reality techniques and tangible interfaces. Wolfgang Broll was program chair of the Web3D/VRML 2000 symposium and served on several program committees including CVE?98, Web3D / VRML 2000, CVE 2000, DISTEL 2000 and Web3D 2001. He is the author of more than 30 reviewed papers and has presented research at several conferences including ACM SIGGRAPH and IEEE VRAIS.

Tobias Höllerer

Tobias Höllerer is a Ph.D. candidate and Graduate Research Assistant in the department of Computer Science at Columbia University. Since he joined the Columbia's Ph.D. program in the fall of 1995 he is working with Professor Steven Feiner on virtual worlds and 3D user interfaces. He is writing his Ph.D. thesis on user interfaces for mobile augmented reality systems (MARS). Tobias received M.Phil. and M.S. degrees from Columbia University and a German Diplom in Computer Science from the Berlin University of Technology. He spent a summer each at Microsoft Research and Xerox PARC, where he was a summer intern working on 3D user interfaces and information visualization. Prior to his work at Columbia University he was doing research in scientific visualization and natural language programming. Tobias's academic services include serving on the program committees for IEEE Virtual Reality 2000 and 2001. His main research interests lie in augmented reality, mobile and wearable computing, and 3D user interfaces.

Léonie Schäfer

Léonie Schäfer is a Researcher at the Institute for Applied Information Technology at the German National Research Center for Information Technology (GMD). Léonie's research interests are in human factors in virtual environments. In particular she is interested in Synthetic Characters and Communication in Virtual Worlds. Léonie received her Diploma in Computer Science from the Berlin University of Technology. She has had several years' working experience in Multimedia as Software Developer, in Advertising and Film as Creative Team Member, and in Event Coordination as Project Manager. The expertise she acquired in these branches motivated her to return to research in order to develop new innovative approaches in Multimedia and Virtual Reality. She participated in several national and international research projects, and she has presented research at conferences including ACM SIGGRAPH and Eurographics.