User Evaluation of See-Through Vision for Mobile Outdoor Augmented Reality

Benjamin Avery Bruce H. Thomas Wayne Piekarski

Wearable Computer Laboratory
School of Computer and Information Science
University of South Australia
Mawson Lakes SA 5095

ABSTRACT

We have developed a system built on our mobile AR platform that provides users with see-through vision, allowing visualization of occluded objects textured with real-time video information. We present a user study that evaluates the user's ability to view this information and understand the appearance of an outdoor area occluded by a building while using a mobile AR computer. This understanding was compared against a second group of users who watched video footage of the same outdoor area on a regular computer monitor. The comparison found an increased accuracy in locating specific points from the scene for the outdoor AR participants. The outdoor participants also displayed more accurate results, and showed better speed improvement than the indoor group when viewing more than one video simultaneously.

KEYWORDS: Outdoor Augmented Reality, Wearable Computers, Telepresence, Image-based Rendering, Occlusion

INDEX TERMS: I.3.6 [Computer Graphics]: Methodology and Techniques – Interaction Techniques; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism – Virtual Reality.

1 INTRODUCTION

Augmented Reality (AR) can be used to augment the user's view of the world with both virtual information and virtual views of real-world information [1]. This paper investigates augmenting the user's view with occluded real locations, using videos captured at remote locations and 3D geometric models of the environment. We have developed a system that can render photo-realistic views of occluded locations that are displayed relative to the user's physical real-world location. In this case an occluded object or location could be a car or building hidden behind another building as seen in Figure 1. The system has been designed so that texture information is sourced from a video stream from the occluded location that is captured from a robot [1], other AR users, or surveillance camera [5]. It is assumed that the source of video information is equipped with position and orientation sensors to aid the rendering system.

Previous research has investigated visualizing occluded objects for outdoor AR [6, 11] and systems capable of rendering photo-realistic 3D scenes of real environments intended for indoor use at a desktop computer [7, 9, 10]. When users view occluded objects in their real-world locations using AR, they can easily comprehend the position, orientation and size.

Viewing remote video images by rendering them on the user's display has been shown to be usable and understandable [11]. When a user is able to see their own surroundings with correctly registered occluded locations directly overlaid (as with AR), they are easily able to determine spatial relationships between the relevant locations. The extreme alternative for the user is to view multiple remote videos on a regular display unaltered. This would pose problems to users as they have to manually determine the spatial relationships between videos. While this requires increased cognitive load for the user, the video images are unaltered, and so are at the highest quality possible.

In this paper we present a study investigating how well users understand video sequences recorded at various locations by comparing current techniques with an image-based rendering technique on an outdoor wearable AR system. While previous research has evaluated user interface techniques for viewing occluded objects in AR, none have compared AR see-through vision with the current conventional method of viewing objects.

The paper is organized as follows. Section 2 discusses related research followed by Section 3 that discusses the visualization system. Section 4 describes the design for our formal evaluation. The results and discussion are presented in Sections 5 and 6. The paper is concluded in Section 7.

2 RELATED WORK

Image-based rendering techniques have been used for many years to generate photo-realistic 3D reconstructions of real objects. Neumann *et al.* [7] developed the Augmented Virtual Environment (AVE) that allowed multiple video sources to be rendered as textures onto a 3D model that the user could navigate using a desktop computer. This was extended by Hu *et. al* [4] to support texturing from a video sequence.

This use of real images in a virtual environment was brought to augmented reality by Kameda *et al.* [5]. The see-through vision tool allows users to 'see through walls' by rendering images from the other side of the wall onto a hand-held display. Video images were captured from fixed surveillance cameras. This system was

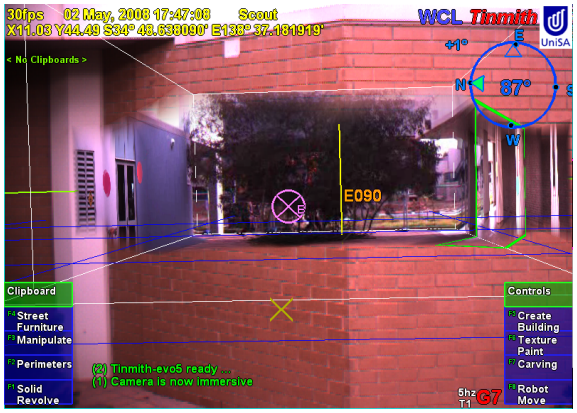

Figure 1 – An AR view showing an occluded area through a building.

e-mail: benjamin.avery@unisa.edu.au
e-mail: bruce.thomas@unisa.edu.au
e-mail: wayne@tinmith.net

IEEE International Symposium on Mixed and Augmented Reality 2008
15 -18 September, Cambridge, UK
978-1-4244-2859-5/08/$25.00 ©2008 IEEE

evaluated by Tsuda *et al.* [11] where they determined that transparency, ground planes, and wire-frame overlays help the users to understand the images they are presented with. A study by Livingston *et al.* [6] further investigated the problem of viewing occluded objects in an immersive AR scenario, and the user's ability to judge spatial relationships between them. Bane and Höllerer [3] investigated X-Ray vision techniques for outdoor augmented reality.

To date there have been few user evaluations on immersive outdoor augmented reality systems. Studies have investigated enjoyment of outdoor AR games [2], interaction techniques and users ability to determine depth [6].

Providing see-through vision has obvious applications for surveillance. The original see-through vision system by Kameda [5] was designed using surveillance camera videos. Wang *et al.* developed a range of desktop visualizations [12] suitable for handling 50 live surveillance cameras installed in a large multi-storey building. Visualizations included 2D images, billboarded images and projected textures. A qualitative study was performed to observe which visualizations people use when performing surveillance tasks.

3 Visualization System

Our previous work presented in [1] was an early implementation of an occluded object visualization system, coupled with a modeling system capable of generating the required 3D geometric models. It was built as an extension to Tinmith [8]. The user uses hand-gesture based tools to create 3D models that are used to render viewpoint corrected images of occluded objects. The video information and modeling of remote or occluded buildings were provided by a mobile robot platform that the user could remotely navigate to suitable locations.

Some additions were made to the system to facilitate the user evaluation. Rather than rendering only the latest frame of a video stream, an image snapshot feature was added to record past frames and render them in the scene also. We chose to use the relative angles between the current image in the video stream, and the previous snapshot angle. Experimentation revealed a total change of 25° to both heading and pitch of the camera ($(\Delta h + \Delta p) > 25$) was a suitable value to ensure coverage of the scene while avoiding storing excess images. The current live video texture is outlined with a green border so that users can easily see which part of the display is being updated, and avoid possible confusion when sections of the display may suddenly update.

It was found during initial testing that occluded objects would appear very small on the display as they were typically quite a distance from the user. We implemented a digital zoom function to allow the user to see more detail of objects at a distance. We found a zoom of 3x to be appropriate for use in our scenario. The zoom mode is activated by holding a button on our wireless remote unit.

4 Evaluation

We conducted a between-subject study comparing user's understanding of a scene in two different conditions: using a wearable computer outdoors to observe a rendered view of the scene, and watching the source video unaltered on a LCD monitor.

We believe that this is a good way of comparing the system presented above with the currently employed alternative. A regular monitor is commonly used for surveillance camera monitoring [12]. The AR see-through vision system has the advantages that information is placed in-situ and it should be highly intuitive for the user to use. It also has disadvantages such as limited HMD resolution, tracker error, image alignment, and that the viewpoint is limited to what the user can physically move to. The purpose of this study is to determine if the AR see-through vision system is a suitable alternative to traditional video monitoring techniques. While alternative video displaying techniques exist for desktop computers [7, 11, 12] they do not leverage the users understanding spatial relationships between locations as AR does.

4.1 Hypotheses

The study will be testing the following hypotheses:

Hypothesis 1: Users are able to understand a video more quickly, and comprehend its contents more accurately when displayed in-situ with a see-through vision system compared to watching it unaltered on a LCD display not co-located with the environment.

Hypothesis 2: Users are also able to understand multiple simultaneous videos more quickly and accurately.

Hypothesis 3: Users will be able to complete a task requiring them to compare and align multiple real-world locations more quickly and accurately.

4.2 Participants

The study was conducted with 34 participants split into two groups of 17. They were assigned to viewing either videos on a desktop computer or the outdoor wearable AR computer. These two groups are referred to as the *indoor* and *outdoor* participants from hereafter. Participants were from a variety of age groups, but were primarily under 50 (0-21: 9, 22-25: 5, 26-30: 9, 31-50: 8, 50+: 3). There were 25 males and 9 females. Of the participants using the wearable AR system, over 75% had experience with outdoor AR

4.3 Tasks

In order to evaluate our hypotheses we created three tasks for each participant to complete. For these tasks a set of pre-recorded videos were created. Two locations (Location A and Location B) were selected from our university campus. Photos of these locations can be seen in Figure 2. In each of the locations we affixed 4 different brightly colored markers to the walls, 2 on each side. The markers were made from 50cm diameter cardboard.

For the study we wanted to ensure that every user viewed exactly the same video sequences to maintain consistency of the results. Using a robot or surveillance camera would have lead to inconsistent camera paths, frame-rates and lighting conditions Videos were captured prior to running the study using a camera, orientation sensor, and GPS. The video files were captured at a resolution of 320x240 at 15fps. The AR visualization system was modified to read MPEG videos instead of a live wireless video stream. The videos had an average length of 16 seconds and consisted of panning horizontally to capture the markers.

4.3.1 Single video task

Participants were randomly assigned either Location A or Location B for the first task. They were given a simple top-down line-drawing map of the location. The outdoor participants were shown the assigned location rendered on the AR display. An example of the user's display can be seen in Figure 1. They were instructed to find the location of the 4 brightly colored dots on the occluded area and mark the location of the markers on the map, and write down the estimated height of the marker.

The indoor participants were shown a video on a computer monitor while sitting at a desk. They were provided with the same map and instructed to mark the same information as the outdoor participants. For both groups the video was played on a continuous loop and the participants could watch the video as many times as needed.

4.3.2 Double video task

In the second task participants were shown a different location from the one in the first task. Multiple locations and ordering were chosen to avoid learning effects. The location was presented to the participant as two simultaneous videos playing at the same time.

Figure 2 - Location A (top) and Location B (bottom) used in the user study. Colored markers were placed on the walls and videos were recorded.

Each video only observed half the markers. The indoor participants watched two videos playing side-by-side on the display. The outdoor participants saw multiple textures being updated at the same time. The participants were asked to indicate the marker locations and heights on a map as in the first task

4.3.3 Scenario task

The third task required the user to complete a more complex set of steps. The task was designed to simulate an emergency rescue situation where the participant had to determine the location of three injured people, and the best point to reach each person from through an adjacent building. The courtyard containing the three people was occluded by a building with 20 windows on it, any of which could be possible rescue points. The participants watched a video captured from a courtyard containing 3 colored markers simulating locations of injured people. Their task was to determine for each of the 3 markers visible in the video, which windows on the adjacent building were directly in front of the markers (at a perpendicular angle to the surrounding walls).

The outdoor participants were taken to an area where they could clearly see a wall with the evenly spaced windows, but the courtyard location was occluded. The occluded area was visible using the see-through vision. By walking along the visible wall, they could line up the real windows with the occluded markers. Participants had to select a window (numbered 1-20) for each marker.

The indoor participants were provided with the recorded video, a satellite photo of the area, and a photograph of the windows. In order to complete the task they had to determine marker locations on the satellite map, then a location on the windowed wall, and from that determine the correct window.

5 RESULTS

Each participant filled in answers on the provided maps, and also completed a questionnaire. The time taken to complete each task was recorded by the researcher. Analysis of study results was performed by a t-test for independent samples.

5.1 Accuracy

For each participant there were a total of 8 markers (4 for each of the first two tasks) that they selected a distance along the wall and a height for. We found no significant difference in accuracy between the single to double video tasks in either the indoor group or outdoor group (p=0.42 and 0.24 respectively).

Comparing indoor and outdoor participants across accuracy of all markers placed, indoor participants showed an error of 4.14m (SD:4.69m) and outdoor participants with an error of 3.27m (SD:3.00m). This was a significant improvement in accuracy ($p < 0.04$) for the AR system. We also determined the average answer for each group (these have been adjusted against the actual answer). A positive number indicates the participants gave answers that were further along the wall than the actual marker, and negative values indicate closer. As the average results for both groups were positive (indoor=3.76m and outdoor=1.50m), it indicates participants believe the markers to be further from the camera/viewpoint than they actually were.

Height values were compared separately. Indoor participants were slightly more consistent at estimating height. They were on average 0.43m (SD:0.36m) from the correct answer, while outdoor participants were 0.56m (SD:0.50m) ($p < 0.01$). However when observing the average value (ie the average of answers compared against the correct heights), indoor participants selected average of 0.25m below the actual height (SD:0.50m) while outdoor participants had an average of only 0.01m above the marker (SD:0.75m). This is a statistically significant difference ($p < 0.001$).

For the scenario task the results were the opposite of what we expected. The indoor participants produced an average error of 1.51 windows (SD:1.25) (the windows were on average 1.8m apart from each other) while the outdoor group had a higher error of 1.98 (SD:1.50) ($p < 0.05$).

5.2 Timing

The time taken to complete the task did not drop from single video to the double video task for the indoor participants (p=0.43). However, there was a significant time decrease for the outdoor participants from 3m:57s (SD:1:36) to 2:32 (SD:1:36) ($p < 0.002$).

For the scenario task there was a significantly lower time shown when using the AR system (2:10, SD:0:50) over the indoor desktop (2:49, SD:1:10) ($p < 0.05$).

5.3 Questionnaire results

All participants were asked to complete a questionnaire after completing the study. There were 11 common questions asked to participants from both indoor and outdoor groups and an additional 4 questions asked to the outdoor group. The questions were answered using a 5-point Likert scale from 1 (easy, agree, low) to 5 (hard, disagree, high). The results of the questionnaire are shown in Table 1. The results were analyzed using a Mann-Whitney U test for independent samples. Significant differences ($p < 0.05$) are shown in bold.

Table 1 - Questionnaire Results. Significant differences (Mann-Whitney U test: $p < 0.05$) are shown in bold.

	Indoor	Outdoor
I felt I completely understood the layout of the locations presented to me.	2.00	2.00
The speed of the videos was (1=too slow, 5=too fast)	**3.12**	**2.35**
Understanding a single video was	2.06	2.00
Understanding double videos was	**3.12**	**2.00**
Understanding the scenario task video was	**2.41**	**1.76**
The scenario itself was hard to understand	3.88	4.24
The resolution of the display was sufficient for the task	**1.47**	**2.82**
The technology was helpful to complete the tasks	1.60	1.69
The viewpoints available were appropriate	2.00	2.20
The technology was intuitive to use	1.93	1.88
Being able to move around was helpful	N/A	1.12
Do you believe the tasks would have been easier if the videos were shown normally on a television screen?	N/A	3.31
The use of transparency was confusing	N/A	3.81
Rate the strain on your eyes	**1.82**	**2.65**
Rate the strain on your back/hips	N/A	2.41

5.4 Feedback

In addition to the fixed questions on the questionnaire form, participants were asked to provide feedback on each of the three tasks and comment on the system. 13 of the 17 outdoor participants provided written feedback after the study. Participants made a range of comments on tracker jitter and HMD visibility which are caused by the hardware and difficult to avoid. The wire-frame model was sometimes difficult to make out on the HMD. Using thicker lines would overcome this.

One interesting topic discussed in the feedback related to difficulty differentiating between the view of the real world and the occluded object overlays. When the first-person view is combined with the occluded scene, any similar colors or textures between the occluded and occluding areas can make it difficult to discern where the separation between the two lies. Participants suggested the ability to remove or reduce the visibility of the real world to help in viewing the occluded area. This would have the effect of switching between an AR and VR view. A more obvious boundary between the visible and occluded areas on the display was also suggested. A border could be displayed on the screen around the occluded information, with thick lines and bright colors. In this way the edge between the occluded and visible environment would be obvious.

Although the scene was being updated by a simulated live video feed, some users felt it was too slow and would have preferred the information be presented as quickly as possible. This was reinforced by the questionnaire which showed a significant difference in option of the video speed between the groups ($p < 0.005$).

The known difficulty in judging depth of occluded objects in AR [6] was reflected in participants' comments. Some participants suggested alternate ground plane rendering techniques to assist with depth judgment. These included aligning the ground grid to real buildings, extending the edges of occluded objects, and providing more on-screen distance information. There were requests for alternate viewpoints of the virtual scene. Our system is capable of rendering these but they were disabled to simplify the study.

6 Discussion

The results indicate that the outdoor users were more accurate at locating the positions of the markers on a map compared to the indoor group. However, across both single and double video tasks the time taken was much longer. So *Hypothesis 1* and *Hypothesis 2* were partially valid.

We suspect that the longer times could be avoided with additional training on the AR system. We observed learning effects between the first and second task by the significant time decrease. We can assume that the participants were learning and that the second task was not simply 'easier', as on the questionnaire the outdoor group indicated that the single and double video tasks were of equal difficulty (identical results for Q3 and Q4). This suggests that the difference between single and multiple videos using the AR system is negligible. We believe this finding would scale to a larger number of videos. However, further research would be needed to determine how well users can interact with a system with very large numbers of simultaneously updated videos. The significantly different responses between Q3 and Q4 in the questionnaire ($p<0.001$) from the indoor group suggests that observing multiple videos on a desktop is more difficult than only one. This is supported by the lack of a time decrease between tasks for the indoor group as the task learning is countered by increased difficulty.

The third task was designed to exploit the benefits of AR; being able to intuitively compare the real and virtual worlds by looking at them. The indoor participants performed surprisingly well considering the provided maps were of very low resolution. 47% of indoor participants providing answers that were no more than 1 window from the correct answer. For the outdoor group there was only a single participant to obtain this result. We believe the lack of accuracy for the outdoor groups was primarily caused by tracking problems. The outdoor participants' accuracy at aligning real and virtual objects can never exceed the accuracy of the GPS and orientation sensors. At the distance of 50m from the markers even minor orientation errors resulted in large rendering offsets. As outdoor tracking technology improves these problems should be reduced. The time taken to complete the task was significantly lower than the indoor group. So *Hypothesis 3* was partially correct. The optimal solution so such a task may be an AR system to enable quick results, combined with GPS assisted mapping for accuracy.

7 Conclusion

We have presented an evaluation of our see-through vision system that demonstrates that AR is a viable alternative to viewing videos of occluded objects or areas. We conducted a study to compare speed and accuracy of finding markers between see-through vision outdoors and watching unaltered videos on a traditional display. The outdoor participants demonstrated a higher accuracy judging the position of the markers but took a longer time than those indoors with traditional display. Results indicated the see-through vision system makes observing multiple videos significantly easier to understand than when displayed simultaneously on a monitor. At more complex tasks requiring users to compare and align two real-world locations, outdoor users were faster, but tracker error caused a lower accuracy.

References

[1] B. Avery, W. Piekarski, and B. Thomas. Visualizing Occluded Physical Objects in Unfamiliar Outdoor Augmented Reality Environments. In *6th Int'l Symposium on Mixed and Augmented Reality*. p 285-286. Nara, Japan 2007.

[2] B. Avery, W. Piekarski, J. Warren, and B. Thomas. Evaluation of User Satisfaction and Learnability for Outdoor Augmented Reality Gaming. In *7th Australasian User Interface Conference*. p 17-24. 2006.

[3] R. Bane and T. Höllerer. Interactive Tools for Virtual X-Ray Vision in Mobile Augmented Reality. In *3rd Int'l Symposium on Mixed and Augmented Reality*. p 231-239. Arlington, VA, USA 2004.

[4] J. Hu, S. You, and U. Neumann, *Texture Painting from Video*. Computer Graphics, Visualization and Computer Vision, 2005. 13 p 119-125.

[5] Y. Kameda, T. Takemesa, and Y. Ohta. Outdoor See-Through Vision Utilizing Surveillance Cameras. In *3rd Int'l Symposium on Mixed and Augmented Reality*. p 151-160. Washington DC, USA 2004.

[6] M. A. Livingston, J. E. Swan, J. L. Gabbard, T. Höllerer, D. Hix, S. Julier, Y. Baillot, and D. Brown. Resolving Multiple Occluded Layers in Augmented Reality. In *2nd Int'l Symposium on Mixed an Augmented Reality*. p 56-65. Tokyo, Japan 2003.

[7] U. Neumann, S. You, J. Hu, B. Jiang, and J. Lee. Augmented Virtual Environments (AVE): Dynamic Fusion of Imagery and 3D Models. In *IEEE Virtual Reality*. p 61-67. Los Angeles, CA, USA 2003.

[8] W. Piekarski and B. Thomas. Interactive Augmented Reality Techniques for Construction at a Distance of 3D Geometry. In *Immersive Projection Technology / Eurographics Virtual Environments*. p 19-28. Zurich, Switzerland 2003.

[9] H. S. Sawhney, A. Arpa, R. Kumar, S. Samarasekera, M. Aggarwal, S. Hsu, D. Nister, and K. Hanna. Video Flashlights - Real Time Rendering of Multiple Videos for Immersive Model Visualization. In *13th Eurographics Workshop on Rendering*. p 157-168. Italy 2002.

[10] M. Tory, A. Kirkpatrick, M. Atkins, and T. Möller, *Visualization Task Performance with 3D and Combination Displays*. IEEE Transactions on Visualization and Computer Graphics, 2006. 12(1) p 2-13.

[11] T. Tsuda, H. Yamamoto, Y. Kameda, and Y. Ohta. Visualization Methods for Outdoor See-Through Vision. In *15th Int'l Conference on Artificial Reality and Telexistence*. p 1-8. Christchurch, NZ. 2005.

[12] Y. Wang, D. M. Krum, E. M. Coelho, and D. A. Bowman, *Contextualized Videos: Combining Videos with Environment Models to Support Situational Understanding*. IEEE Transactions on Visualization and Computer Graphics, 2007. 12(6) p 1568-1575.