NATURAL INTERACTION IN A DESKTOP MIXED REALITY ENVIRONMENT

René de la Barré, Siegmund Pastoor, Christos Conomis, David Przewozny, Sylvain Renault, Oliver Stachel, Bernd Duckstein, Klaus Schenke

Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, Einsteinufer 37, 10587 Berlin

ABSTRACT

We have developed a combined 2D/3D workstation for desktop Mixed Reality applications which consists of a 100" rear-projection screen with a centre-mounted 30" 3D Accommodation Display and a control desk with video-based finger, eye and gaze tracking systems. The stereo viewing zone (sweet spot) of the 3D display is about 100 mm in height and 130 mm in width. The user's lateral head movements are optically tracked, providing for a comfortable working area. The various tracking data in combination with speech input allow multimodal interaction with Mixed-Reality applications.

1. INTRODUCTION

In the mixed3D project, we develop novel display and interaction techniques for desktop Mixed Reality applications. The term "Mixed Reality" (MR) was coined by Paul Milgram about a decade ago [1]. It describes a system concept where both real and virtual (computer-generated) things appear to coexist in the same space. The potential value of MR systems is increasingly recognized and appreciated in a wide variety of fields, including simulation, maintenance, medicine, architecture, driving, industrial design and entertainment [2]. A great potential lies in the capability of not just mimicking or enhancing properties of the real world, but of exceeding the physical laws governing reality, such as the possibility of stepping back and forth in time, performing an "undo" function, rendering hidden things visible, and using and switching between various artificial tools used for interaction.

A mixed environment is usually composed of the user's real surroundings viewed through a semi-transparent Head Mounted Display with the virtual objects optically or electronically superimposed onto vision. Recently, system concepts have been proposed for non-wearable, desktop MR displays based on auto-stereoscopic free-viewing 3D technologies [3]. Such displays are supposed to be better for long-term use, because the user does not experience the fatigue and discomfort of wearing a heavy head-mounted device. On the other hand, free-viewing displays are generally limited by a restricted working volume and potential space constraints.

This paper presents a prototype desktop system developed in our lab. Figure 1 illustrates the system concept and Fig. 4 shows a current implementation, respectively. A special free-viewing 3D display with 30" screen diagonal is surrounded by a large 100" 2D rear-projection screen. The 3D display projects high-resolution stereoscopic images (two times UXGA) "hovering" 25 cm in front of the display. The user may virtually touch the aerial images floating in reach with his/her hand. The surrounding 2D screen serves to visualise 3D objects, too large or too many for the limited size of the 3D display, as well as virtual tools which the user may select by various combinations of speech and pointing. Pointing is possible with the finger or simply by looking at the object selected (gaze input). Speech input confirms a selection. Selected 3D objects may be grasped with the naked hand and moved into the 3D area for manipulation by a simple drag & drop operation.

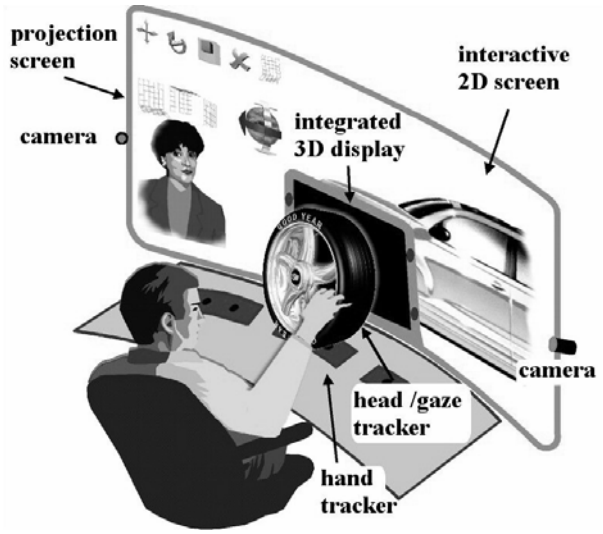

Fig. 1: Concept of a Mixed Reality workstation developed in the mixed3D project [4].

In summary, this concept is based on the vision of a workplace environment allowing users to virtually "dive" into the technical structures of a complex system to be developed, such as a new car concept, and to visualize selected depth layers by intuitive input gestures (pointing and grasping with the naked hand). Moreover, the system should enable direct manipulation of virtual 3D objects with the hands, thus enhancing current 3D object generation tools (specialized CAD and 3D rendering software) through natural sculpturing operations.

2. SYSTEM SETUP

The basic system setup consists of a special high-resolution display generating floating 3D images within the user's arm reach, a conventional video projector used in combination with a curved rear-projection screen, a large workbench with cameras and infrared illuminators embedded into the desktop, and various cameras and infrared diodes for gaze and head tracking mounted into the frame of the 3D display. For voice control, the system employs a commercial speech recognition engine. Haptic feedback is provided by a PHANToM^{TM} force feedback device. The WORKBENCH^{3D} software provides a joint 3D data space for the displays and generates audio-visual and haptic output. Moreover, this special software integrates various unconventional interaction and input devices. In the following we will describe the main components in more detail.

2.1. The combined 2D/3D Display

Recent approaches to 3D display work without the need of special glasses. Here, the optical elements required to separate the stereo views are included in the monitor. Most of the practical free-viewing display technologies have the typical drawbacks of conventional displays: the accommodation distance is fixed, irrespective of where the virtual objects appear, leading to an accommodation conflict, when a user tries to touch the virtual objects with his/her hand or a real tool. The 3D Accommodation Display developed in the mixed3D project solves this problem by generating aerial 3D images projected in front of the display screen. Fig. 2 illustrates the basic optical setup.

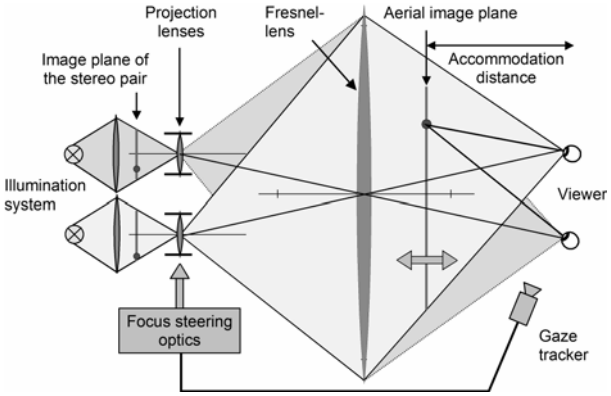

Fig. 2: Top view of the principle of operation of the 3D Accommodation Display. The viewer accommodates on the aerial image plane and perceives a floating 3D object. The Fresnel lens images the exit pupils of the stereo projector at the locations of the left and right eye, respectively, and hence separates the constituent stereo images. The aerial image plane is dynamically aligned with the current position of the viewer's fixation point.

Two video projectors in a stereo projection arrangement are used to create a very bright stereo image pair. The stereo projector is focused on an aerial plane floating in front of (or behind) a large convex lens. The aerial image constantly changes its location according to the fixation point of the user. Hence, the viewer perceives an image of the currently observed virtual object in his/her accommodation distance [5]. This MR display creates holography-like images with a large depth range and supports the link of accommodation and convergence. On the other hand, it requires on-line measuring of the user's gaze direction and adaptation of the optical components of the display.

The developed display prototype creates a single aerial image plane 25 cm in front of the Fresnel lens for direct MR interactions within the user's arm reach. The display is mounted on a laterally movable stage controlled by a video head tracker. The displays sweet spot has a size of 100 (vertical) x 130 (horizontal) mm; the head tracker makes the sweet spot follow the user's eye position when moving laterally (Fig. 3).

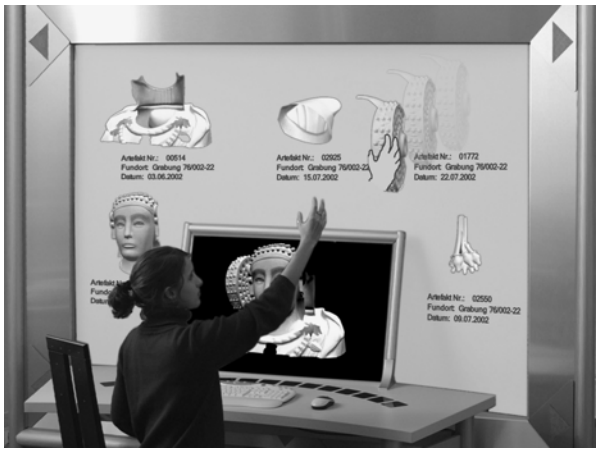

Fig. 3: Combined 2D/3D MR display. The prototype workplace allows users to "drag" computer-generated or scanned 3D objects from the high-resolution 2D rear-projection screen and "drop" them into the centre 3D display for direct manipulation. Video-based hand-gesture recognition devices are embedded in the desktop. Cameras behind the front plate of the 3D display serve to recognize the viewer's gaze direction and head position. Both computer vision devices are used for multimodal interaction.

Normally, human stereoscopic vision is limited to a central binocular region where the visual fields of the left and right eye overlap. The binocular field is flanked by two monocular sectors within which objects are visible to one eye only. In these sectors, spatial vision relies on monocular depth cues such as shading, perspective and occlusion. The idea of the combined 2D/3D display was derived from this basic concept of vision. The display shows 3D images in the centre part of a wide-screen display. The surrounding area reproduces only monocular depth cues, but subtends a major portion of the user's visual field. The entire

display area belongs to a joint data space. Hence, it is possible to make the 3D imagery merge seamlessly with the surrounding 2D region. Optionally, the 2D area may offer additional room where the user temporally places virtual objects and relevant tools.

With the current prototype the 3D display has a diagonal of 30" and provides a spatial resolution of 2x 1600x1200xRBG (2x UXGA). The 100" 2D image projector provides UXGA resolution, too.

2.2. Devices for multimodal interaction

The workplace offers various interaction modalities. Voice input combined with gaze tracking allows the user to visually and verbally address interactive objects. The display tracks the user's movements and adapts the optics as well as the 3D perspective to the current head position (motion parallax). Hand gestures are recognized by devices embedded in the desk. When grabbing objects from the far 2D screen, a realistic looking simulated hand is shown, virtually extending the user's arm; direct MR interaction with the real hand is possible if objects are within arm's reach. Optionally, conventional devices such as a mouse and a keyboard are available. In the case of concurrent (conflicting) inputs, rules defined by the application decide what device will get the focus.

2.2.1. Hand Tracker

The hand tracker employs active infrared illumination from the bottom (desktop) in order to facilitate segmentation of the hand in the video images. The most frontal segments of the hand are recognized as the finger tips. The cameras in the desktop form a multiple-baseline stereo setup (Fig. 4). Depending on the current 3D position of the user's hand a subset of cameras is automatically selected to form a temporary stereo unit. This allows flexible tracking and precise recognition of finger positions and simple gestures over a wide tracking range. The measurement rate is 50 Hz at a precision of 5x5x15 (x,y,z) mm in a lateral tracking angle of 100 degrees.

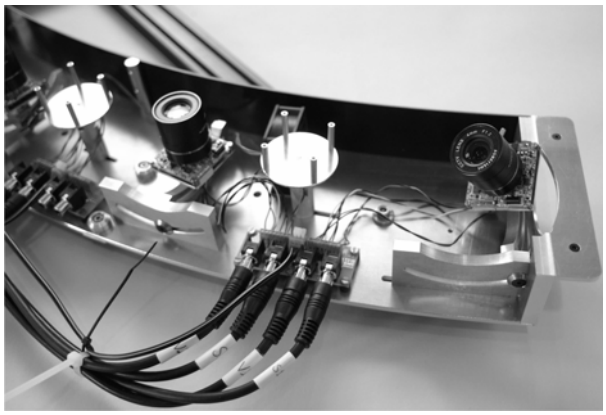

Fig. 4: Detail of the hand tracker hardware

2.2.2. Head Tracker

The head tracker initially analyses the edges in the video image and compares the orientation patterns with a database containing hundreds of pre-stored human eye pattern using a neural net approach. After the eye position has been found, the current user's individual eye pattern are used for tracking with a simple block matching algorithm (implemented in assembler using the Pentium IV SSE2 multimedia instruction set). This two-stage approach was used to speed-up and simplify tracking. With high-speed stereo cameras the tracking rate is 120 Hz at a precision of 3x3x10 (x,y,z) mm (pupil position). The stereo cameras baseline is 800 mm, since the cameras are mounted to the left and right side of the 30" 3D display.

2.2.3. Gaze Tracker

The gaze tracking algorithm applies a special cornea-reflex method. It senses the user's current point of fixation (lines of sight of both eyes) at a rate of 50 Hz and a precision of about 1 degree with a single stationary camera. Due to the wide-angle optics in combination with a 1280x1024 resolution camera, the user may move in a range of 300x300x300 mm. A two-step calibration procedure requires minimal individual calibration for the end-user. All cameras and (pulsed) infrared illuminators used for the various trackers are synchronized, in order to avoid mutual interferences.

2.2.4. Sensor data fusion

Using sensor fusion techniques, the various tracking modules can be combined to setup novel interaction tools allowing smart application. A new and very natural pointing tool was created by combining the head (pupil) tracker with the finger tip tracker [6, 7]. With such a tool the user can easily control the position of a cursor on the screen by pointing in the desired direction. Tracking the user's gaze direction allows the speech recognition engine to react more reliably and context sensitively. Tactile feedback when touching a virtual object is provided by a commercial force-feedback device. Within the 3D area there is a perfect correspondence between the visual and tactile space. Corresponding spatialized audio is produced by an array of loudspeakers.

2.3. The WORKBENCH^{3D} software

WORKBENCH^{3D} [6] is a development tool which helps software designers to create multimodal multimedia applications with the Microsoft Visual Studio® .NET environment. In order to guarantee flexible adjustment to different hardware configurations, the WORKBENCH^{3D} supports visualization on multiple monitors (combined 2D/3D displays), communication between computers via .NET, environmental sound as well as voice input and output and user guidance using an avatar. It supports the PHANToM^{TM} force-feedback device from SensAble Inc. The inheritance of the .NET Framework makes it

possible to enhance the API with other technologies. The software supports DirectX® technologies and includes interfaces for various unconventional interaction devices (Fig. 5). The video tracker software communicates with the API using a DCOM-interface.

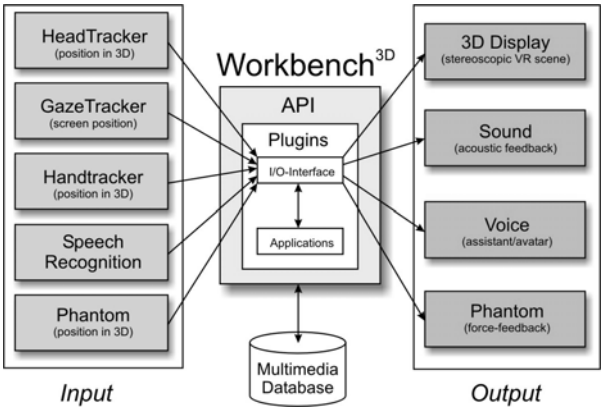

Fig. 5: Structure of the WORKBENCH^{3D} software

3. SAMPLE APPLICATIONS

The first prototype application developed was a computer simulation, where fragments of an archaeological finding could be assembled (see Fig. 2). The fragments displayed in the 2D area were selected by simply pointing at them with the finger tip. A spoken "affirmative" command (Yes, OK) would lift the selected object from its 2D screen position, so that it could be moved with the hand into the perception and interaction area of the 3D Accommodation Display. There, it could be shifted and rotated until it matched with the other fragments already assembled in the 3D area. The user could switch to other applications by pointing at and selecting the respective icons. The user could call for help by an avatar while using an application. Further sample applications are under development and will be presented at the workshop.

4. CONCLUSIONS

One of the essential requirements of Mixed Reality is the availability of adequate visualization technologies that allow users to perceive the mixed worlds and interact with them in a natural and familiar way. The challenge here is not only to correctly blend the virtual and the real scene components, but at the same time to support the natural perception mechanisms of human vision. The last aspect is of great importance in the context of comfortable user interaction and long term visual perception. In the mixed3D project a display system has been developed that provides support for direct interaction with virtual objects in close range distances (within the user's arm reach). Apart from the display problem, new solutions were found for multimodal interaction with virtual objects. These include techniques for remote and unintrusive sensing of the user's gaze direction, head and eye position, as well as finger positions and hand gestures. A special software package called WORKBENCH^{3D} allows combining various modalities including speech and fuses these modalities in order to guarantee intuitive and reliable interaction with mixed reality applications. Moreover, the software supports multiple screen setups such as the one implemented in the combined 2D/3D display presented in this paper.

The desktop system was successfully tested in a sample application allowing users to virtually grasp and combine archaeological artefacts. Systematic testing of user experiences (viewing comfort, usability of combined interaction modalities) is planned. Our earlier approach aiming at non-command, continuous human-computer interaction [9] showed that stability and unnoticeable delay of the tracking devices are of utmost importance. A future challenge is to support two-handed input allowing users to create and manipulate virtual 3D objects the same way as artisans form sculptures from clay. This will require a more elaborate tracking scheme reading the gestures of both hands.

5. REFERENCES

[1] Milgram, P. and Kishino, F., "A Taxonomy of Mixed Reality visual displays", *IEICE Transactions on Information Systems*, E77-D (12) pp 1321-29, 1994.

[2] Tamura, H., "Steady Steps and Giant Leap Toward Practical Mixed Reality Systems and Applications." Proc. Int. Status Conf. on Virtual and Augmented Reality, Leipzig, pp. 3-12, 2002.

[3] Pastoor, S. and Conomis, C., "Mixed Reality Displays", in O. Schreer, P. Kauff and T. Sikora (eds.), *3D Videocommunication*, Wiley, Chichester, 2005 (in print).

[4] Pastoor, S., "mixed3D – 3D-Techniken für Mixed-Reality-Systeme", project description, Fraunhofer-Institut für Nachrichtentechnik, HHI, Berlin, 10.8.2001.

[5] Pastoor, S. and Boerger, G., "Autostereoskopisches Bildwiedergabegerät (Autostereoscopic Display)", Patent DE 195 37 499.

[6] Przewozny, D. and de la Barré, R., "Fusion von 3D-Finger- und 3D-Augenposition für ein Mensch-Maschine Interface", 3D-NordOst 2003, Berlin, 5.12.2003, pp. 99-101.

[7] de la Barré, R., Przewozny, D. and Neumann, F., "magic pointing - Berührungslos Bedienen mittels Interpretation der Augen- und Fingerposition", HHI, 2004, http://www.hhi.fraunhofer.de/german/im/projects/mixed3d/

[8] Renault, S. and Stachel, O., "Unterstützung multimodaler und natürlicher Interaktionen durch die WORKBENCH^{3D} bei der Modellierung und Visualisierung von VR-Umgebungen auf autostereoskopischen Endgeräten", GFaI, 2004. http://salza.gfai.de/3d_display_if/

[9] Pastoor, S. and Liu, J., "3-D Display and Interaction Technologies for Desktop Computing", in B. Javidi and F. Okano (eds.), *Three-Dimensional Television, Video and Display Technology*, Springer, Berlin, 2002.