To appear in: *Eye Tracking Research & Applications Symposium 2004, San Antonio, TX, March 22-24, 2004, ACM*

Visual Deictic Reference in a Collaborative Virtual Environment

Andrew T. Duchowski* Computer Science
Nathan Cournia* Computer Science
Brian Cumming* Computer Science
Daniel McCallum* Computer Science
Anand Gramopadhye† Industrial Engineering
Joel Greenstein† Industrial Engineering
Sajay Sadasivan † Industrial Engineering
Richard A. Tyrrell ‡ Psychology
Clemson University

Abstract

This paper evaluates the use of Visual Deictic Reference (VDR) in Collaborative Virtual Environments (CVEs). A simple CVE capable of hosting two (or more) participants simultaneously immersed in the same virtual environment is used as the testbed. One participant's VDR, obtained by tracking the participant's gaze, is projected to co-participants' environments in real-time as a colored lightspot. We compare the VDR lightspot when it is eye-slaved to when it is head-slaved and show that an eye-slaved VDR helps disambiguate the deictic point of reference, especially during conditions when the user's line of sight is decoupled from their head direction.

1 Motivation

The lack of eye contact in tele-communication systems has long been recognized. Ishii and Kobayashi [1992] showed the importance of eye contact in shared drawing and conversation systems while Garau et al. [2001] demonstrated the importance of eye gaze in humanoid avatars representing people engaged in conversation. In dyadic conversation experiments, where the avatar conditions differed only in their treatment of eye gaze, Garau et al. showed that the inferred-gaze avatar significantly outperformed the random-gaze model and also outperformed the audio-only avatar on several response measures. The inferred-gaze avatar's head movement was determined by tracking of the remote participant and eye movement was inferred from conversational turn taking. Clearly, if eye tracking were available, the avatar's eye movements could be replicated directly without needing to infer their direction.

The problem of deictic reference in Collaborative Virtual Environments (CVEs) has been demonstrated by several authors. Hindmarsh et al. [2000] eloquently describe the problem as that of users having difficulties in understanding others' perspectives. More specifically, the authors report that, in general, individuals could not determine what a co-participant was referring to, where, and at what, they were looking or pointing. The authors further argue that a visual deictic reference is one of the critical and foundational elements of collaborative work—i.e., the reference to, and discussion of, objects and artifacts. The authors demonstrate the problem by reporting users' (Sarah and Karen's) comments during a furniture rearrangement task (see Figure 1). Clearly, a visual aid symbolizing what a participant in a CVE is looking at would be beneficial.

Conveyance of gaze direction in multi-party systems has been studied for some time. Vertegaal [1999] showed the benefits of conveyance of gaze direction in a cooperative document sharing and multi-party conversation system. In Vertegaal's GAZE Groupware System, gaze direction of multiple participants is symbolized

```
S:  You know this desk-thing?
K:  Yeah?
S:  Can you see–| what I'm pointing at now?
    ((K Turns to Find S))
K:  Er I can't see you, but [I think–
S:  [It's like a desk-thing.
K:  Er–where've you gone?  [heh heh heh
S:  [Erm, where are you?
K:  I've– th– I can see
S:  Tur– (.) oh, oh yeah.  You're near the lamp,
    yeah?
K:  Yeah.
S:  And then, yeah turn around right. (.) and then
    it's like (.) I'm pointing at it now, but I
    don't know if you can see what [I'm pointing
    at?
K:  [Right yeah I can see.
```

Figure 1: Example of lack of visual deictic reference in a CVE ([]'s indicate overlapping utterances; (.) indicates a short pause in talk). From Hindmarsh et al. [2000].

by an eye-tracked animated color dot over shared documents. We adopt this "laser pointer metaphor"[1] and implement a form of visual deictic reference similar to that of Vertegaal's in our CVE. We display each participant's deictic reference point as a small red dot in the environment. There is no restriction on dot colors; these can be changed per participant and can be embellished with other information such as participant names. However, unlike Vertegaal, we can only track one participant's gaze at a time (only one of our Head Mounted Displays is equipped with an eye tracker). Therefore, we explore head-slaved deictic reference as an alternative to an eye-slaved representation. We hypothesize that either head- or eye-slaved representation will better identify the reference target than no form of visual deictic reference.

2 Methodology

2.1 Apparatus

Each of our VEs is driven by a 1.5 GHz dual-processor PC running Red Hat Linux (v8.0, kernel v2.4.20) equipped with 1 G RAM and an NVidia GeForce4 Ti 4600 graphics card. Multi-modal hardware components in our immersive VR system include a binocular eye tracker mounted within a Virtual Research V8 Head Mounted Display (HMD). The V8 HMD offers 640 $\times$ 480 pixel resolution per eye with individual left and right eye feeds, giving the user a field of view of 75.3° $\times$ 58.4° visual angle [Watson et al. 1997].

*{andrewd | acnatha | brcummi | dcmccal}@vr.clemson.edu
†{agramop | joel.greenstein | ssadasi}@ces.clemson.edu
‡tyrrell@clemson.edu

[1] Vertegaal referred to the lightspot as the "miner's helmet"—we choose the laser pointer since a miner's helmet suggests the lightspot is slaved only to head direction.

HMD position and orientation tracking is provided by each display device's own Ascension 6 Degree-Of-Freedom (6DOF) Flock Of Birds (FOB). A 6DOF tracked, hand-held mouse provides either a means of representation of a virtual tool for the user in the immersive environments or a means of navigation. As a navigation tool, participants press the left or middle mouse button to move forward or backward, respectively, along the current line of sight. In the present experiment, the mouse was configured as a navigation tool.

The eye tracker is a video-based, corneal reflection unit, built jointly by Virtual Research and ISCAN. Each of the binocular video eye trackers is composed of a miniature camera and infrared light sources, with the dual optics assemblies connected to a dedicated PC. The ISCAN RK-726PCI High Resolution Pupil/Corneal Reflection Processor uses corneal reflections (first Purkinje images) of infra-red LEDs mounted within the helmet to measure eye movements. Figure 2 shows the dual cameras and infra-red LEDs of the binocular assembly. The HMD is shown in Figure 2 (inset), with the FOB sensor just visible on top of the helmet. Mounted below

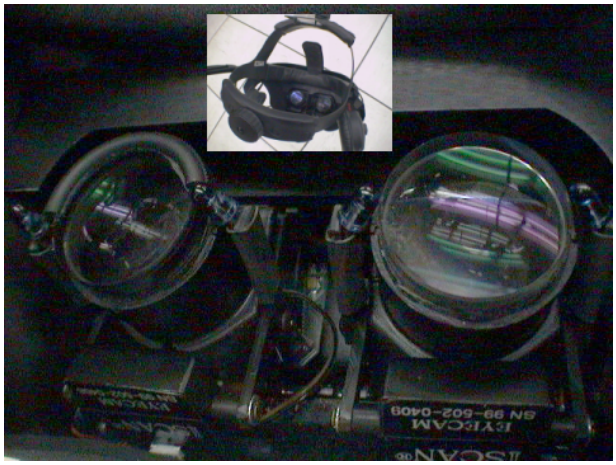


Figure 2: Binocular eye tracker optics (w/HMD inset).

the HMD lenses, the eye imaging cameras peer upwards through a hole cut into the lens stem, capturing images of the eyes reflected by a dichroic mirror placed behind the HMD lenses. The processor typically operates at a sample rate of 60 Hz, however while in binocular mode our measured sample rate decreases to 30 Hz. The user's eye position is determined with an accuracy of approximately 0.3 degrees over a ± 20 degree horizontal and vertical range using the pupil/corneal reflection difference. The maximum spatial resolution of the calculated Point Of Regard (POR) provided by the tracker is 512×512 pixels per eye. Using the vendor's proprietary software and hardware, the PC calculates the user's real-time POR from the video eye images. In the current VR configuration, the eye tracker is treated as a black box delivering real-time eye movement coordinates over a 19.2 Kbaud RS-232 serial connection, and can be considered as an ordinary positional tracking device.

2.2 The Collaborative Virtual Environment

To extend a single-user virtual environment to two participants, our main concern is the implementation of a shared state repository [Capps 2001]. Our CVE is limited to only two fully immersed participants due to the availability of two HMDs in our laboratory. Furthermore, our two participants are co-located in the same lab, about 6 feet apart. The physical arrangement of our laboratory therefore precludes implementation of a geographically disjoint CVE.

Our shared state repository model is realized by a client-server architecture (star topology), where many clients can connect to the central server. The server contains the only truly valid copy of the world and all world logic runs on the server. Clients enqueue user/world input, and then send these input requests to the server. The server then processes each input, and pushes a new world state to each client. To compensate for the delay between sending new input to the server and receiving an updated world state, each client attempts to predict future world states (e.g., via dead-reckoning).

Multithreaded clients manage the FOB and eye tracker devices only when these devices are active (this allows interaction via mouse and keyboard, e.g., for remote non-immersed participants). The client can be thought of as a specialized dumb terminal since the only calculations it performs (in addition to marshaling user input) is how best to represent the world to the user.

The representation of an avatar's eye direction from the tracked Point Of Regard (POR) of the avatar's human counterpart is relatively simple. Since POR data gives coordinates of the user's gaze on the viewing planes in front of the human's eyes, it is a relatively simple matter to derive the orientation of the eye given an assumed center of rotation. Furthermore, since our eye tracker returns $(0,0)$ during blinks, we also model the avatar's eye blinks by texture mapping a closed eyelid during blinks, thus providing a compact representation of an "anthropomorphic humanoid" [Luciano et al. 2001].

2.3 Immersion in the CVE

Current Collaborative Virtual Environment systems can be described as either first- or second-generation. Ragusa and Bochenek [2001] classify first-generation systems as those including a Head-Mounted Display or a Binocular Omni-Orientation Monitor (BOOM). Second-generation systems employ large projectors and stereoscopic glasses, and include such well-known systems as the Cave Automatic Virtual Environment (CAVE), the Powerwall system (called Immersive Work-Walls by Fakespace Systems), and the ImmersaDesk. Our CVE falls under the first-generation classification.

The virtual environment used for deictic reference experiments is a simple room populated with road sign targets. The CVE is shown in Figure 3 with both referrer's ("trainer's") and referee's ("trainee's") viewpoints. In Figure 3 the referrer is looking at the "Pilot Mountain State Park" sign. Visual feedback to the referrer is given as a transparent eye symbol and is visible in the referrer's view shown in Figure 3(a). In Figure 3(b) the referee sees the referrer's avatar along with the referrer's visual deictic reference lightspot. Figure 3(a) also shows the referee's lightspot, in this case head-slaved since the referee's HMD is not equipped with an eye tracker, projected just left of the "Do Not Enter" sign. Our CVE thus permits display of co-participants' VDRs.

2.4 Participants

Eight participants (7 male, 1 female) volunteered to take part in the experiment. All participants were undergraduate or graduate students. All participants were knowledgeable about computer graphics and half were experienced virtual reality users and/or developers.

2.5 Experimental Design

A $4 \times 3 \times 4$ factorial design was used with independent variables of deictic reference, target density, and selection modality. Deictic reference refers to the visibility of the projected lightspot over targets selected by the referrer. Thus this variable pertains to what the referee saw. These ranged from: "head only", "head+eye", "head+hdot", "head+eye+edot". "Head only" means that the referee could only see the referrer's avatar's head rotations. The "head+eye" condition means that the avatar's head and eye rotations were made visible to the referee. In both of these cases,

(a) Referrer's view (with referee's lightspot visible). (b) Referee's view (with referrer's lightspot visible).

Figure 3: Deictic reference in the CVE.

no deictic reference lightspot was shown. "Head+hdot" means that the avatar's head rotations were visible and that the deictic reference lightspot was slaved to the avatar's head direction. "Head+eye+edot" means both the avatar's head and eye rotations were visible and the visible deictic reference lightspot was slaved to the referrer's gaze direction.

Target density refers to the target distribution on the walls of the environment. To facilitate the referee's task of verbally naming the target selected by the referrer, we chose easily identifiable targets (road signs) texture mapped onto the walls of a simple environment. We varied the referrer's pointing task (selection of targets) by choosing either a large target (the wall) vs. small targets (the signs) on either sparsely or densely populated walls. On densely populated walls, the signs cover 27.2% of the wall area, on average, while on sparsely populated walls they cover 14.7%, on average. The textured environment is shown in Figure 4, where half the number of walls are densely populated.

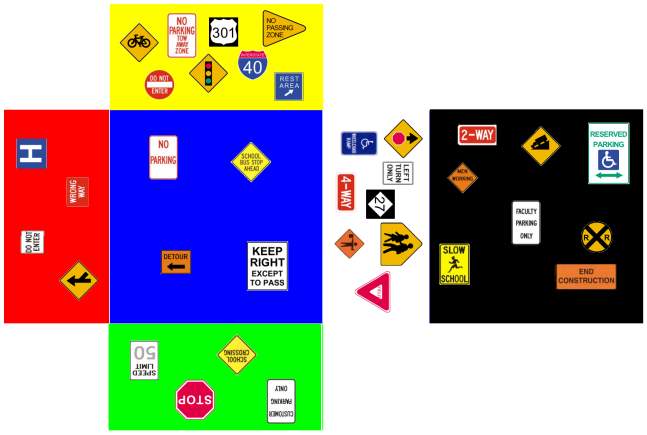


Figure 4: Target stimuli.

Selection modality refers to the separation between the referrer's head and gaze directions. The motivation here was to control the ambiguity between head and gaze direction for the referee ensuring that the referrer's gaze would, under some conditions, be fairly far away from head direction so as not make the referee's task obvious. To prevent the referrer biasing the condition, the referrer was never aware whether the referee would see the referrer's light spot, or whether the light spot would be slaved to either head or gaze. As far as the referrer was concerned, the referee would always see the referrer's eye-slaved lightspot. To provide this functionality, for each target (wall or sign), the referrer was given two large crosshairs indicating the required orientation of both the head and eyes. For example, when instructed to select the stop sign on the green wall, the referrer was asked to center their head at the sign center while fixating their gaze 19.64° away from the head direction. Separation between head and gaze direction was thus varied between 0°, ∼10°, ∼15°, and ∼20°.

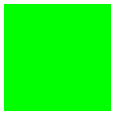
A within-subjects design was used to test deictic reference. To counteract ordering effects, deictic reference order was chosen from a 4 × 4 Latin square (repeated for each group of 4 subjects). Three targets for each of the target density conditions were given to subjects in an initially randomized order. The 9 target combinations are shown in Table 1, arranged in groups of three different target types: wall target, signs on sparsely populated walls, and signs on densely populated walls. Each referee subject responded to the same randomized sequence of questions (9 total) but under different deictic reference conditions resulting in 36 trials per subject. In all, data was recorded from 288 trials, with 72 trials recorded for each form of deictic reference (9 trials per deictic reference variant for each of 8 subjects).

2.6 Procedure

Each experimental trial was fairly simple: in each trial the referee was asked to verbally identify the target selected by the referrer within the virtual environment. Although the referrer knew which target to select, s/he was not aware of the referee's experimental condition. As far as the referrer knew, the referee could always see the referrer's gaze. The referee was then asked to respond to the question "Which wall am I looking at?" or "Which sign am I looking at?" based on either wall or sign target condition (see above).

At the beginning of each trial, participants were introduced to the two VR devices and environment they were to be immersed in (along with the avatar representations they would be seeing). The referee was told that each of the walls was a valid target, i.e., ceiling, floor, and 4 side walls. Training was given on the use of the 6DOF mouse which provided navigation within the environment (forward or backward movement along the line of sight, a typical "flying" navigation strategy employed in VR). The referee was

Table 1: Target × selection modality, with separation between gaze and head direction (visual angle) given below each sign.

Wall Targets			Sign Targets on Sparse Walls			Sign Targets on Dense Walls		
								
15.86°	11.50°	0°	19.64°	16.09°	0°	16.50°	14.39°	0°

also told about the avatar's expressiveness: the avatar body rotates, sometimes the eyes rotate as well. The referee was free to "fly" up to the referrer avatar to inspect the avatar's body position and head/eye orientation. A short 5-point eye tracker calibration sequence was performed for the referrer before display of the VE.

3 Results

Repeated-measures 1-way ANOVA was performed to examine the main effect of Visual Deictic Reference (VDR). The mean number of guesses is given in Table 2 and depicted in Figure 5.[2] The effect

Table 2: Results: mean number of guesses.

Condition	Mean
A: head rotation only	1.71
B: head & eye rotation	1.26
C: head rotation & head-slaved lightspot	1.53
D: head & eye rotation & eye-slaved lightspot	1.14

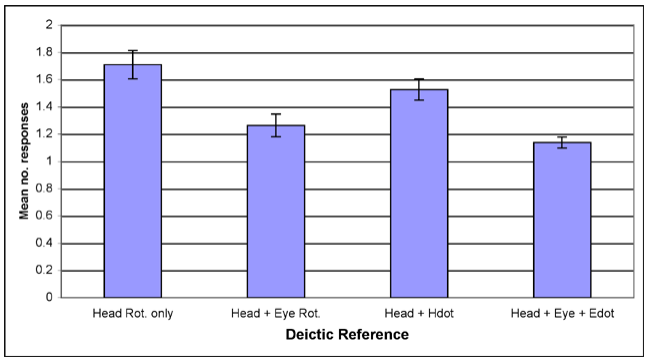


Figure 5: Effect of deictic reference (error bars represent ± one standard error of the mean).

of VDR is significant, F(3,21) = 9.96, p < .001 ($\eta^2 = .587$). Pairwise comparisons reveal that condition A is significantly different from both B (p = .017) and D (p = .004) but not C (p = .174). B is marginally different from C (p = .057) but is not different from D (p = .196). C is significantly different from D (p = .001).

Note that the above pairwise comparisons are uncorrected for the multiple tests being performed (i.e., overall $\alpha > .05$). Applying Bonferroni correction (more conservative), the only significant pairwise comparisons are between A and D (p = .022) and between C and D (p = .009).

[2]We chose not to record the time to guess since the task was unpaced.

Condition D (eye-slaved lightspot) appears to provide the best performance, although it is not significantly better than B (head and eye rotation).

A 1-way ANOVA of mean responses pooled per given deictic reference condition suggests there is a significant effect of selection modality (see Table 1) under each condition except B (when head and eye rotations are visible but no VDR lightspot is displayed). Mean number of guesses per selection modality under each VDR condition is shown in Figure 6.

When only the avatar's head rotation is observable and no VDR lightspot is shown (condition A), performance degrades significantly when direction of gaze and head differs by 16.5°, F(8,63) = 19.10, p < .001.

When the visible VDR lightspot is head-slaved (condition C), performance degrades significantly when gaze and head direction differ by 14.39° or 16.5°, F(8,63) = 27.63, p < .001.

When the visible VDR lightspot is eye-slaved (condition D), performance degrades significantly when gaze and head direction differ by 19.64°, F(8,63) = 4.06, p < .001.

4 Discussion

Performance results indicate that display of the VDR lightspot is quite effective, especially when it is coupled with gaze direction. The head-slaved VDR is clearly less effective. It is interesting to note no significant difference between the avatar's head rotation and the head-slaved VDR lightspot. Similarly, there appears to be no difference between head and eye rotation and eye-slaved VDR lightspot. It appears, therefore, that conveyance of head and eye rotation is as effective as the presence of an eye-slaved lightspot.

Pooled per-condition analysis suggests that users may perceive a "deictic reference ambiguity" when head and gaze direction are sufficiently incongruent when gazing at densely populated targets. In the case when the avatar's eye rotations are not visible, users experienced significant performance degradation when gaze and head direction differed by 16.5°. Surprisingly, the addition of the head-slaved VDR lightspot degraded performance further when head and gaze incongruity decreased to 14.39°. The addition of the head-slaved VDR lightspot may have served to confuse users if they were expecting the VDR lightspot to be eye-slaved. It should be noted that both 14.39° and 16.5° selections were made at targets situated on densely populated walls within the environment. Users did not appear to experience deictic reference ambiguity when gaze and head incongruity increased to 19.64° since this target (the stop sign) was positioned on a low density wall.

The only time users experienced significant deictic reference ambiguity when head and gaze incongruity reached 19.64° was when the eye-slaved lightspot was visible on a low density wall. This may appear puzzling since it is in this case that one would expect the eye-slaved deictic reference to be of most benefit. This case, however, may expose a problem with the user making the target selections (the referrer in our experiment) and not the referee since

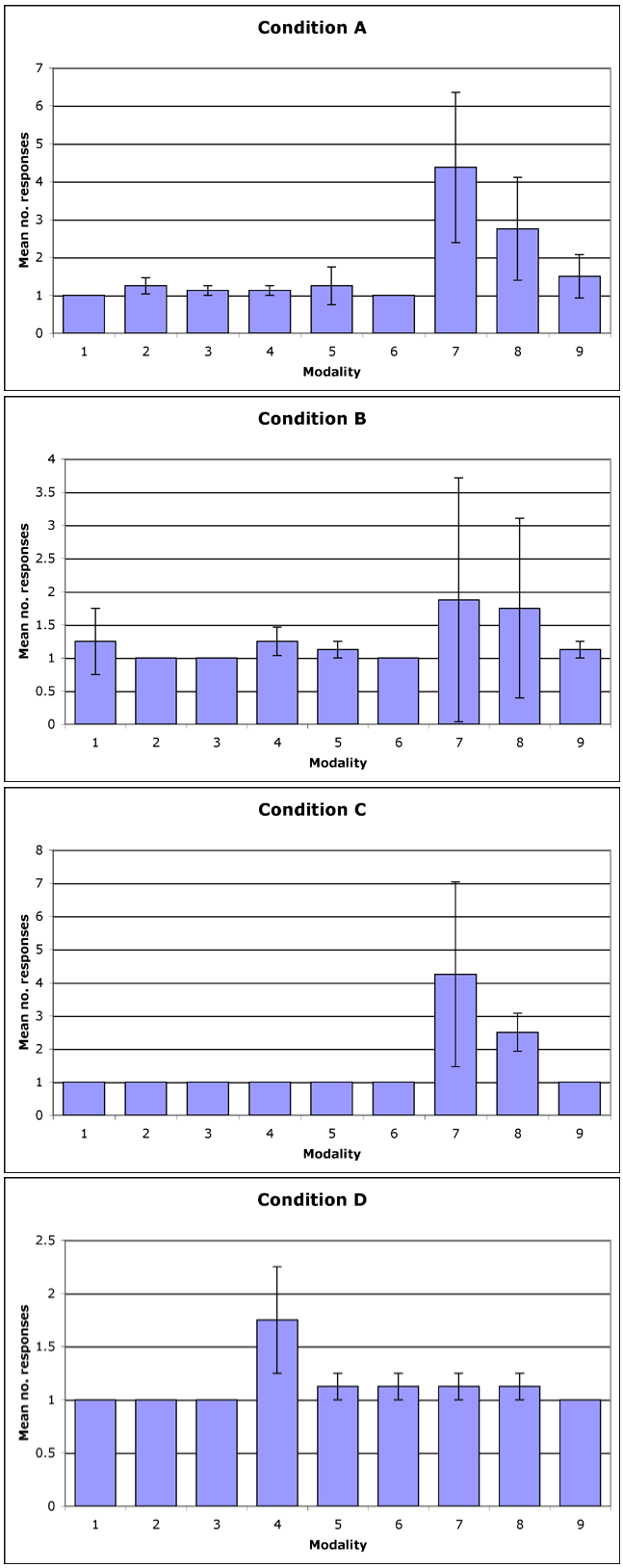


Figure 6: Per-condition modality effects.

it is fairly difficult to voluntarily move one's gaze 20° off-axis in a VR helmet. Previous research suggests that the eyes generally do not deviate more than 30° from the head-centric view direction [Barnes 1979; Watson et al. 1997; Murphy and Duchowski 2002].

5 Conclusion

We have presented a mechanism for the display of an explicit visual deictic reference in a collaborative virtual environment. Our lightspot representation is analogous to a laser pointer's dot projected from the avatar's head or eyes. We have shown that an eye-slaved lightspot is quite effective for disambiguating deictic (e.g., "look at this") reference, especially when it is coupled with gaze direction. The head-slaved VDR is clearly less effective although still provides some benefit than when no lightspot is shown at all. Alternatively, it may be almost as effective to simply provide sufficiently expressive avatars capable of accurately representing head and eye rotations.

Our research suggests that the expressibility (animated direction) of the avatar's torso, head, and eyes may be as suggestive as an explicit depiction of the projected point of regard. That is, participants may obtain as much information by looking at co-participants' body language as by looking at the dot projected by a virtual laser pointer affixed to the co-participant's head or eyes. Indeed, the avatars used in our study are capable of representing a variety of expressive poses, a few of which are shown in Figure 7. The top row shows the poses that we have successfully coupled to our 6DOF head tracker, e.g., when looking down, up and to the left, and up and to the right. The middle row shows a face-forward avatar (first column) which is the avatar's typical appearance in the CVE. The reason for the avatar's arm placement is due to the original avatar holding a weapon which we have deleted for our purposes. The avatar is constructed from 4 submodels (legs, torso, head, eyes) connected through a series of bones enabling us to independently rotate each part of the body. In addition to this simple skeletal system, the legs and torso submodels contain vertex animation data such as walk, taunt, kneel, etc., as shown in the two poses in the middle row of Figure 7. The last row reveals the gaming source for our inspiration: these poses reflect the avatar's various postures related to exhaustion and/or death. A full skeletal-based model including human joint positions and orientations (e.g., elbows, wrists, etc.) where each is tracked or at least inferred would offer an idealized model of avatar motion providing additional expressiveness.

In summary, we believe that an eye-slaved visual representation of deictic reference is beneficial and should be provided in collaborative environments. Representing the tracked head and eye movements of a user by expressive avatar body language also appears to provide visual benefit to the referee in the CVE. That is, expressive avatars capable of accurately rotating the head and eyes may offer an effective alternative to an eye tracked visual deictic reference. We conclude with the following recommendations:

1. Endow avatar representations with as much body expressiveness as possible, e.g., realistic torso, head, and eye rotations, depending on tracking capabilities.
2. For collaboration, provide a mechanism to allow disambiguation of deictic reference through visual means, e.g., a laser pointer metaphor appears to be effective.

In future studies we intend to examine other means for communicating deictic reference such as animation of eye gaze vectors drawn in 3D, as well as evaluating more expressive avatar animations.

Figure 7: Avatar.

6 Acknowledgments

This work was supported in part by NSF ITR award # 0217600 and NASA Ames task # NCC 2-1114. Avatar (MD3) model format courtesy of Id Software, Inc., used by permission. The "Hooligan" model used in our system is courtesy of Toni "Cornix" Daniele and is available on-line at <http://www.planetquake.com/polycount/>.

References

BARNES, G. R. 1979. Vestibulo-Ocular Function During Co-ordinated Head and Eye Movements to Acqurie Visual Targets. *Journal of Physiology*, 127–147.

CAPPS, M. 2001. *Course 42: Developing Shared Virtual Environments*. ACM SIGGRAPH, New York, NY. URL: <http://sharedvr.org/learn/sig00/> (last accessed 01/21/01).

GARAU, M., SLATER, M., BEE, S., AND SASSE, M. A. 2001. The Impact of Eye Gaze on Communication using Humanoid Avatars. In *Human Factors in Computing Systems: CHI 01 Conference Proceedings*. ACM Press, 309–316.

HINDMARSH, J., FRASER, M., HEATH, C., BENFORD, S., AND GREENHALGH, C. 2000. Object-Focused Interaction in Collaborative Virtual Environments. *ACM Transactions on Computer-Human Interaction 7,* 4 (December), 477–509.

ISHII, H. AND KOBAYASHI, M. 1992. ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact. In *Human Factors in Computing Systems: CHI '92 Conference Proceedings*. ACM Press.

LUCIANO, C., BANERJEE, P., AND MEHROTRA, S. 2001. 3D Animation of Telecollaborative Anthropomorphic Avatars. *Commun. ACM 44,* 12 (December), 65–67.

MURPHY, H. AND DUCHOWSKI, A. T. 2002. Perceptual Gaze Extent & Level Of Detail in VR: Looking Outside the Box. In *Conference Abstracts and Applications (Sketches & Applications)*. ACM, San Antonio, TX. Computer Graphics (SIGGRAPH) Annual Conference Series.

RAGUSA, J. M. AND BOCHENEK, G. M. 2001. Collaborative Virtual Design Environments. *Commun. ACM 44,* 12 (December), 41–43.

VERTEGAAL, R. 1999. The GAZE Groupware System: Mediating Joint Attention in Mutiparty Communication and Collaboration. In *Human Factors in Computing Systems: CHI '99 Conference Proceedings*. ACM Press, 294–301.

WATSON, B., WALKER, N., AND HODGES, L. F. 1997. Managing Level of Detail through Head-Tracked Peripheral Degradation: A Model and Resulting Design Principles. In *Virtual Reality Software & Technology: Proceedings of the VRST'97*. ACM, 59–63.

WATSON, B., WALKER, N., HODGES, L. F., AND WORDEN, A. 1997. Managing Level of Detail through Peripheral Degradation: Effects on Search Performance with a Head-Mounted Display. *ACM Transactions on Computer-Human Interaction 4,* 4 (December), 323–346.