Objective Measures for the Effectiveness of Augmented Reality

Mark A. Livingston* Catherine Zanbaka† J. Edward Swan II‡ Harvey S. Smallman§

Naval Research Laboratory, Washington D.C.

Abstract

Augmented reality (AR) systems present a mixture of virtual and real objects. The challenge for AR system evaluators is how to tell whether the virtual world is *effective* at conveying the sense of reality. It may never be possible or even necessary to determine whether the user is truly fooled in all situations or is merely "suspending disbelief," but one can objectively measure the effectiveness of an AR environment with a task-based approach. We present the results of our first such experiment, involving low-level perceptual tasks of recognition and depth matching.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation, Graphical user interfaces; I.3.7 [Computer Graphics]: 3D Graphics—Virtual reality

Keywords: Augmented reality, Presence, Perception, Depth matching, User interface, Evaluation

1 Introduction

A number of prototype augmented reality (AR) systems show the possibilities the paradigm creates. One difficulty in the acceptance of AR is knowing whether the AR system is truly *effective* in its presentation. The AR system evaluator must determine whether the user can perform the task for which the AR system is designed.

This is similar to the concept of presence in virtual environments (VEs) [7]. However, the appropriate measures for AR are different due to the fundamentally different experience of users of AR systems from those of VEs. Subjective metrics do not directly measure the effectiveness towards performing the task for which the AR system is designed. An objective method of measuring effectiveness has the user perform tasks that rely on perception and/or cognition of the graphical objects within the real environment.

A number of experiments have been conducted on depth perception in AR. In perception of nearby objects (0.821–1.810 m) with stereo video-based AR, users have been observed to place a virtual pointer with greater variance than a real pointer [1]. Overall, these users placed both real and virtual pointers in front of the targets, which could also be either real or virtual.

A doctor performed ultrasound-guided needle biopsies with and without the assistance of AR [6]. A second physician evaluated needle placement by objective medical standards for placement. Needle localization was 35% better when using AR. There is no precise analog possible in the real world of the capability this AR system provides the physician: to see a lesion through the patient's skin. However, analogous tasks can be constructed in which the user manipulates combinations of real and virtual objects.

*Virtual Reality Laboratory, Naval Research Laboratory. Corresponding email: mark.livingston@nrl.navy.mil

†Dept. of Computer Science, Univ. of North Carolina at Charlotte

‡Dept. of Computer Science and Engineering, Mississippi State Univ.

§Pacific Science & Engineering Group, San Diego

IEEE Virtual Reality 2005
March 12-16, Bonn, Germany
0-7803-8929-8/05/$20 ©2005 IEEE

2 Current Experiments

This work concentrates on two low-level tasks required for perception in our motivating application [2]: recognizing objects and matching depth against other objects. The latter task we hope will lead us to a sufficient understanding of far-field depth perception in AR in order to allow us to provide a usable perception of depth between real and virtual objects [3].

Both tasks use a Sony Glasstron LDI-D100BE stereo optical see-through display. This display focuses the image at 1.2 m from the user, has a fixed inter-pupillary distance (IPD) of 62 mm and fixed vergence angle of 0°. Virtual images were generated using a Pentium 4 3.06 GHz processor with an Nvidia Quadro4 900 XGL graphics card. The display was fixed in the world and calibrated through manual alignment; it was not tracked.

2.1 Task 1: Resolving Objects

Optical properties in head-worn displays reduce the user's effective visual acuity [5]. Snellen charts are a convenient and widely-used tool for determining real-world visual acuity; resolving one minute of arc of visual angle is normal [4]. Our Sony Glasstron yields 2.205 min, implying a Snellen score of 20/40. We would expect to see further degradation due to blur from the optical elements.

We tested eight subjects on three conditions: natural vision, natural vision through the HMD optics, and vision of HMD graphics. We implemented a virtual version of the Snellen chart with the same letters and apparent size. All users had normal or corrected vision (20/20 or better). All users suffered decreased acuity looking at the real target through the HMD, up to a factor of two. Surprisingly, all subjects tested at 20/30 acuity on the virtual chart; in most cases, this matched their score on the real chart viewed through the HMD.

We note two confounds. We used a standard-size chart, for viewing at 20 feet. While we render the eye chart with the correct apparent size, the Glasstron display focuses only at an apparent distance of 1.2 m from the user. This difference may account for the unexpected performance of the users. The next version of this test will use an eye chart sized for viewing at 1.2 m. The virtual eye chart was anti-aliased, which also might have improved user performance beyond the predicted score.

2.2 Task 2: Depth Matching

We created a depth matching task with which we could test the user's perception of virtual objects against that of real objects. We set eight referents along a hallway (Figure 1) at distances ranging from 5.3 m to 44.2 m, owing to our far-field motivating problem [2]. We asked eight subjects to position virtual and real objects (one object per trial) at the distance of an indicated referent. The users moved a trackball to control both the real and virtual targets and pressed a mouse button to indicate the response. The experimenter pedaled a bicycle using a Wizard-of-Oz method to move the real target. The virtual target was matched in size to the real target. We used a Leica TotalStation to measure the distance to the real target.

Overall, there was no significant difference between the users' accuracy with the real and virtual target. As expected, the users'

Proceedings of the IEEE Virtual Reality 2005 (VR'05)
1087-8270/05 $20.00 © 2005 IEEE

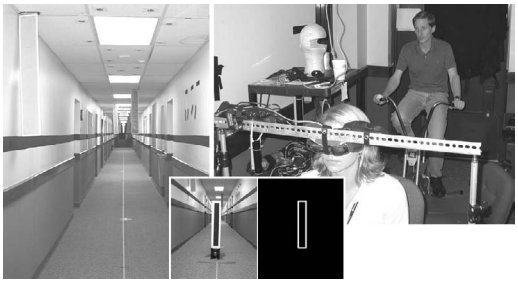

Figure 1: Experimental set-up. The users moved a real (*left-center*) or virtual (*right-center*) target in the hallway (*left*) to match the distance of a colored referent on the ceiling. The experimenter rode a bike (*right*) to move the real target; the users moved a trackball (not visible) to control the virtual target and to cue the researcher.

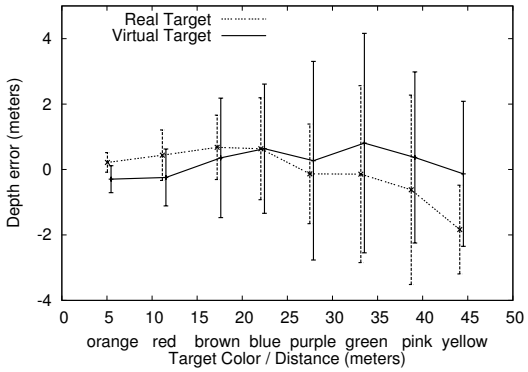

Figure 2: Performance in depth matching. Error bars are one standard deviation and swamp the difference between the real and virtual targets. Distance error is measured in meters; negative errors mean the user left the target in front of the referent (closer to self).

performance was best for the nearest target and degraded as a function of distance (Figure 2). We view the lack of a significant difference between the real and virtual target as a positive result; we expected the real target would yield an easier task.

We measured the standard deviation of the users' responses for each target type (1.826 m) and the difference in the means (0.305 m) and correlation of the means (0.214) for each referent in order to perform a power analysis. Even using these optimistic assumptions and approximations to population statistics, we lack sufficient power to argue for the null hypothesis at significance level $\alpha = 0.05$ ($\delta \approx 2.32$, power $\approx$ 0.64).

We asked users to assess their own performance after the real and virtual task conditions (Figure 3). Three users were unable to accurately assess their performance; one was somewhat inaccurate. This shows the difficult nature of subjective assessment and argues for objective measures. Users may convince themselves of untruths, be unable to identify the factors affecting their performance, and be unable to assess their performance. These conditions do not correlate with actual performance.

3 Conclusions

We have argued for an objective measure of the effectiveness of augmented reality that derives its claims of objectivity and measuring effectiveness from a task-based approach. The crucial aspect of the tests envisioned and performed thus far is that we compare users' performance with virtual objects against users' performance with real objects. The comparison with the real task will enable us

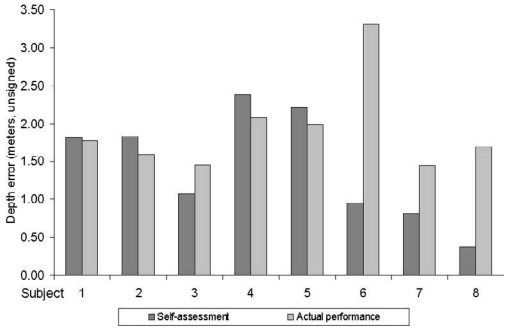

Figure 3: Subjects' self-assessment compared to actual performance on the virtual target. Self-assessment is normalized to match actual performance on the real target (not graphed). Bars indicate self-assessment and actual performance with the virtual target.

to differentiate between an inadequate representation for the virtual objects and subjects' innate difficulty with the task. The former would limit performance only on the virtual task; the latter would limit both. Note that we do not assess whether the user "suspends disbelief" of the virtual objects; such a belief would not prevent a subconcious cue from interefering with the user's performance. Similarly, maintaining disbelief in the "realness" of the graphical objects would not prevent the user from successfully using the information presented in AR.

In being unable to conclude that users were able to perform a task better with a real object than a virtual object, we can maintain hope that AR systems provide useful and natural cues to a user performing a task. Our task-based approach makes strides towards creating objective measures of effective augmented reality systems.

Acknowledgements

We thank Warren Robinett and Jannick Rolland for recollections of early eye chart tests on an HMD, Steven Feiner for the Sony Glasstron, and Jennifer King for advice on statistical testing. Thanks also to Yohan Baillot, Dennis Brown, Simon Julier, Elliot Cooper-Balis, and Erik Tomlin for experimental support, Larry Rosenblum and Ruth Willis for funding support, and our anonymous subjects.

References

[1] David Drascic and Paul Milgram. Positioning accuracy of a virtual stereographic pointer in a real stereoscopic video world. In *SPIE Volume 1457: Stereoscopic Displays and Applications II*, pages 58–69, September 1991.

[2] Mark A. Livingston, Lawrence J. Rosenblum, Simon J. Julier, Dennis Brown, Yohan Baillot, J. Edward Swan II, Joseph L. Gabbard, and Deborah Hix. An augmented reality system for military operations in urban terrain. In *Interservice/Industry Training, Simulation, and Education Conference*, December 2002.

[3] Mark A. Livingston, J. Edward Swan II, Joseph L. Gabbard, Tobias H. Höllerer, Deborah Hix, Simon J. Julier, Yohan Baillot, and Dennis Brown. Resolving multiple occluded layers in augmented reality. In *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003)*, pages 56–65, October 2003.

[4] Lorrin A. Riggs. Visual acuity. In Clarence H. Graham, editor, *Vision and Visual Perception*, pages 321–349. John Wiley and Sons, 1965.

[5] Warren Robinett and Jannick P. Rolland. Personal communication, August 2004.

[6] Michael Rosenthal, Andrei State, Joohi Lee, Gentaro Hirota, Jeremy Ackerman, Kurtis Keller, Etta D. Pisano, and Henry Fuchs. Augmented reality guidance for needle biopsies: A randomized, controlled trial in phantoms. In *Medical Image Computing and Computer-Assisted Interventions (MICCAI) 2001*, October 2001.

[7] Mel Slater. A note on presence terminology. *Presence-Connect*, 3(1), January 2003.

1087-8270/05 $20.00 © 2005 **IEEE**