Cascading Hand and Eye Movement for Augmented Reality Videoconferencing

István Barakonyi[1*] Helmut Prendinger[2] Dieter Schmalstieg[1] Mitsuru Ishizuka[3]
[1]Graz University of Technology, Austria [2]National Institute of Informatics, Japan [3]The University of Tokyo, Japan

ABSTRACT

We have implemented an augmented reality videoconferencing system that inserts virtual graphics overlays into the live video stream of remote conference participants. The virtual objects are manipulated using a novel interaction technique cascading bimanual tangible interaction and eye tracking. User studies prove that our user interface enriches remote collaboration by offering hitherto unexplored ways for collaborative object manipulation such as gaze controlled raypicking of remote physical and virtual objects.

CR Categories: I.3.6 [Methodology and Techniques]: Interaction techniques; H.5.1 [Multimedia Information Systems] - Artificial, Augmented, and Virtual Realities.

Keywords: augmented reality, remote collaboration, eye tracking, tangible interface, multimodal interaction

1 INTRODUCTION

While computer-supported collaborative work (CSCW) is one of the evident application domains that Milgram et al.'s definition of Augmented Reality (AR) [23] suggests, "classic" AR applications require that users are co-located, sharing the same physical space. We have created an augmented reality-based videoconferencing tool that allows users to discuss and manipulate real and virtual objects over great distances while preserving non-verbal communication and part of the conference parties' physical environment. We have experimented with user interface techniques that make communication and interaction smoother while discussing real and virtual objects with a remote videoconference party. Physical objects are pose-tracked with handheld fiducial markers and virtual objects are assigned to tracked physical placeholders.

In our application scenario videoconference parties use both hands to carry out object manipulation and interaction tasks such as translation, rotation or selection with fiducial markers. While two-handed tangible interaction may rely solely on complex 2D and 3D gestures aided by traditional input devices such as mouse and keyboard, we have found that exploiting the human eye as a natural input device during bimanual object manipulation yields faster, richer and more intuitive communication between partners in remote collaboration tasks. To support this statement, we have implemented and evaluated an interaction technique cascading bimanual tangible interaction and eye tracking.

Figure 1 shows a schema of our AR videoconferencing system enhanced by eye tracking. This paper first discusses how tangible augmented reality and non-intrusive eye tracking enhance remote collaboration while comparing our system with related work, then presents our application scenario and interaction technique with implementation details. We conclude our paper with the results of our user study.

* Contact author's email address: bara @ icg.tu-graz.ac.at

2 REMOTE COLLABORATION IN AR

Users of AR applications can see the real world, which provides a reference frame for their actions. They can see themselves and their collaborators, enabling smooth communication with non-verbal cues during collaborative work. Moreover, a virtual space with synthetic objects is aligned with and superimposed onto the real world and shared among the users. Thus changes made to manipulated objects during the collaborative session are distributed and immediately visible to all participants.

Unfortunately, this form of shared AR requires that the collaborators are sharing the same physical space, making it incompatible with remote collaboration over greater distances. State-of-the-art remote collaboration tools include audio/video conferencing and application sharing to help bridge the distance by displaying the remote parties' real environments. Application sharing provides a synchronized view into a 2D/3D workspace, while videoconferencing tools enable the use of natural conversation, facial expression and body gestures.

From a rich set of existing remote collaboration tools for VR and AR environments, we focus on videoconferencing since AR requires constant display of the real environment for correct registration and overlay of virtual objects.

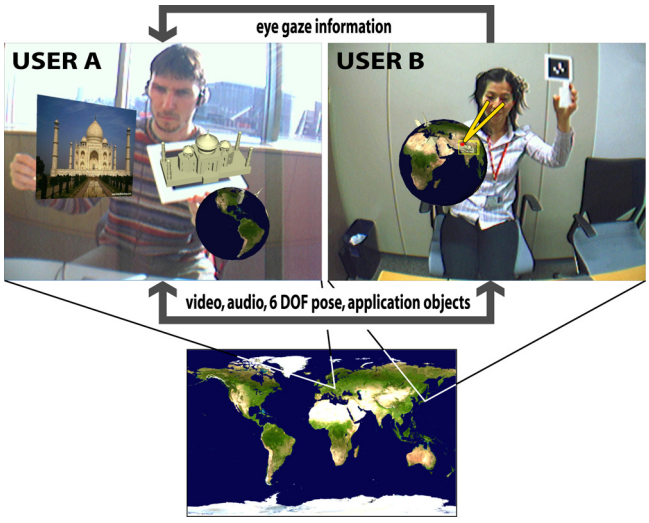

Figure 1. AR videoconferencing system enhanced by eye tracking (the rays emanating from user B's eyes were added only for visualization purposes)

2.1 Videoconferencing in Mixed Reality

An early work by Billinghurst and Kato [3] presents remote collaborators by live video images texture-mapped onto a flat polygon, which are attached to tangible objects that can be freely positioned in the user's AR space. A similar approach was used in the cAR/PE! system [28], which places live video streams of meeting participants into an augmented virtuality environment. The latter video texture techniques were extended by Prince et al. [27] into live 3D actors. Their system called 3D-Live captures the remote collaborator in real time using a shape-from-silhouette algorithm. The system is able to generate a virtual image of the real person from arbitrary viewpoints of AR users, who thus perceive the whole body of the remote user in 3D.

Our system is the first desktop-based videoconferencing tool using standard, low-cost hardware that inserts 3D graphics overlays from a 3D application directly into the image of the remote participant. We combine the assets of videoconferencing and application sharing: the shared virtual objects are superimposed on the live video stream of the conference party. Technical and implementation details can be found in our previous work [2].

Each AR conferencing client is equipped with two monitors to display the local and remote workspaces. A webcam placed on top of the screen rendering the local workspace records the user's physical environment. The video stream is horizontally flipped, thus the screen acts as an augmented mirror. Audio headsets are used for voice communication. One of the clients is equipped with an eye tracker to monitor eye gaze.

3 Tangible Interaction

Tangible [17] and graspable [12] interfaces have become prominent interaction techniques for applications using the metaphor of assigning digital information to physical artifacts. AR applications [4] [22] greatly benefit from these techniques since real objects often serve as representations and control handles for virtual objects in the users' physical environment.

Users of our AR videoconference system are untethered, they only need to wear an audio headset, therefore both hands can be used in the communication process for object manipulation and gesticulation. Handheld fiducial markers serve as tangible user interface (TUI) elements and are used for positioning virtual objects in the environment and for gestural commands. The gestures require that the left and right hand work synchronously since the relative marker position and orientation convey important information in the application context.

3.1 Bimanual interaction and object selection

Bimanual interaction with virtual objects may appear natural to users since there are numerous everyday tasks where we need a coordinated movement of both hands. However, as Guiard [15] and Hinckley et al. [16] demonstrate, two-handed interaction does not necessarily yield better or smoother performance without exploiting the asymmetric roles of the dominant and non-dominant hand. Positive examples for employing the principles of asymmetric two-handed interaction include the Toolglass technique [19] and the Voodoo Dolls technique [26].

Dynamic collaborative environments such as AR applications contain numerous stationary and mobile objects, therefore a selection mechanism is needed to highlight the subjects of collaborative tasks. As the analysis of Pfeiffer et al. [25] points out, verbal reference (e.g. "Show me that blue round object on your left.") may be ambiguous since humans use several competing reference frames in task-oriented dialogs including a user-oriented, communication partner-oriented and object feature-oriented view. This deictic ambiguity imposes undesired and tedious synchronization between the two users ("Do you mean this?", "No, not that on your right, it's on your left.")

During bimanual marker manipulation both hands are occupied, therefore using an extra pointing device such as a mouse to disambiguate object selection and referencing would be cumbersome. Switching between the TUI elements (in our case the handheld markers) and an extra interaction device would result in breaking eye contact with the collaboration partner and potential misrecognition of interaction gestures.

4 Eye Tracking

Videoconference users tend to look at the point of interest or subject of the conversation during a collaborative session. Bolstering experiment results by Velichkovsky et al. [33] suggesting a strong correlation between visual fixations and user attention, we speculate that using the human eye as a natural input device provides us with an effective tool for selecting objects on videoconference displays.

4.1 Eye-based interaction in videoconferencing and VR

Eye tracking research embraces more than two decades of user experience to date. As the overviews of Duchowski [9] et al. and Wilder et al. [35] demonstrate, gaze-based interaction has been successfully used as input device for both human and computer in various VR applications ranging from rendering virtual scenes with level-of-detail degradation scheme matching foveal acuity, through industrial inspection to the assessment of advertisement effectiveness.

Eye tracking has been used in videoconferencing applications mostly to enhance face-to-face discussions. The Hydra project [31] and the GAZE-2 [34] system are prominent examples of improving eye contact and solving the parallax problem in present videoconferencing systems. Fussell and colleagues [14] generated visual clues about the gaze targets of a helper user providing assistance for a remote collaboration partner using a scene-oriented and a head-mounted camera with an integrated eye tracker. Unlike our work, none of the previous videoconferencing projects has exploited eye tracking for interaction with objects within the videoconference window.

Tanriverdi and Jacob [32] evaluated eye movement-based interaction in virtual environments and found it faster than conventional 3D pointing, although users reported limited spatial information perception. Asai et al. [1] combined a head tracker and an eye tracker to create a pointer for immersive stereo projection screens. Their user studies showed that eye tracking was significantly faster than joystick-based interaction for selecting static virtual objects. Ciger et al. [7] experimented with gaze tracking for social interaction with virtual humans.

4.2 Eye-based interaction in AR

Before eye tracking technology entered AR systems, gaze direction was estimated from head pose. Kaiser et al. [20] used gaze direction derived from the head's normal vector to disambiguate multimodal 3D interaction in immersive VR and AR environments. The AR Workbench by Novak et al. [26] used gaze tracking instead of a head pointer, and developed an attentive user interface for AR-based automotive environments. They built a custom-made eye tracker into a head-mounted display (HMD). The EyeCatcher project [10] and Curatu et al. [8] used an eye tracking system also integrated into an HMD. Both systems require a lengthy and precise calibration process and are prone to suffer from HMD slippage.

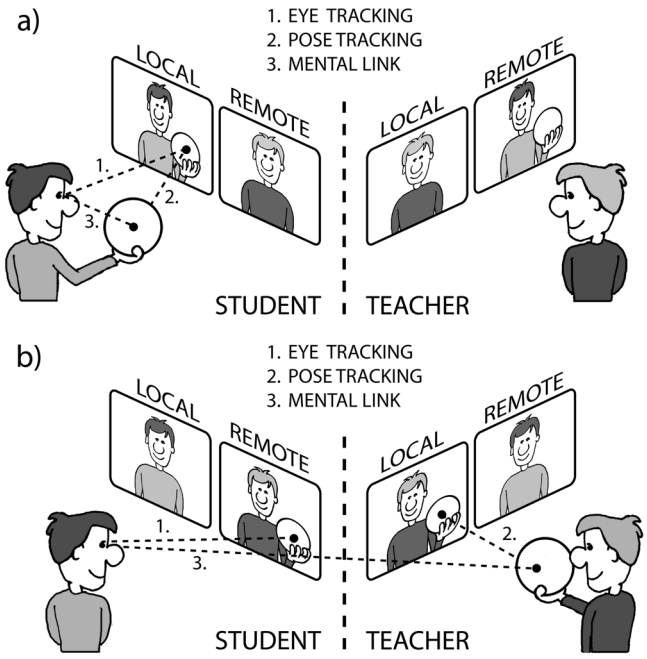

Figure 2. Student-teacher collaborative object selection scenarios: a) selection of a local object, b) selection of a remote object

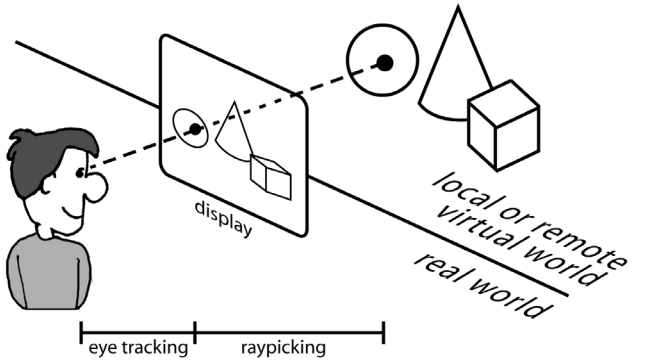

Figure 3. Combining eye tracking and raypicking

Our system is the first desktop-based AR system to integrate gaze tracking as an input device. Although the working volume is smaller than with HMD-based systems, users do not have to wear any head- or body-mounted equipment, increasing user acceptance and comfort. Since users employ their own hands to position selection targets with tangible markers, they are provided with extra spatial cues, eliminating Tanriverdi and Jacob's weak spatial sensing problem [32] described in the previous section.

4.3 Cascading manual and eye-based interaction

Although eye-based interaction has been found significantly faster than manual techniques, users often preferred hand-based input due to numerous problems caused by imposing the constraints of digital input devices onto the human visual sensors. We discuss the potential problems and our solution to them in the "Interaction Design" section.

As the MAGIC pointing technique [36] illustrates, eye tracking may perfectly complement other interaction methods, making their combination more effective than any single approach. The MAGIC technique uses gaze to dynamically warp the "home" position of the pointing cursor to be at the vicinity of the user's fixation point, which greatly reduces cursor movement amplitude to reach target objects in the center of user focus.

Bolt and Herranz [5] used two-handed "coverbal" gestures for multimodal input, and considered modulation of gestural intent with gaze to disambiguate human-computer communication. Although their system design contains a powerful combination of hand gestures, speech and gaze direction to help the computer identify user intention and execute desired object manipulation tasks, eye gaze was not implemented and no collaborative scenarios were considered in their work.

5 Collaborative Object Manipulation

Augmented reality videoconferencing offers novel selection mechanisms including the collaborative selection of local and remote artifacts, and picking mobile virtual and physical objects on a 2D display. Having compared our work with related research, the following sections present our own novel interaction method for dynamic object manipulation tasks that cascades bimanual tangible interaction and eye tracking. We speculate that this combined technique yields superior performance in dynamic environments such as AR, where objects frequently change their pose.

Figure 2 illustrates two basic scenarios using our combined interaction technique for collaborative object manipulation: a selection task for a local and a remote object. We have attempted to create a user interface that diminishes seams between remote collaborators' workplaces: bridging the large distance (physical seam) and using the computer mediating communication in a non-intrusive way (virtual seam).

For both tasks we took a real life situation as reference: an architecture student and a teacher are co-present in the same room studying famous buildings around the world. Each time they want to discuss a certain architectural landmark, they have to look for the corresponding city on a handheld rotating world globe, point at it, and ask questions or explain the chosen building's history. In our test scenarios we transplant this ideal, seamless reference workspace into two remote, distributed workspaces that are connected by our AR videoconferencing system.

5.1 Object selection in local workspace

Figure 2a illustrates the local object selection scenario. We want to preserve the same mental link for partners during remote collaboration sessions as if they were working in the same room: the student holds the globe in one hand, rotates it with the other, and selects a country with a building she wants to learn about by pointing at it with a finger. The teacher perceives and acknowledges the selection, and starts the explanation.

In a computer-supported remote collaboration setting we employ tangible fiducial markers to control the pose of the globe's virtual representation or track the physical globe itself. The student's gaze direction is substituted with an eye pointer controlled by gaze tracking. The eye pointer marks the point on the computer screen from which we shoot a ray into the virtual representation of the student's local workspace (see Figure 4). This technique is called raypicking [6]. Figure 3 shows a schematic view of the combination of eye tracking and raypicking. The student's finger pointing gesture is replaced by a selection gesture with a handheld marker that highlights the current focus of attention estimated by the eye tracker. This corresponds to the point where the virtual ray intersects the virtual world model of the student's workspace, which in our case is the handheld virtual globe. All visual information related to the selection task is transmitted to the teacher's computer including a live video stream of the student's local workspace and the

highlighted object. Naturally, an audio communication channel employing the users' headsets is maintained.

The remote teacher receives a live view of the student's workspace along with information about object manipulation activities, and displays it besides her own local workspace view. The live video capturing the teacher's real environment is augmented with virtual 2D and 3D objects that can be manipulated (translated, rotated, scaled etc.) by the teacher's handheld fiducial markers. See Figure 5 for illustration. The virtual content depends on the student's current object selection, i.e. in our scenario a photo and a detailed 3D model of the selected building enrich the teacher's audio explanations and non-verbal cues. Similarly to the teacher seeing the student's workspace and actions, the student also sees the teacher's augmented workspace in a window next to her local workspace.

5.2 Object selection in remote workspace

Figure 2b depicts a remote object selection scenario. The difference between the real-life mental model of the local selection scenario and the one we want to recreate here is that now the teacher is holding and manipulating the world globe, while the student gives verbal instructions for rotating the globe to the desired geographical location. When an interesting edifice comes into view, the student indicates choice and requests information about the building from the teacher by finger pointing.

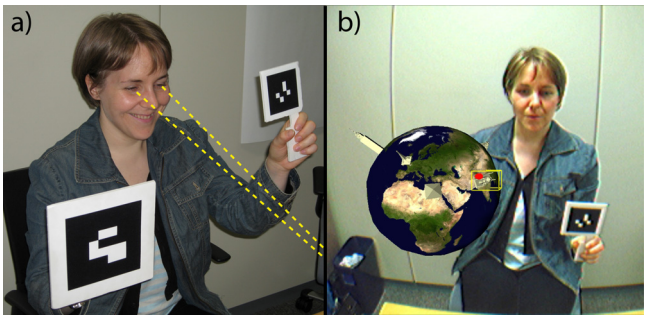

Figure 4. Student a) in real life and b) in the AR videoconference window. Dotted lines were added to visualize gaze direction for explanatory purposes in image a). The red dot in image b) is the eye cursor. The yellow bounding box around the virtual building model (Taj Mahal) displays the result of raypicking, i.e. the currently selected object

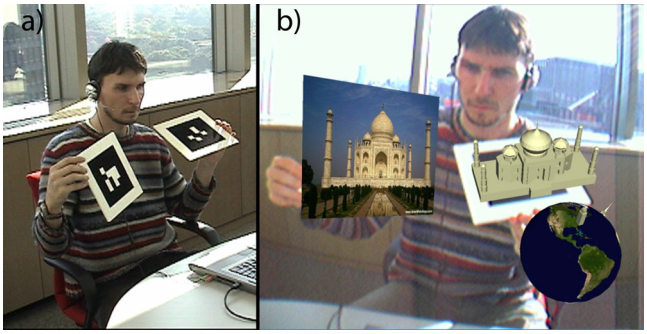

Figure 5. Teacher a) in real life, and b) in the AR videoconference window. A photo and a detailed virtual model of the student's selected building are displayed over the teacher's marker to enhance audio explanation and non-verbal cues

Again, we apply pose tracking and fiducial markers to maintain the virtual workspace model of the teacher and to map the pose of the real globe to the two-dimensional videoconference window. The virtual workspace reappears on the student's side next to the local view of her own workspace. Eye tracking is used to control the eye pointer on the remote workspace display that sets the starting point for raypicking, however, this time we cast a ray into the remote workspace of the teacher.

On the teacher's side a visual highlight is displayed around the selected object, giving a hint about the intersection point of the student's virtual gaze direction ray with the teacher's videoconferencing screen, and the result of raypicking in the teacher's virtual workspace model. Thus the student uses eye gaze to raypick objects in a remote virtual scene. Figure 3 provides illustration again.

As will be presented in following sections, from the two collaborative object selection cases we evaluated the local object selection scenario to observe the careful coordination of hand and eye movement for a single user. The remote object selection scenario requires synchronized movement between the local user's hands and the remote user's eye gaze, and was not considered for this paper for space reasons.

5.3 Interaction sequence

The interaction sequence to select and discuss a building with our AR videoconferencing system consists of a few easy steps. To illustrate a typical interaction flow, let us suppose the student wants to learn architectural details about the Empire State Building. She then does the following:

1. unlocks the globe's rotation
2. rotates the globe to view North America
3. locks the globe's rotation to stabilize view
4. glances at the Empire State Building's virtual model
5. selects the building

After the selection has been made, the system notifies the teacher's computer that triggers the display of a virtual 3D model and a photo of the Empire State Building. Then

6. the teacher moves around the virtual objects with tangible markers and uses non-verbal communication to support and enrich oral explanations

Figure 6 illustrates all steps with annotated screenshots. We use handheld tangible fiducial markers to track the users' hands and to position the virtual objects in the local workspaces. We have considered the principles of asymmetric hand roles during the design of interaction gestures. We employ the non-dominant hand to coarsely position the virtual globe in the user's physical environment, while the dominant hand is carrying out fine movements. We exploit the relative positions of the hands as a kinesthetic frame. The relative horizontal distance of the two markers is mapped to the globe's horizontal rotation angle, and passing a predefined threshold for the relative vertical distance triggers building selection. Since the dominant hand's function is overloaded and hand movements cannot be precisely constrained along orthogonal axes, the globe's rotation angle can be locked in order to avoid accidentally changing orientation during a selection gesture. The globe can be locked or unlocked by penetrating a virtual bounding sphere around the camera with the dominant hand's marker.

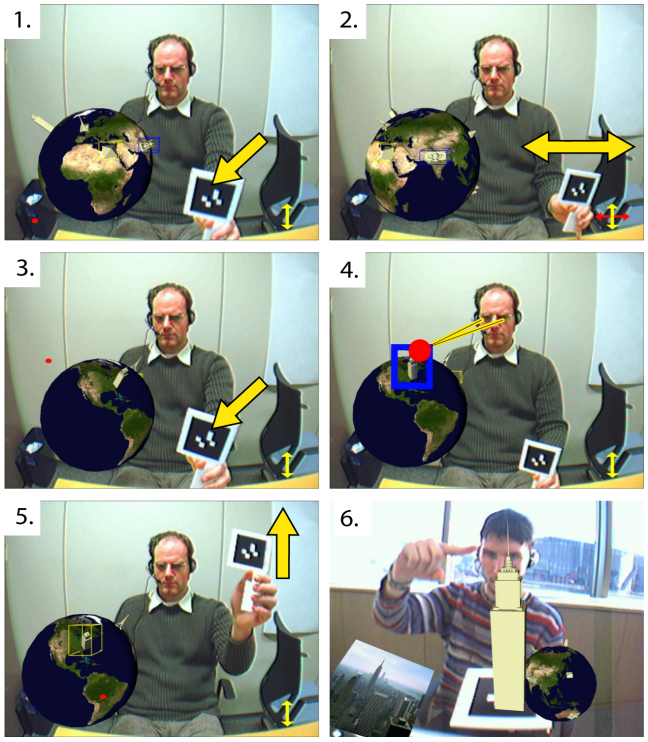

Figure 6. Interaction sequence for object selection (screenshots have been annotated for explanatory purposes)

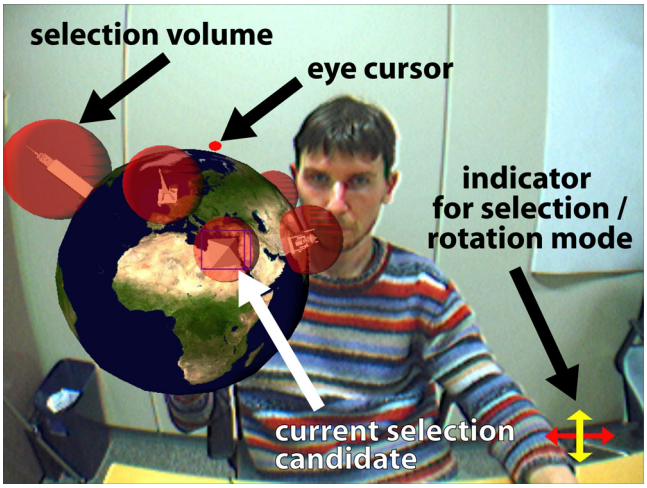

Figure 7. Auxiliary interface elements. The bounding volume visualization and the eye cursor are hidden by default but can be selectively turned on for adjustments and evaluation

Object selection requires coordinated hand and eye movements: after rotating the globe to display a desired building, the user has to look at it and simultaneously make a selection gesture. During our first pilot tests on usability we have found that users can rely on natural, real-life object manipulation patterns instead of applying artificial techniques such as the "zoom-and-pick" method [13] to increase selection accuracy. If an object appears to be too small to be selected, users involuntarily make it bigger by moving it towards their viewpoint (in our case the camera), and if an object obscures parts of their workspace or communication partner, they move it out of the way.

6 Interaction Design

Due to the nearly unconscious and rapid nature of human eye movements, gaze tracking may provide an exceptionally natural and quick communication device between human and computer [18]. However, if used in a naïve way, the eyes may clearly underperform when compared to more traditional input devices such as a mouse or joystick [35]. We have incorporated workarounds into our GUI design to overcome several limitations of eye tracking. Problems and solutions are summarized in the following list.

Jittery eye motion: The human eye does not possess the same controlled behavior as a desktop mouse. Our eyes involuntarily jump between fixation points with rapid saccades. We have developed and experimented with special filters to make eye pointer movement appear stable and smooth: an elastic filter damping jitters and a fixation detection algorithm [29] ignoring eye tracking data during saccades. In our evaluation scenario we used the elastic filter-based approach with damping and strength parameters carefully balanced along the dimensions of responsiveness and stabilization speed.

Tracking inaccuracy: Due to calibration errors and the one-degree size of the fovea, eye tracking is not suitable for operations requiring an accuracy of a few pixels. For enhancing selection, we assign invisible bounding volumes to selectable objects. These selection volumes (SV) are simple shapes that encapsulate their parent object to hide fine details otherwise difficult to catch with an eye pointer. When hit by the raypicking action, they notify their parent object, which then becomes a selection candidate. SVs are automatically assigned to selectable objects, their shape and size are dynamically calculated and updated.

We currently use a simple algorithm that generates cubical and spherical SVs based on the shape and dimensions of the axis-aligned bounding box. Possible additional "smart" generation techniques include scaling the SV inverse proportionally to the object's relative distance to the viewpoint (in our case the camera) and introducing dynamic size constraints to avoid overlapping SVs.

Midas touch: If we let gaze automatically select objects, the "Midas touch" problem [18] occurs, frequently annoying users by unintentional system responses. In our interface gaze only chooses and highlights selection candidate objects that become selected only when an explicit gesture is made.

Staring: To overcome the "Midas touch" problem, some systems apply a dwelling time to trigger selection and thus implicitly impose strain onto the eyes. We have refrained from making users stare at spots on the display by saving the last selection candidate until the next valid object is hit by the eye gaze.

Cursor drift: Visualizing estimated gaze direction with an eye cursor can be distracting. Users tend to automatically look at the cursor and follow it, being misled by the illusion of perfect tracking. Due to aforementioned inevitable tracking inaccuracies, the offset between the real and measured fixation point gets gradually amplified, causing the eye cursor to drift. The drifting problem can be improved by precise calibration adjusting the system to individual users and by selectively turning off eye cursor visualization, for which we provide an easy keyboard shortcut.

Figure 7 visualizes the bounding volumes and the eye cursor. These auxiliary GUI elements are hidden by default but can be selectively turned on for adjustments and evaluation.

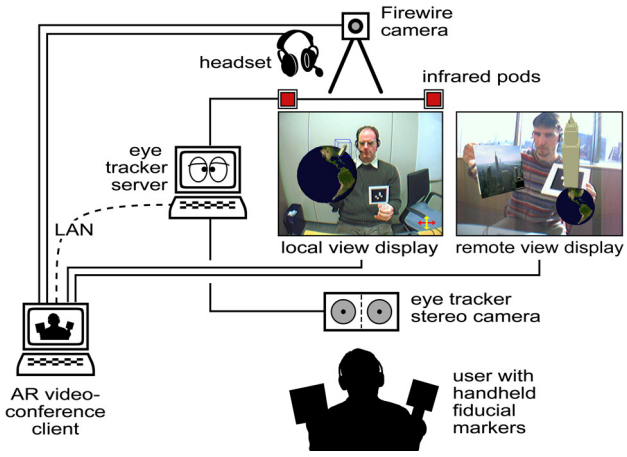

Figure 8. Hardware environment of the AR videoconference user playing the student role

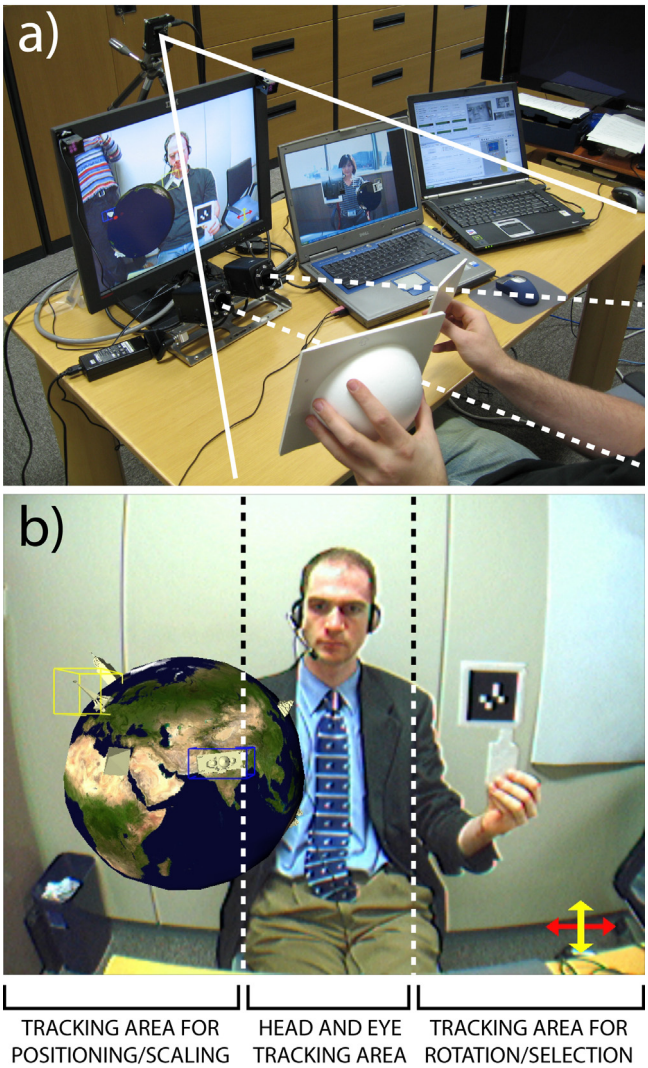

Figure 9. Ergonomic workspace arrangement. Camera angle is marked with a solid line in image a). Image b) shows the video image seen by the camera. The eye tracker's working volume is illustrated with a dotted line in both images

7 Hardware and Software Environment

For the implementation of our student-teacher collaboration scenario, we have built an asymmetric hardware setup. The student's hardware environment is depicted in Figure 8. The setup is divided into two major components: the AR videoconferencing client and the eye tracker server. The teacher's hardware setup is similar, however, it does not include an eye tracker.

Both AR videoconferencing clients contain a PC with two monitors displaying the local and remote workspace. At both sides a PointGrey Research Firefly® color camera is mounted on top the screen showing the local workspace. This display renders a scene-oriented [14] live video stream of the local work environment augmented with virtual objects. The video background is horizontally flipped, turning the local view's display into an augmented mirror (for a study on user preference see [2]). The cameras can be rotated and moved around the user without the need for recalibrating pose and eye tracking.

For pose tracking we used the ARToolKit [3] tracking library, which is a low-cost optical marker tracking system requiring only a webcam and fiducial markers printed on paper. For eye tracking we use the faceLAB™ 4 [11] system. This high-speed, vision-based eye tracker uses a stereo camera system and off-axis infrared illumination to do dark pupil tracking combined with 6DOF head tracking at a 60Hz data rate. The tracker is non-intrusive and untethered; no contact is required with the user allowing them to move around within a defined working volume. The eye tracker has its own server, and sends data to the videoconference client via LAN. Eye tracking and pose tracking integration into the AR videoconferencing client was realized by extending the *Studierstube* AR framework [30].

7.1 Workspace arrangement

As both pose and eye tracking are vision-based, occlusion is an apparent problem. In particular markers and hands may cast a shadow on the face if they are near the infrared pods, obstructing head tracking. Fiducial marker tracking also demands that the marker patterns are visible all times. To satisfy these constraints, an ergonomic physical and virtual workspace has been designed.

As Figure 9 illustrates, we divide our system's working volume into three parts: a marker tracking area for positioning / scaling the virtual globe, head and eye tracking area, and a marker tracking area for rotating the globe and selection gestures. The faceLAB™ system's cameras are zoomed in to monitor facial features, constraining their field of view between 10 and 30 degrees based on tracking accuracy. The 4mm lens of the FireWire camera following hand movements features a wide field of view that enables the tracking of markers in a larger area encompassing the user's desktop. The latter arrangement not only matches the distribution of the hardware components but the natural spatial structure of body parts (left hand – head in the middle – right hand) as well.

The eye tracker software features a fully integrated 3D world model that allows the detection of eye gaze intersections with preregistered physical and virtual objects in the surrounding environment. While this is useful for static objects measured prior to calibration, we had to rely on our own raypicking method to handle intersections with dynamic objects in AR environments.

8 Pilot Evaluation

We have conducted a pilot usability study to gain initial experience both in numerical and anecdotal form how smooth and natural the cascade of bimanual and eye gaze-based interaction appears to users. Our evaluation application was the architecture

scenario described in previous sections. It is important to note that we did not intend to create an actual remote collaboration tool for architects but rather a challenging selection task with interesting content that would appeal to an evaluation group with diverse cultural and educational background. Our test group consisted of 9 participants: 3 women and 6 men ranging in age from 26 to 35 years. 2 of them were Asians, the rest of the group were Caucasians, which determined their facial feature tracking profile. Two users wore glasses and one user had slight strabismus. None of them had ever worked with tangible user interfaces or an eye tracker, and they had only modest experience with videoconferencing or video chat tools such as Skype or Net-meeting.

Prior to the evaluation session each user received a short training on tangible interaction gestures and calibrated the head and eye tracking module with a technician. Then participants were asked to play the student role in the local object selection variant of the student-teacher application scenario described in Section 5.1. An experienced user acted as the teacher in a different room, and moderated a conversation about widely known architectural landmarks through the AR videoconferencing client. Whenever a student wanted to discuss a building or received explicit instructions from the teacher, she rotated the virtual globe to find the building's small-scale virtual model over the corresponding geographical spot and selected it with a pointing method.

We organized two evaluation sessions. In both sessions the student and the teacher used both hands to manipulate objects with tangible markers but in the first session the student relied on our eye tracker-based pointing technique while in the second session a desktop mouse replaced the eye pointer, requiring periodical switching between marker and mouse. During the first session eye pointer visualization could be turned on/off by the user any time. The total length of the sessions was 15-20 minutes. After the sessions participants filled out a questionnaire, and were interviewed about their answers afterwards. The questionnaires used a Likert scale, asking users to rate the following 9 categories from -3 (negative opinion) to +3 (positive opinion):

1. Overall learning curve for tangible interaction
2. Learning curve for globe rotation
3. Learning curve for building selection technique cascading hand and eye movement
4. Learning curve for making selection candidates with eyes
5. Strain imposed on eyes and body
6. Eye tracker accuracy
7. Preference of eye pointer over mouse pointer in combination with bimanual tangible interaction
8. Usefulness of eye cursor
9. Perception of a shared workspace with remote communication partner

We captured videos of all users in all sessions. We simultaneously recorded the student and teacher view and the eye tracker software's window. The videos were manually annotated and analyzed to extract typical execution times for the following interaction elements:

1. Searching for a country by rotating globe with marker
2. Locking / unlocking globe rotation
3. Making a selection with eye and manual gesture after finding a selection candidate on screen
4. Making a selection with mouse after finding a selection candidate on screen

Figure 10 shows the mean and standard error values of evaluation scores and interaction times for each category.

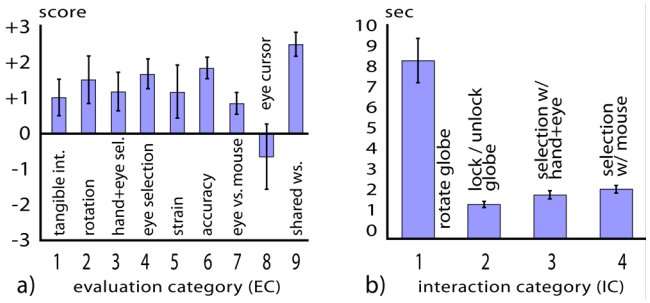

Figure 10. Evaluation results: a) evaluation scores, b) interaction times. Error bars indicate the standard error of the mean

8.1 Outcomes

Mean scores and standard error for evaluation categories (EC) 1-3 indicate that marker-based interaction has been found generally easy and intuitive, although some users had an apparent learning curve problem. All users thought that overlaying virtual objects on top of the videoconferencing window enriched remote collaboration. Without exceptions users quickly learned marker associations, however, the assessment of individual components of two-handed gestures indicate that bimanual synchronization requires considerable time to master. Times for interaction categories (IC) 1-4 show that most time was spent on rotating the globe to "browse" for interesting buildings, while the average execution time of individual gestures was short.

The scores of EC 4-6 suggest that eye tracker-based interaction is smooth and accurate enough to be suitable for dynamic selection tasks in AR environments. Interviews revealed that users found our eye-based interface an easy and natural way to select dynamic objects even with fine details, which justifies our efforts with selection volumes and candidates. Without exception users were disturbed by broken eye contact with the remote collaborator when putting down the markers and grabbing the mouse in the second session. Switching between manual interaction devices clearly interrupts the natural interaction flow and requires spatial and mental reinitialization of the device's reference frame.

The comparison of IC 3 and 4 suggest that making a selection with a manual gesture supported by eye gaze was slightly faster than the combination with the mouse, although there was no significant difference. Scores for EC 7 indicate slight preference for the cascaded interaction technique relying on eye tracking over the mouse-based scenario, however, the mean score is lower than expected. While we acknowledge that it is difficult to compete with the fluency users had acquired with mouse-based interaction through years, usability problems with the eye tracker also significantly contributed to the decreased performance of our eye-based interface. Our biggest problem was the limited working volume causing untrained users to occlude their face or accidentally exit the tracking area. Users tended to choose a "safe", stiff body posture for stable interaction. Some participants also reported fatigue after holding the markers in hand after a longer session. These claims are supported by the significantly large standard error value of EC 5. A possible workaround for the occlusion problem may be the display of visual system warnings to make users aware of the interruptions in eye tracking.

Participants were strongly polarized by their opinion about the use of the eye pointer. We provided a keyboard shortcut to selectively turn the cursor on/off in order to allow the choice of

personal preference. Users for whom eye tracking and working volume problems were not significant preferred hiding it, while those who needed constant reassurance that their gaze was still being tracked preferred turning it on.

9 Conclusion and Future Work

We have implemented an AR videoconferencing system supporting collaboration between remote workspaces and a novel interaction technique cascading fiducial marker-based bimanual tangible interaction and eye tracking. During our pilot evaluation sessions we found that our system and interaction technique are highly suitable for rich, natural and intuitive collaborative object manipulation over great distances.

An interesting addition to our system would be using a real ray such as a laser pointer to illuminate selected physical objects in the remote workspace similarly to the work by Kurata et al. [21]. Additional future work includes extending the current dual-user system and user interface to support several users, carrying out a more comprehensive user evaluation, and experimenting with techniques requiring less accurate eye tracking but offering a larger working volume.

10 Acknowledgements

This research was supported by the Encouragement of Young Scientists Grant (FY2005-FY2007) from the Japan Society for the Promotion of Science (JSPS), an NII Joint Research Grant with the University of Tokyo (FY2005), and the Austrian Science Fund FWF (contract No. Y193). The authors would like to thank Arturo Nakasone for helpful discussions, and all user study participants.

Reference

[1] Asai, K., Osawa, N., Takahashi, H., Sugimoto, Y.Y., Yamazaki, S., Samejima, M., and Tanimae, T., Eye Mark Pointer in Immersive Projection Display. In *Proc. of Virtual Reality 2000*, New Brunswick, NJ, USA, 2000, pp. 125-134.

[2] Barakonyi, I., Fahmy, T., and Schmalstieg, D., Remote Collaboration Using Augmented Reality Videoconferencing. In *Proc. of Graphics Interface 2004 (GI'04)*, London, ON, Canada, May 17-19, 2004, pp. 89-96.

[3] Billinghurst, M., and Kato, H., Real World Teleconferencing. In *Proc. of the Conference on Human Factors in Computing Systems (CHI'99)*, Pittsburgh, USA, May 15th-20th, 1999.

[4] Billinghurst, M., Kato, H. and Poupyrev, I., Collaboration with Tangible Augmented Reality Interfaces. Presented at HCI International'01, August 5-10, 2001, New Orleans, LA, USA.

[5] Bolt, R.A. and Herranz, E., Two-Handed Gesture in Multi-Modal Natural Dialog, In *Proc. of the Symposium of User Interface Software and Technology (UIST'92)*, Monterey, CA, USA, 1992, pp. 7-14.

[6] Bowman, D., Kruiff, E., LaViola, J., and Poupyrev, I., *3D User Interfaces: Theory and Practice*, Addison-Wesley, 2004.

[7] Ciger, J., Herbelin, B., and Thalmann, D., Evaluation of Gaze Tracking Technology for Social Interaction in Virtual Environments. In *Proc. of the 2nd Workshop on Modeling and Motion Capture Techniques for Virtual Environments (CAPTECH'04)*, Zermatt, Switzerland, 2004.

[8] Curatu, C., Hua, H., and Rolland, J., Projection-based Head-mounted Display with Eye Tracking Capabilities. In *Proc. of SPIE*, V. 5875, 2005, pp. 141-149.

[9] Duchowski, A. T., and Vertegaal, R., Course 05: Eye-based Interaction in Graphical Systems: Theory & Practice. In *SIGGRAPH 2000 Course Notes*, New York, NY, USA, 2000.

[10] EyeCatcher project page: http://www.fit.fraunhofer.de/projekte/eyecatcher/

[11] faceLAB™ product page: http://www.seeingmachines.com/

[12] Fitzmaurice, G., Ishii, H., and Buxton, W., Bricks: Laying the Foundations for Graspable User Interfaces. In *Proc. of the Conference on Human Factors in Computing Systems (CHI'95)*, Denver, CO, USA, 1995, pp. 442-449.

[13] Forlines, C., Balakrishnan, R., Beardsley, P.A., van Baar, J., and Raskar, R., Zoom-and-pick: Facilitating Visual Zooming and Precision Pointing with Interactive Handheld Projectors, In *Proc. of the Symposium on User Interface Software and Technology (UIST'05)*, Seattle, WA, USA, 2005, pp. 73-82.

[14] Fussell, S. R., Setlock, L. D., and Kraut, R. E., Effects of Head-mounted and Scene-oriented Video Systems on Remote Collaboration on Physical Tasks. In *Proc. of the Conference on Human Factors in Computing Systems (CHI'03)*, 2003, Fort Lauderdale, FL, USA, pp.513-520.

[15] Guiard, Y., Asymmetric Division of Labor in Human Skilled Bimanual Action: The Kinematic Chain as a Model. In *Journal of Motor Behavior*, 19(4), 1987, pp. 486-517.

[16] Hinckley, K., Pausch, R., Proffitt, D., and Kassell, N. Two-handed Virtual Manipulation. In *ACM Transactions on Computer-Human Interaction (TOCHI)*, 5(3), 1998, pp. 260-302.

[17] Ishii, H., and Ullmer, B., Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. In *Proc. of Conference on Human Factors in Computing Systems (CHI'97)*, Atlanta, GA, USA, 1997, pp. 234-241.

[18] Jacob, R.J.K., The Use of Eye Movements in Human-Computer Interaction Techniques: What you Look at is What you Get. In *ACM Transactions on Information Systems*, 9(3), 1991, pp. 152-169.

[19] Kabbash, P., Buxton, W. and Sellen, A., Two-handed Input in a Compound Task. In *Proc. of the Conference on Human Factors in Computing Systems (CHI'94)*, Boston, MA, USA, 1994, pp. 417-423.

[20] Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Feiner, S., and Cohen, P., Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality. In *Proc. of International Conference on Multimodal Interfaces (ICMI 2003)*, Vancouver, Canada, 2003, pp. 12-19.

[21] Kurata, T., Sakata, N., Kourogi, M., Kuzuoka, H., and Billinghurst, M., Remote Collaboration Using a Shoulder-worn Active Camera/Laser, In *Proc. of the International Symposium on Wearable Computers (ISWC'04)*, Arlington, VA, USA, 2004, pp. 62-69.

[22] Lee, G.A., Nelles, C., Billinghurst, M., and Kim, G.J., Immersive Authoring of Tangible Augmented Reality Applications. In *Proc. of International Symposium on Mixed and Augmented Reality 2004 (ISMAR 2004)*, Arlington, VA, USA, 2004, pp. 172-181.

[23] Milgram, P., Takemura, H., Utsumi, A. and Kishino, F., Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum. In *SPIE Vol. 2351, Telemanipulator and Telepresence Technologies*, 1994, pp. 282-292.

[24] Novak, V., Sandor, C., and Klinker, G., An AR Workbench for Experimenting with Attentive User Interfaces, In *Proc. of International Symposium on Mixed and Augmented Reality (ISMAR'04)*, Arlington, VA, USA, 2004, pp. 284-285.

[25] Pfeiffer, T., Latoschik, M.E., Resolving Object References in Multimodal Dialogues for Immersive Virtual Environments. In *Proc. of Virtual Reality (VR'04)*, Chicago, IL, USA, 2004, pp. 35-42.

[26] Pierce, J.S., Stearns, B.C., and Pausch R., Voodoo Dolls: Seamless Interaction at Multiple Scales in Virtual Environments. In *Proc. of the Symposium on Interactive 3D Graphics*, Atlanta, GA, USA, 1999, pp. 141-145.

[27] Prince, S.J.D., Cheok, A.D., Farbiz, F., Williamson, T., Johnson, N., Billinghurst, M. and Kato, H., Real-Time 3D Interaction for Augmented and Virtual Reality. In *Proc. of SIGGRAPH'02 Sketches and Applications*, San Antonio, USA, July 21-26, 2002.

[28] Regenbrecht, H., Lum, T., Kohler, P., Ott, C., Wagner, M., Wilke, W., and Mueller, E., Using Augmented Virtuality for Remote Collaboration. In *Presence: Teleoperators and Virtual Environments*, 13(3), 2004, pp. 338-354.

[29] Salvucci, D.D., Goldberg, J.H., Identifying Fixations and Saccades in Eye-tracking Protocols. In *Proc. of the Symposium on Eye Tracking Research and Applications*, Palm Beach Gardens, FL, USA, 2000, pp. 71-78.

[30] Schmalstieg, D., Fuhrmann, A., Hesina, G., Szalavári, Zs., Encarnação, M., Gervautz, M., Purgathofer, W., The Studierstube Augmented Reality Project, In *PRESENCE - Teleoperators and Virtual Environments*, MIT Press, 2002.

[31] Sellen, A., Buxton, W., and Arnott, J., Using Spatial Cues to Improve Videoconferencing. In *Proc. of the Conference on Human Factors in Computing Systems (CHI'92)*, Monterey, CA, USA, 1992, pp. 651–652.

[32] Tanriverdi, V., and Jacob, R.J.K., Interacting With Eye Movements In Virtual Environments, In *Proc. of the Conference on Human Factors in Computing Systems (CHI'00)*, The Hague, The Netherlands, 2000, pp. 265-272.

[33] Velichkovsky, B.M., Dornhoefer, S.M., Pannasch, S., and Unema, P.J.A., Visual Fixations and Level of Attentional Processing. In *Proc. of the International Conference of Eye Tracking Research and Applications*, Palm Beach Gardens, FL, USA, 2001.

[34] Vertegaal, R., Weevers, I., Sohn, C., and Cheung, C., GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction. In *Proc. of Conference on Human Factors in Computing Systems (CHI'03)*, Fort Lauderdale, FL, USA, 2003, pp. 521-528.

[35] Wilder, J., Hung, G., Tremaine, M.M., and Kaur, M., Eye Tracking in Virtual Environments. In *Handbook of Virtual Environments*, Lawrence Erlbaum Associates, 2002.

[36] Zhai, S., Morimoto, C., and Ihde, S., Manual and Gaze Input Cascaded (MAGIC) Pointing, In *Proc. of the Conference on Human Factors in Computing Systems (CHI'99)*, Pittsburgh, PA, USA, 1999, pp. 246-253.