Indoor vs. Outdoor Depth Perception for Mobile Augmented Reality

Mark A. Livingston* — Naval Research Laboratory
Zhuming Ai† — Naval Research Laboratory
J. Edward Swan II‡ — Mississippi State Univ.
Harvey S. Smallman§ — Pacific Science & Engineering Group

Abstract

We tested users' depth perception of virtual objects in our mobile augmented reality (AR) system in both indoor and outdoor environments using a depth matching task. The indoor environment is characterized by strong linear perspective cues; we attempted to re-create these cues in the outdoor environment. In the indoor environment, we found an overall pattern of underestimation of depth that is typical for virtual environments and AR systems. However, in the outdoor environment, we found that subjects *overestimated* depth. In addition, our synthetic linear perspective cues met with a measure of success, leading users to reduce their estimate of the depth of distant objects. We describe the experimental procedure, analyze the data, present the results of the study, and discuss the implications for mobile, outdoor AR systems.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/Methodology; H.1.2 [Models and Principles]: User/Machine Systems—Human factors

1 Introduction

It is well-known, though not well-understood, that perception of depth in virtual environments (VEs) suffers from *depth compression*, leading to an underestimation of egocentric depth. There have been a number of efforts to understand the analogous situation in augmented reality (AR), in which the depth of a virtual object must be understood. Though some of these efforts have been in support of outdoor applications, previous depth studies did not test the depth perception of users in outdoor environments.

Our main goal in this experiment was to fill in this gap in the understanding of depth perception within AR systems: how is the depth of a virtual object perceived relative to the real environment in an outdoor scene? To solve this, we designed a test that placed users in an indoor AR environment (Figure 1, top) and (in a second session) an outdoor AR environment (Figure 1, bottom). We replicated our previous depth matching task [20]. Our secondary goal was (under the assumption that the indoor environment's strong linear perspective cues would improve depth estimation) to determine whether a virtual analog of these cues would provide a similar improvement in the outdoor AR environment.

2 Related Work

Among the numerous depth cues that the human visual system uses to interpret depth from the projection of the three-dimensional world onto the retina is relative size. One may also consider depth perception in near-field (within arm's length), medium-field (within conversational distance), and far-field distances [2]. Cues may vary in availability, potency, and saliency between these spaces. Relative size, however, is considered to be nearly constant in saliency across depth, but also weak compared to the strength of the occlusion cue at all distances and weak compared to binocular disparity, motion parallax, and height in the visual field at distances typical for interaction with an object. (Such distances correspond to the nearer half of our experiment.)

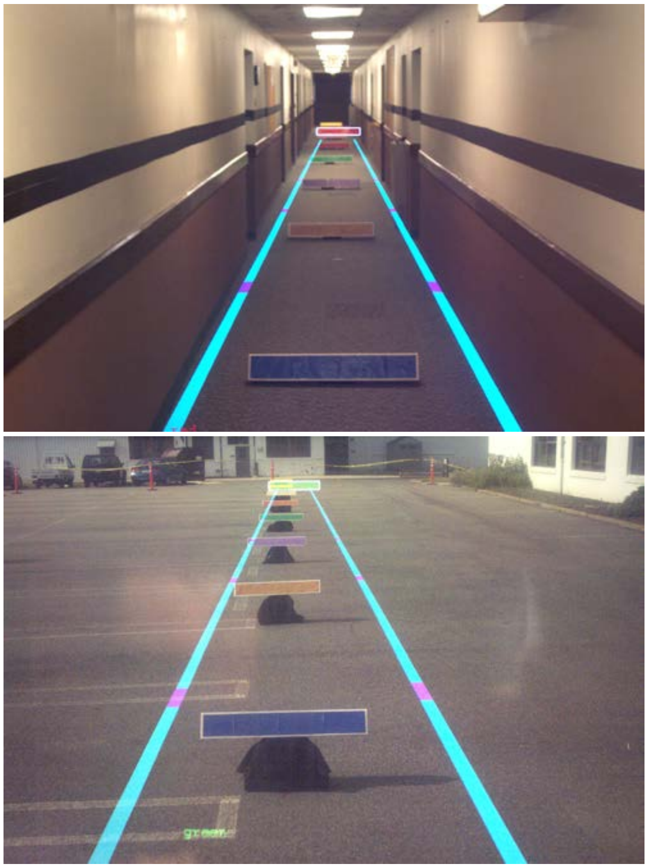
Figure 1: The indoor (top) and outdoor (bottom) scenes for the experiment. Visible in these images are the colored referents and a virtual object (floating above the ground plane) which the user was to match in depth with the referent of the same color. (See Figure 10 in the color plate to see the color scheme.) These images show both types of linear perspective cues we introduced.

A smaller number of studies have been conducted to understand the perception of depth in head-worn AR displays. Rolland et al. [17] found in a pilot study that at near-field distances (0.8-1.2 meters), depth of virtual objects was overestimated. A follow-up experiment [18] compared forced-choice and depth-matching tasks with an improved AR display; they found improved accuracy and no consistent bias in the estimated depth.

Ellis and Menges [4] ran a series of AR depth perception experiments in the near field (0.4-1.0 meters). They studied occlusion of the virtual depth location, convergence, accomodation, observer age, and display using monocular, bi-ocular, and stereo visualiza-

*e-mail: mark.livingston@nrl.navy.mil
†e-mail:ai@itd.nrl.navy.mil
‡e-mail:swan@acm.org
§e-mail:HarveySmallman@pacific-science.com

IEEE Virtual Reality 2009
14-18 March, Lafayette, Louisiana, USA

tion. They found that monocular viewing increased error (since stereo is a potent cue in near-field depth perception), and that the occlusion of the virtual location caused a change in the vergence angle, biasing the depth judgment towards the user. Opening a virtual hole in the occluding surface decreased this error. A follow-up study incorporated motion parallax and AR system latency as variables. Depth perception errors increased with increasing distance and latency.

Kirkley [9] studied occluding surfaces, the ground plane (a form of linear perspective), and object type (real, realistic virtual, and abstract virtual) in monocular AR viewing at medium-field distances (3-33.5 meters). He found that occluding surfaces increased error, placing objects on the ground plane decreased error, and judging the depth of real objects was most accurate.

Our first experiment at far-field distances [12] used graphical parameters such as drawing style (fill, wire-frame, or both), opacity, and intensity on occluded virtual objects at far-field distances (60-300 meters). We found that using wire-frame outlines of filled shapes, decreasing opacity with distance, and decreasing intensity with distance resulted in better ordinal depth perception. A follow-up experiment showed that similar errors were made when matching the depth of real objects and unoccluded virtual objects against real referents [13].

Jurgens et al. [8] tested several cue in video-overlay head-worn ARfor a task involving depth-to-ground judgments (where distance is up to approximately two meters). Subjects were asked to put a stake in the ground at a particular location, which relies on registration primarily and depth as a supporting cue. They found a preference for a "cast circle" whose position was directly below the stake and whose radius was equal to the stake's height off the ground, but no performance improvement.

Lappin et al. [10] compared environmental context for their effects on egocentric depth judgments with real environments and people as the items being judged for depth. They found a wide-open outdoor environment to yield nearly veridical judgments, but an indoor hallway to produce an expansion of the perceived distances and greater variability in the estimates. Bodenheimer et al. [1] extended this protocol to VEs and found a typical distance compression in the VE at 30m, but not at 15m; they found no significant difference between the indoor and outdoor environments.

Our most recent experiments [20] explored medium- and far-field distances (5-45 meters) with variables such as position in the visual field, occlusion of the site of the virtual object, and practice on the task. We found a switch in the error of depth matching of virtual objects, from underestimation inside of 23 meters to overestimation beyond. We estimated an 8% increase in the rate at which error increased with distance from the unoccluded to occluded conditions. A follow-up experiment [19] examined the effect of the experimental protocol and object type in medium-field distances (3-7 meters). We found less underestimation than previous studies, but a consistent bias nonetheless. Another follow-up study [7] studied AR depth perception against VR depth perception, suggesting that the virtual background contributes to the underestimation of depth in immersive virtual environments.

3 Experimental Design

The design of this experiment was based on the experiment in [20]. The main goal in this experiment was to test the depth perception on indoor and outdoor versions of the task, with a secondary goal to see if AR visualizations could replicate the powerful linear perspective depth cues provided by the indoor environment.

Our image generation system was a Dell Precision workstation equipped with a Pentium4 processor and nVidia Quadro4 900 XGL graphics processor. Images were displayed on an nVisorST optical see-through display. This display offers dual SXGA displays (1280 × 1024@60Hz) with a 48° × 36° field of view (FOV) and adjustable inter-pupillary distance (IPD). We estimate that the experienced vertical FOV of the see-through region is reduced by 3.4° due to the position of the optics and LCD relative to the housing of the display. We used 40° for the horizontal FOV in the rendering for all trials (indoor and outdoor); it may be that the way the display sat on the user's head caused this mismatch, but this enabled the best registration of all graphics. Users were tracked with an IS-1200 vision-inertial hybrid tracking system.

3.1 Subjects

We drew a pool of twelve subjects from the laboratory and clerical staff at our site. Eight men and four women between the ages of 22 and 51 (mean=35.6) completed an indoor and an outdoor session of the experiment. All volunteered and received no compensation. Our subjects reported being heavy computer users, having no known problems with depth perception, and being comfortable with geometric reasoning. Most subjects did not report having any trouble learning or completing the experiment. One subject reported (prior to the experiment) being very susceptible to motion sickness; this subject took an extended rest during the indoor session, but otherwise had no trouble. It should be noted that the indoor session for each participant was slightly longer; it was during this portion that the stereo test was administered and the user's inter-pupillary distance (IPD) and vergence were measured. These latter calibrations were not repeated for the second (outdoor) session. Also, less training on the task was necessary, since this was identical in both sessions.

We screened our users for proper stereo acuity with a task in which they were asked to identify which of four targets was closer. One target was presented with a slight stereo disparity that made it appear closer than the others in the set. Users were shown nine such sets, and most users correctly identified all nine targets that were closer. Some users made one or two mistakes, but these were on targets that required high acuity; thus all subjects were judged to have passed the stereo fusion screening test.

We asked users if they were color-blind, given that we use color to identify our targets; no subjects said they were, and each subject was told the color names that would be used. This turned out to be necessary in the outdoor condition, as the sunlight faded the colors of the targets. Users were told to ask if they were unsure of which target was a particular color. The experimenter responded with an ordinal direction (e.g. "third" or "second-to-last"). With only eight colors, this was rarely requested after the first occurrence of a particular color; however, as we shall discuss later, there was evidence of incorrect matching.

3.2 Experimental Task

Subjects used a trackball to maneuver a (virtual) target in one dimension: by pushing the trackball away, the virtual rectangle moved further from the subject's location. Subjects were shown a set of eight (real) referents (visible in Figure 1) that were identified by color. For each trial, a color name appeared in the lower-left of the subject's augmented view. The virtual target appeared in that color at a random initial depth within the experimental area. The subject then moved the virtual target until its depth matched the depth of the referent of the named color. We mapped pixels on the screen directly to distance in the virtual world; subjects could move the target with a resolution of approximately 1mm. The target would disappear behind the subjects if they pulled it too far forward. There was no limit to how far they could move it away from them; the farthest outlier among the analyzed data was about 74m. This lack of limits on the target position was true in both the indoor and outdoor tasks, implying that subjects could push the target beyond a physical occluder with no effect on its presentation. No feedback was provided to subjects on their performance at any time during the trials. Thus the subjects had no way of knowing whether their placement of the target was correct or not.

Subjects were tracked and could thus move horizontally or vertically with respect to the central line of vision towards the referents, but they were asked not to move much forward or backward so as not to significantly change the depth to the referents. A table in front of the user aided them in keeping a constant position with respect to the referents; this table held the trackball. When subjects believed the target was at the proper depth, they pressed the space bar on a keyboard to indicate the conclusion of the trial. The graphics would disppear, leaving only the real scene visible; to begin the next trial, the subject pressed the space bar again. Most subjects learned to press the space bar twice after a few trials, so that there was virtually no rest between most of the trials. Subjects were instructed that they may at any time rest between trials, and they did so as needed.

For each user, we calibrated our system as follows. First, we set the system for the user's IPD. We then asked the user to focus on a virtual calibration object at a distance beyond the farthest referent and determine if the calibration object was closer or farther than real object. Indoors, this calibration object was at the end of the hallway. Outdoors, this object was the nearest building, approximately 61 meters away. We then adjusted the vergence angle between the user's eyes until the user said the two distances were equal. We also aligned this virtual object to a specific real object in order to measure any offset between the orientation within the ground plane as reported by the tracking system and as specificied in the model used to generate the graphics. These two angles were applied to all subsequent renderings.

3.3 Independent Variables and Counterbalancing

We used the following independent variables in the experiment with the following counterbalancing methods. All variables were within-subjects.

- **Environment**$\in \{\text{indoor}, \text{outdoor}\}$ *within subjects*
  All subjects completed the test in both indoor and outdoor environments (Figures 1 and 10). In order to minimize disruption to our co-workers, this variable was *not* counterbalanced; all users completed the indoor session, then returned 7-14 days later to complete the outdoor session. Also, the outdoor environment was not completely flat. We propped the referents up on platforms to match the angle of declination from the user's position in the indoor and outdoor environments. This meant that we were unable to match the height off the ground; there was a difference of 28 cm between the height of the lowest and highest ground points that were used for positions of the referents in the experimental area. (For comparison, the referents are 97 cm wide and 14 cm tall.)

- **Tramlines**$\in \{\text{on}, \text{off}\}$ *within subjects*
  In both environments, subjects saw tramlines at floor level that were at the width of the hallway used for the indoor environment. Thus in the indoor condition if our system achieved perfect registration, these lines would have been aligned with the intersection of the hallway floor and side walls (Figure 1). This variable was randomly permuted with the gridpoints, distance, and repetition.

- **Gridpoints**$\in \{\text{on}, \text{off}\}$ *within subjects*
  In both environments, we used gridpoints placed along the tramlines evenly in distance (Figure 10). This variable was completely crossed with the other variables; thus subjects would see conditions with only tramlines, only gridpoints, both tramlines and gridpoints, and neither tramlines nor gridpoints. This variable was part of the randomly permuted set.

- **Distance**$\in \{4.83, 9.66, 14.49, 19.32, 24.15, 28.98, 33.81, 38.64\}$ *within subjects*
  The referents were placed at these distances (in meters); the same order for the colors was used in both environments. This variable was part of the randomly permuted set.

- **Repetition**$\in \{1,2,3,4,5\}$ *within subjects*
  Users completed each combination of the above variables five times. Each combination of the above three variables was counted and the repetition thus labeled when the result of the trial was stored in the output data.

To emphasize, all variables were completely crossed. All variables except the environment were randomly permuted; each user completed an indoor session and then an outdoor session 7-14 days after the indoor session. Thus each user completed $2(tramlines) \times 2(gridpoints) \times 8(distance) \times 5(repetitions) = 160$ indoor trials and 160 outdoor trials, for a total of $320 \times 12(subjects) = 3840$ data points.

3.4 Dependent Variables

For each trial, we measured the virtual distance at which the user placed the target; this gives us the ability to compute a signed error in distance. We measured the time to complete the trial and the user position at the time the response was entered. Each user completed the NASA TLX [6] after each of the indoor and outdoor sessions. Users also completed a subjective questionnaire before and after each session. Users were asked about their physical health both before and after each session, as well as to assess their performance after each session.

3.5 Hypotheses

We hypothesized that the outdoor environment, with the loss of the linear perspective provided by the hallway, would result in greater error and less precision in the placement in depth of the virtual object. We did not think the relative size cue would be strong enough to guide users to the correct answer, and thus the tramlines and gridpoints would reduce this error and increase the precision, with tramlines proving more useful than gridpoints. Also, the relative size of the target and the referents would be the same for the indoor and outdoor environments; thus we hoped to see the tramlines and gridpoints compensate for the linear perspective cues provided by the hallway on the indoor trials. As is typical of depth perception, we expected increasing errors and decreasing precision with increasing depth. Based on results of our previous studies [19], we expected users to improve their performance with repetition of the task, a typical learning effect. This occurred in the previous studies with the same task, despite the lack of feedback on performance.

4 Results

We analyzed the data with a repeated-measures 2(environment) × 2(tramlines) × 2(gridpoints) × 8(distance) × 5(repetition) analysis of variance (ANOVA) with the |STAT analysis package [15]. We considered three forms of error. *Signed error* considers the direction of the error, whereas *absolute error* does not. *Normalized error* attempts to counteract the well-known decay of accuracy and precision with increasing distance and is a preferred measure when looking for effects not due to distance. Normalized error equals the estimated distance divided by the distance of the referent.

4.1 Main Effects

Environment on Normalized Error There was a main effect of environment on normalized error – F(1,11)=11.032, p=0.007 (Figure 2). Users tended to underestimate the depth in the indoor environment and overestimate the depth in the outdoor environment. One can see in the graph that veridical perception (horizontal line labeled "1") separates the graphs of the indoor and outdoor

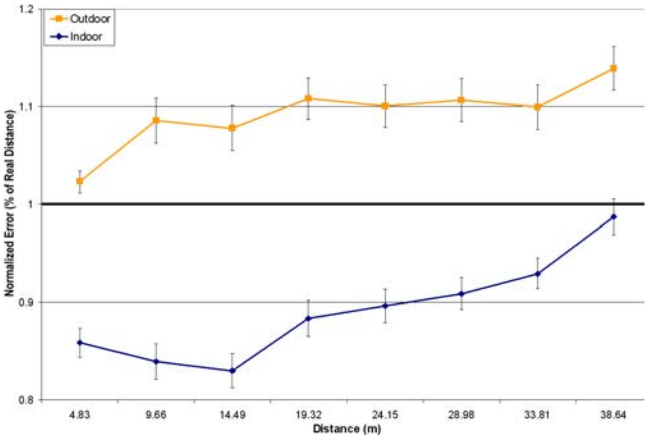

Figure 2: The main effect of environment on normalized error shows the underestimation of depth indoors and overestimation of depth outdoors. Though there is variation with distance, it is less than the difference between the two environments. The thick line at 1.0 denotes veridical perception.

data. There is a trend of increasing the depth estimate visible in this graph, but the result here is not affected by the distance, since we are using normalized error. It should be noted that environment also showed a significant main effect on signed error, F(1,11)=11.072, p=0.007.

Repetition on Normalized Error There was a main effect of repetition on normalized error – F(4,44)=6.613, p=0.000 (Figure 3). Normally, one would expect to see such an effect only if feedback were provided on the user's performance after each trial. This would indicate learning on the part of the user. It should be emphasized that we did not give the users any feedback during the test. There was no way for them to know whether they were correct or not other than their own perception. Also, our users simply increased their estimate of depth. In the indoor environment, where they were underestimating the depth, this meant an improved estimate. But in the outdoor environment, where they were already overestimating depth, they continued to increase their estimate of depth (and thus their error).

Gridpoints on Signed Error In our initial analysis, we found a main effect of gridpoints on signed error – F(1,11)=5.279, p=0.042 (Figure 4). However, upon examination of the graph, it appeared to us that the effect was almost entirely due to the sixth referent (at 28.98 meters). We re-ran the analysis without this referent present and found that the presence of gridpoints was no longer considered significant. This, and the lack of a significant main effect of the tramlines, is a disappointing result in that we had hoped to re-create the powerful effect of linear perspective with these graphical aides. However, it appears that the users did not require them as much as we had thought, so the effect is there, but in a more limited case than a main effect for the whole of the data, as discussed below.

Distance on Absolute Error There was a main effect of distance on absolute error – F(7,77)=13.391, p=0.000 (Figure 5). We see the typical approximately linear increase in error with increasing distance in both the indoor and outdoor environments. This means that performance on our task is consistent with findings on depth perception tasks in the literature. The slight difference in the rate of increase in error has to do with differences between the indoor and outdoor environments discussed below. This helps validate the task (used also in [20]) as a measure of depth perception.

Distance on Time There was a main effect of distance on time – F(7,77)=25.620, p=0.000 (Figure 6). Users were, as is typical of depth perception, slower with increasing distance. In order to remove the effect of the starting position on the response time, we chose the starting position for each trial from a uniform random distribution across the space in which the referents sat. We did force the starting position to be at least five meters (i.e. about the distance between referents) away from the correct position, however. This ensured that the subjects would have to do something to the target in order to complete the task.

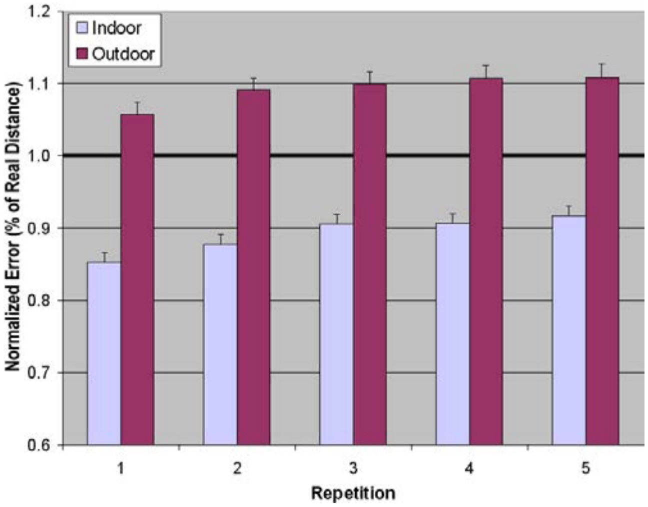

Figure 3: The main effect of repetition on normalized error shows the increasing distance of the estimated depth in both indoor and outdoor environments. Note that only in the indoor condition would this increasing estimate represent a learning effect. The thick line at 1.0 denotes veridical perception.

Repetition on Time There was also a main effect of repetition on time – F(4,44)=16.717, p=0.000 (Figure 7). Users were notably slower on the first repetition than on succeeding ones. This can be attributed in part to learning the order of the referents and partly due to a standard practice effect (familiarity with equipment, etc.). We note that the difference between the fourth and fifth (final) repetition is quite small, and that for some distances, the fourth repetition was actually faster on average. So we feel that users became as comfortable with the task as could be expected.

4.2 Result for Tramlines

In the interactions, we discover one of the more interesting results, and the result that indicates our graphical aides were indeed making a difference, if perhaps in a more restricted case than we had initially hoped.

We noticed in the graph of normalized error separated by distance and environment (Figure 8) that the graphs seem to separate. We ran an analysis using just the data from referents 6, 7, and 8 (28.98, 33.81, and 38.64 meters, respectively). Note that these referents were clearly beyond the reach of the use of stereo (convergence) as a depth cue. Also, we note that these were beyond the depth at which our previous study found a switch in bias in an indoor environment from underestimation to overestimation [20].

This restricted analysis showed a main effect from the presence of the tramlines on normalized error – F(1,11)=5.356, p=0.041. (There was also a main effect of distance, even though only three distances were in the data.) When a similar restricted analysis was performed with the data from the first five referents, there was no main effect from the tramlines.

In considering analysis techniques in previous work, we graphed the performance of the individual users in our study. Since some users are simply not proficient at depth perception, it is possible that the inclusion of a user who performs poorly may produce "results"

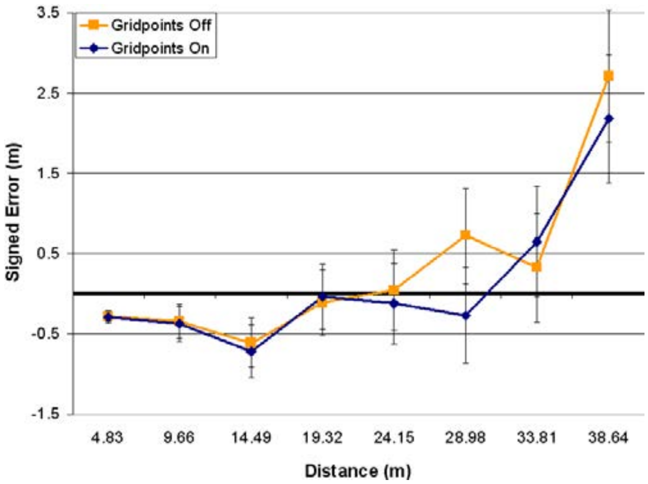

Figure 4: The main effect of gridpoints on signed error is entirely due to the sixth referent; when it was removed from the data, the effect was no longer significant. Since this graph shows signed error, the thick line at 0.0 represents veridical perception.

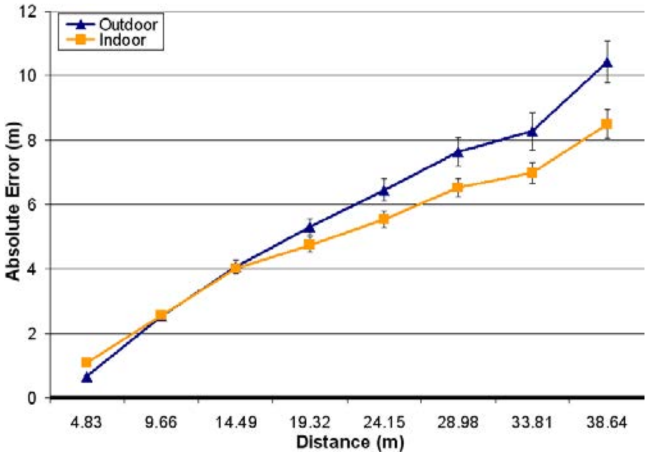

Figure 5: The main effect of distance on absolute error shows the typical decay of performance with increasing distance. Though the rates are slightly different for the indoor and outdoor environments, this demonstrates that performance on our task is consistent with depth perception. For absolute error, veridical perception is also represented by 0.0, at the bottom at the graph.

that are not truly significant or mask results that are significant but are not recognized as such. Figure 11 shows the performance graph for all users. Based on this graph, we eliminated users S01, S03, and S12 from our analysis and re-ran the ANOVA.

The main effect of environment on normalized error remained – F(1,8)=13.744, p=0.006. The main effect of repetition on normalized error also remained – F(4,32)=4.641, p=0.005. Thus we are confident that these effects are not due to outliers in our subject pool. Some interesting interactions appeared in this analysis, such as a tramlines-by-repetition interaction – F(4,32)=2.855, p=0.039 – that we have yet to explore fully.

4.3 Subjective Results

We asked users to complete the NASA Task Load Index (TLX) [6] after both the indoor and outdoor sessions. While there is little data to analyze with this coarse sampling, we do note that the average workload for the indoor sessions was 52.0 (range$\in [35.3, 69.4]$) whereas for the outdoor sessions, it was 44.2 (range$\in [18.9, 77.4]$). With the extremely high ranges for both types of sessions, we find it difficult to draw any conclusions from this data. We do note that nine of the twelve users found the second (outdoor) session to have a lower workload.

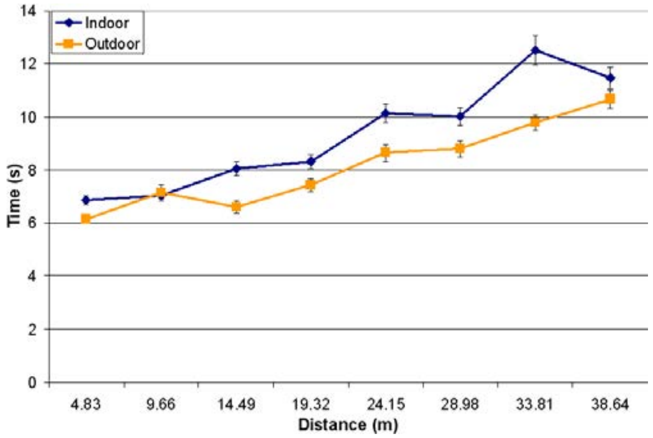

Figure 6: The main effect of distance on response time is typical; users were slower with increasing distance. This is not due to the starting position of the target, which was drawn from a uniform random distribution.

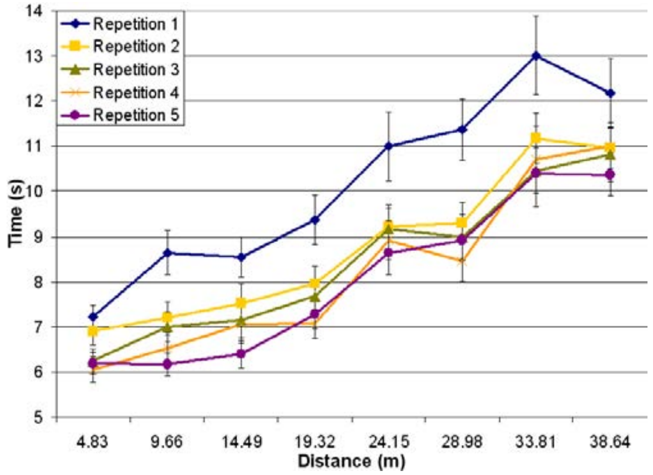

Figure 7: The main effect of repetition on response time is also typical; users were faster with repetition. Note that the greatest difference is between the first repetition and the others.

We separately asked users to rate their own performance after both sessions. Ten users thought that their performance was better on the outdoor session than on the indoor session. Users also tended to note a slight increase (1-2 points on a ten-point Likert scale) in eye strain, fatigue, dizziness, and pain during their indoor and outdoor sessions. There was little difference between the environments, though we do note a consistently (if slightly) higher increase in these types of discomfort on the indoor sessions. This was most likely due to the shorter time required in the second (outdoor) session, in which we did not need to spend as much time on calibration. (We did not re-measure IPD, re-test stereo, or re-measure the user's height.) Counterbalancing the order of indoor and outdoor sessions would have been a desirable study design, albeit difficult to secure the sites.

These results may reflect familiarity with the equipment gained during the experiment and tolerance of the ergonomic discomfort associated with it.

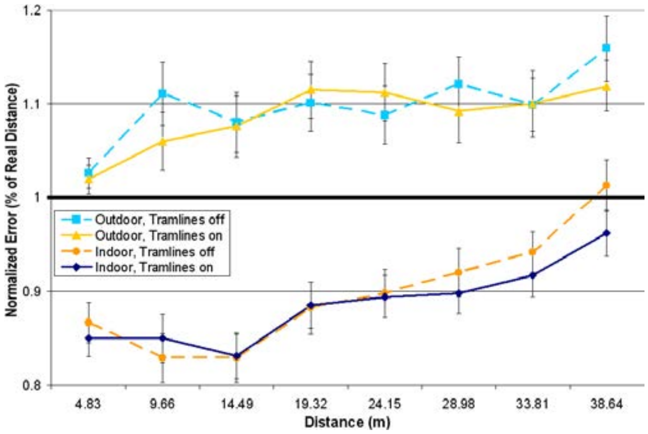

Figure 8: The graph of normalized error separated by the use of tramlines and by environment led us to investigate if the tramlines were useful, but in a more restricted case of distance referents than we had intially hoped to find. When restricted to referents 6, 7, and 8, there is indeed a significant main effect of the presence of the tramlines. Again, the thick line at 1.0 represents veridical perception.

5 Future Work

The analysis we conducted used all data from the subjects who were judged to be sufficiently accurate, but we have noticed in the data from these subjects some apparent outliers in the distance estimates. Many of these are readily interpretable as matching the target to an incorrect referent. It may be that users did not remember the order of the colors correctly and were having trouble seeing them. This was most notably a problem with some of the outdoor sessions near the end of the experiment, when the colors were quite faded. An initial analysis identified 23 outliers from the remaining nine subjects over both indoor and outdoor sessions. However, it is not clear when an accurate but inconsistent distance estimate (Figure 12) should be considered an outlier and when it should be considered simply an instance when the subject overcame other difficulties. We need a good criterion for identifying outliers in distance estimates so that we may test whether these outliers created or masked any effects.

There are some limitations that should be noted about our current experiment. The order of the indoor and outdoor sessions was not counterbalanced. This was due to an unfortunate limitation in the availability of the spaces we used. Though subjects had from seven to fourteen days between their indoor and outdoor sessions, it is theoretically possible that the increase in their estimated depth is due to an order effect. Order has been shown to have an effect on distance estimates [16, 21]; however, it is not clear how these results comparing completely real and completely virtual environments with a time-to-walk estimate apply to our results with AR scenes and depth matching. Also, order dependence seems unlikely given the perceptual nature of the task and the time between our indoor and outdoor sessions, but it is something that would be wise to test in a future experiment. However, a new indoor space will have to be identified if new data is to be gathered from an indoor environment. Another goal for future experiments is to find a flatter outdoor experimental area, so that it does not introduce a confound of the height-off-ground between the indoor and outdoor sessions. It would be nice to test the tramlines in a less confined indoor environment as well.

The current data shows an underestimation of depth in the indoor condition for all distances we tested. Our previous work [20] showed a switch from underestimation to overestimation in this condition (the only one tested in the previous study). Despite the modest changes in the distances, we expected to see a similar change at least in the indoor condition. We should investigate why this pattern of responses was not present in the current experiment.

Another interesting question for further study is how the mis-registration of the graphics affected the users' estimates of depth. The mis-registration visible in Figure 10 was typical performance for our system. Registration error has been shown to affect other tasks at similar distances with our AR system [11]. A controlled experiment would be of great practical value for outdoor AR systems.

6 Discussion

The most interesting result from this experiment is the difference in depth estimation between the indoor and outdoor environments. We see an underestimation indoors but an *overestimation* outdoors; this is a curious result. For outdoor AR systems, this is clearly an important consideration. Previous studies of depth perception in virtual environments consistently indicate a depth compression in immersive environments. Studies of depth perception in AR have shown mixed results, with some underestimation (especially at closer distances) and some overestimation. But we found a consistent overestimation at all distances (including some used in those previous experiments) when users were looking at an outdoor scene. This overestimation was surprising and appears to conflict with our own previous work and other depth perception findings. However, there are no good direct comparisons with other work, since depth perception at these distances in outdoor AR systems is not well-studied, and we used a single protocol of depth matching. It should be noted that this protocol most closely matches the task that we expect users to perform in our AR applications; Gibson [5] argues for this type of ecological validity.

The big difference between the two environments is quite obvious and known to be an important factor in depth perception: strong linear perspective cues provided by the structure of the indoor environment clearly must be affecting the users' depth perception. One possibility may be related to the classic Ponzo illusion (Figure 9).

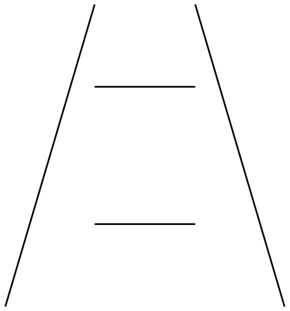

Figure 9: The Ponzo illusion tricks the user into thinking that the upper line is longer because the outside lines induce linear perspective and cause the upper line to be interpreted as farther away.

The Ponzo illusion has two lines that induce a sense of linear perspective [3], just like our hallway and our tramlines. This causes horizontal lines (like our referents) to be interpreted as varying in depth. Our referents of course do vary in depth, but still it may be that the sense of depth is controlled not only by the real cues, but also by the illusory cues that create this illusion. However, we do not see the amount of increase in depth estimate as size constancy would predict [14]. We also wonder if any small variations in other depth cues, such as apparent brightness (since our colors were not initially controlled for brightness, and at any rate were heavily affected by the sunshine during the outdoor sessions) may have created an interaction with our subjects' depth perception that we do not yet understand.

Using the size and the distances of the referents, we can compute their angular size (allowing for the 60° tilt of the front face directly away from the user) and, from that, the number of pixels the target would occupy on the AR display at the same size and distance. These values appear in Table 1. We can see that the relative size constraint gets progressively less precise in matching the size of the target to the referents. Also, we did not account for the reclined pose of the front face of the referents when rendering the target; it was drawn as a front-facing rectangle of 14cm. This resulted in an error of 12-20% in the height of the target, depending on the distance and accounting for the pixel grid. We do not know that subjects were attempting to use this dimension in their size matching; it would seem more likely that they were using the horizontal dimension, in which the target and referents were about seven times larger. But this will be corrected in future experiments.

Distance	Angular Size	Pixel Size
4.83	$10.86° \times 1.44°$	337×49
9.66	$5.67° \times 0.72°$	175×25
14.49	$3.81° \times 0.48°$	117×16
19.32	$2.87° \times 0.36°$	88×12
24.15	$2.30° \times 0.29°$	71×10
28.98	$1.92° \times 0.24°$	59×8
33.81	$1.64° \times 0.21°$	50×7
38.64	$1.44° \times 0.18°$	44×6

Table 1: The angular and pixel size of the referent and the target at the correct distance are computed from the known size, approximate user height, angle of the front face with the user's line of sight, and the measured FOV for the display.

We did succeed in our initial goal in one sense: the tramlines appeared to help users in judging the depth of distant objects in the outdoor environment. Due to misregistration (apparent in Figure 1) and conflict in brightness (among other cues), the tramlines appear to have actually degraded performance in the indoor environment, where they are redundant. But outdoors, the presence of the tramlines caused users to behave against their general tendency to overestimate the depth and thus improve their depth perception. That this was effective (in the statistical sense) only for distant objects may merely indicate that depth judgments at closer distances are possible through other cues, such as convergence and binocular disparity, that are ineffective at the distances of our farthest referents.

These results give us hope that we can ultimately characterize the depth perception of AR users in both indoor and outdoor settings, and that we are beginning to elucidate the factors that create the underestimation and overestimation present in various AR environments.

Acknowledgements

The authors wish to thank the anonymous subjects for their participation and the reviewers for helpful comments. Support provided by the NRL Base Program.

References

[1] B. Bodenheimer, J. Meng, J. Wu, G. Narasimham, B. Rump, T. P. McNamara, T. H. Carr, and J. J. Rieser. Distance estimation in virtual and real environments using bisection. In *Symposium on Applied Perception in Graphics and Visualization*, pages 35–40, July 2007.

[2] J. Cutting. *Reconceiving Perceptual Space*, pages 215–238. MIT Press, 2003.

[3] S. A. DeCaro. Classic line illusions and the depth hypothesis. Retrieved 05 September 2008 from `http://psychology.sdecnet.com/illusion.htm`.

[4] S. R. Ellis and B. M. Menges. Localization of virtual objects in the near visual field. *Human Factors*, 40(3):415–431, Sept. 1988.

[5] J. J. Gibson. *The Perception of the Visual World*. Houghton-Mifflin, 1950.

[6] S. G. Hart and L. E. Staveland. *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*, pages 239–250. North Holland Press, 1988.

[7] J. A. Jones, J. E. Swan II, G. Singh, E. Kolstad, and S. R. Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, 2008.

[8] V. Jurgens, A. Cockburn, and M. Billinghurst. Depth cues for augmented reality stakeout. In *Design-Centred HCI (CHINZ)*, pages 117–124, July 2006.

[9] S. Kirkley. *Augmented Reality Performance Assessment Battery (ARPAB)*. PhD thesis, Instructional Systems Technology, Indiana University, 2003.

[10] J. S. Lappin, A. L. Shelton, and J. J. Rieser. Environmental context influences visually perceived distance. *Perception & Psychophysics*, 68(4):571–581, 2006.

[11] M. A. Livingston and Z. Ai. The effect of registration error on tracking distant augmented objects. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 77–86, Sept. 2008.

[12] M. A. Livingston, J. E. Swan II, J. L. Gabbard, T. H. Höllerer, D. Hix, S. J. Julier, Y. Baillot, and D. Brown. Resolving multiple occluded layers in augmented reality. In *Proceedings of International Symposium on Mixed and Augmented Reality*, pages 56–65, 2003.

[13] M. A. Livingston, C. Zanbaka, J. E. Swan II, and H. S. Smallman. Objective measures for the effectiveness of augmented reality. In *IEEE Virtual Reality 2005 (Poster Session)*, 2005.

[14] S. P. McKee and H. S. Smallman. *Size and Speed Constancy*, chapter 14, pages 373–408. Cambridge University Press, 1998.

[15] G. Perlman. Data analysis in the unix environment. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, pages 130–138. Springer-Verlag, July 1982.

[16] J. M. Plumert, J. K. Kearney, J. F. Cremer, and K. Recker. Distance perception in real and virtual environments. *ACM Transactions on Applied Perception*, 2(3):216–233, July 2005.

[17] J. P. Rolland, W. Gibson, and D. Ariely. Towards quantifying depth and size perception in virtual environments. *Presence: Teleoperators and Virtual Environments*, 4(1):24–49, Winter 1995.

[18] J. P. Rolland, C. A. Meyer, K. Arthur, and E. J. Rinalducci. Method of adjustments versus method of constant stimuli in the quantification of accuracy and precision of rendered depth in head-mounted displays. *Presence: Teleoperators and Virtual Environments*, 11(6):610–623, Dec. 2002.

[19] J. E. Swan II, A. Jones, E. Kolstad, M. A. Livingston, and H. S. Smallman. Egocentric depth judgments in optical, see-through augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 13(3):429–442, May/June 2007.

[20] J. E. Swan II, M. A. Livingston, H. S. Smallman, D. Brown, Y. Baillot, J. L. Gabbard, and D. Hix. A perceptual matching technique for depth judgments in optical, see-through augmented reality. In *Proceedings of IEEE Virtual Reality*, pages 19–26, 2006.

[21] C. Ziemer, J. Plumert, J. Cremer, and J. Kearney. Making distance judgments in real and virtual environments: Does order make a difference? In 3^{rd} *Symposium on Applied Perception in Graphics and Visualization*, page 153, July 2006.

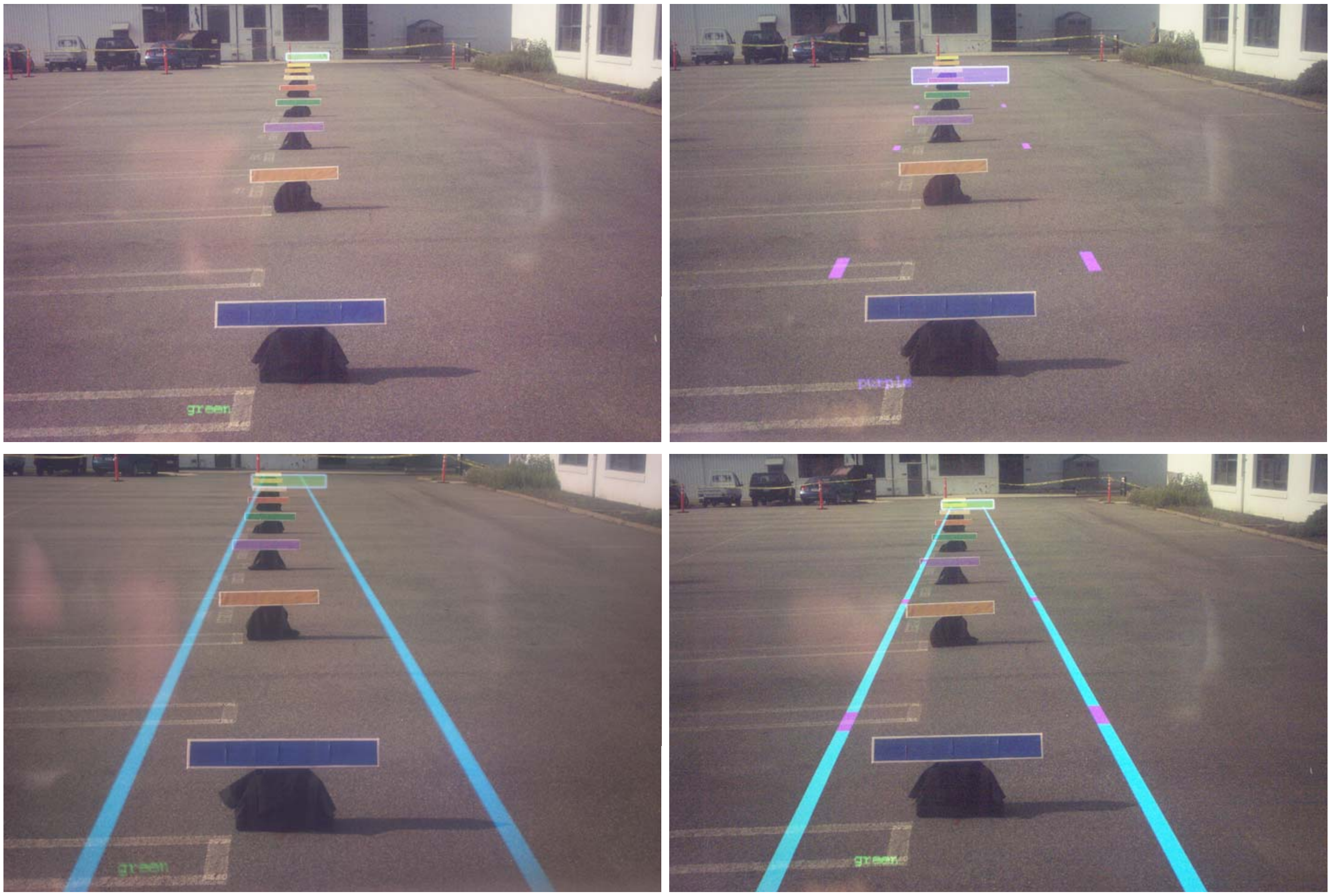

Figure 10: The experimental conditions included no graphical cues (upper left), gridpoints (upper right), tramlines (lower left), and both gridpoints and tramlines (lower right). These graphical cues were meant to mimic the strong linear perspective cues that the hallway provided in the indoor environment for users when observing the outdoor environment pictured here. The mis-registration of the graphics is clearly seen; how much this inhibited the correct estimation of depth is unknown.

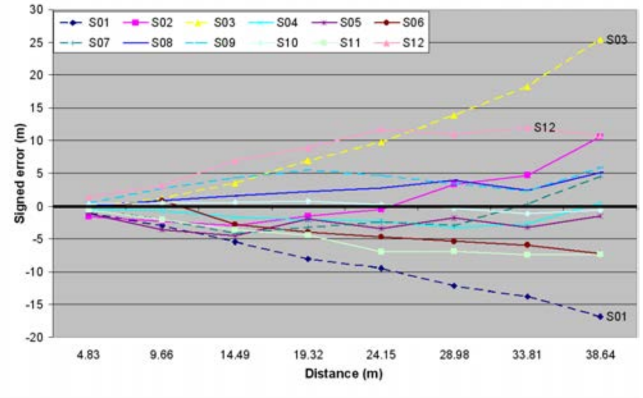

Figure 11: The graph of error for each subject led us to run an analysis without users S01 (lowest estimates), S03 (highest estimates), and S12 (high error in the medium field). This was done to ensure that our results were not due to or masked by poor performance of a subset of users. Again, the thick line at 0.0 represents veridical perception.

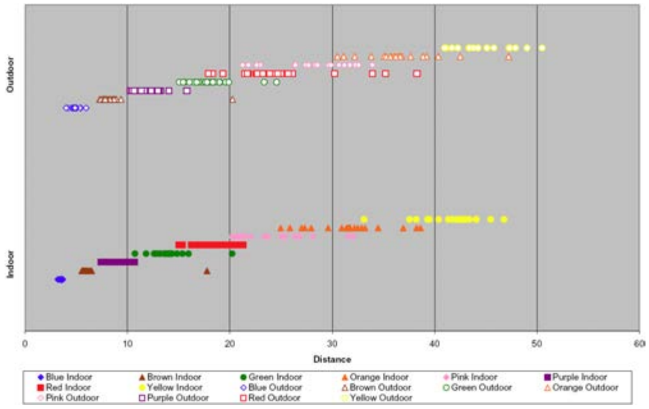

Figure 12: This graph shows one subject's data with individual points denoting depth estimates for each referent, separate by indoor (lower) and outdoor sessions. One difficult question in our analysis is whether to label the right-most data point for the fourth (green) referent of the indoor session as an outlier. It is clearly inconsistent with the other data points, but it is by far the most accurate; most of the data demonstrates severe depth compression.