Evaluating Wide-Field-of-View Augmented Reality with Mixed Reality Simulation

Donghao Ren* Tibor Goldschwendt† YunSuk Chang‡ Tobias Höllerer§

University of California, Santa Barbara

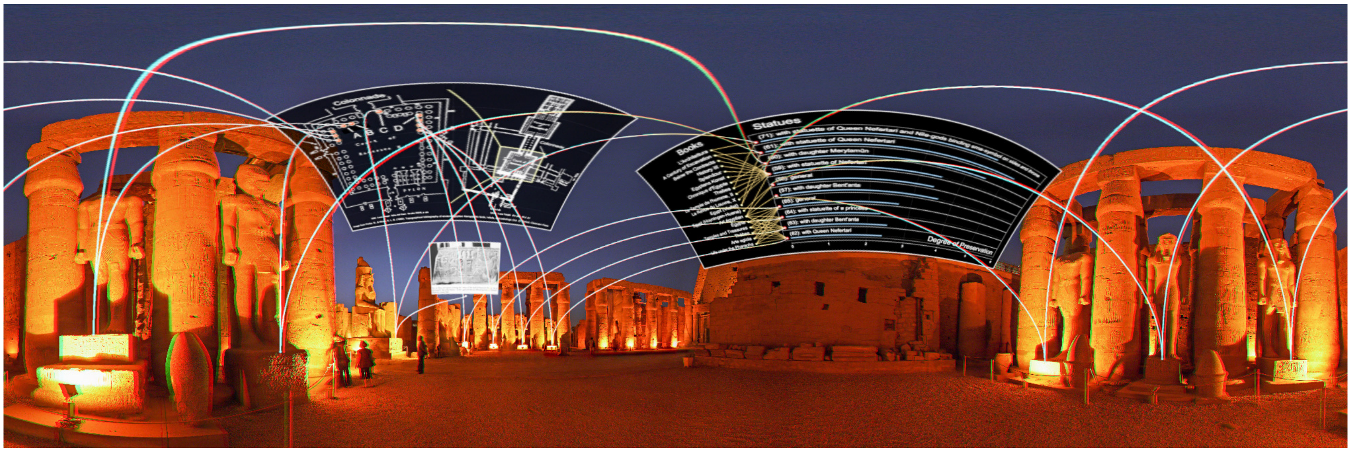


Figure 1: The Luxor Temple scene used in our experiment, here shown in red-cyan anaglyph mode with equirectangular projection (the ground part cropped for better fit here). The annotations consists of a chart, a map and a photograph, as well as 3D links among them and the statues. Users experience full and small field-of-view augmented reality views into this scene, with and without mild tracking artifacts, and perform information-retrieval tasks relating to the charts and the scene. The stereo panorama of the Luxor Temple was provided by Tom DeFanti, Greg Wickham, and Adel Saad, with stereo rendering by Dick Ainsworth [1, 35].

Abstract

Full-surround augmented reality, with augmentations spanning the entire human field of view and beyond, is an under-explored topic since there is currently no hardware that can support it. As current AR displays only support relatively small fields of view, most AR applications to-date employ relatively small point-based annotations of the physical world. Anticipating a change in AR capabilities, we experiment with wide-field-of-view annotations that link elements far apart in the visual field. We have built a system that uses full-surround virtual reality to simulate augmented reality with different field of views, with and without tracking artifacts. We conducted a study comparing user performance on five different task groups within an information-seeking scenario, comparing two different fields of view and presence and absence of tracking artifacts. A constrained field of view significantly increased task completion time. We found indications for task time effects of tracking artifacts to vary depending on age.

Index Terms: H.5.1 [Information Interfaces and Presentation (e.g., HCI)]: Multimedia Information Systems—Artificial, augmented, and virtual realities; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual reality

*e-mail: donghaoren@cs.ucsb.edu
†e-mail: goldschwendt@cs.ucsb.edu
‡e-mail: ychang@umail.ucsb.edu
§e-mail: holl@cs.ucsb.edu

1 Introduction

Augmented Reality (AR) has been a thriving topic within the field of human-computer interaction in recent years, and its promise is felt both in terms of academic research [7] and industry (e.g., Microsoft HoloLens, MagicLeap, etc.). AR provides an intuitive and direct connection between the physical world and digital information. Application areas range from consumer smartphone apps annotating the starry sky to intricate augmentation gear providing simulations of training scenarios in the real world [29]. Implementing augmented reality imposes complex technical challenges, such as registration in the real world, real-virtual object occlusion, or device wearability. In this paper, we investigate the impact of two fundamental technical parameters of AR devices on user task performance in a tourism-inspired information seeking scenario: field of view (FOV) and tracking artifacts (jitter and latency).

Full-surround AR with augmentations extending beyond a small window around a point of interest, or even beyond the entire human field of view, is an under-explored topic, since there is currently no practical viewing hardware that can support it. Because of this limitation, most AR applications to-date employ relatively small point-based annotations of the physical world, such as textual labels, notes, or arrows.

As technology evolves, AR devices will provide wider fields of view that hopefully will cover the entire human field of view [17, 26, 13]. Anticipating such a change in the not-too-distant future, we investigate AR tasks involving wide-field-of-view visualizations and annotations that link elements far apart in the visual field. AR designers and device manufacturers should be aware of the potential impact that important immersion parameters [8] have on such design choices. We demonstrate a simulated wide-field-of-view AR user interface that met with widespread acclaim by users and we evaluate the impact of field of view and tracking accuracy on user performance in different task groups within a simulated augmented art-historic tourist site scenario.

It is hard to run experiments to test future capabilities when current hardware does not support the envisioned applications. One way to get around this difficulty is to use virtual reality to simulate AR and run a controlled experiment. We review the existing literature on AR simulation in the following Section.

We have built an AR simulation system using the AlloSphere [21], a large-scale full-surround immersive VR system, implementing AR with varying field of view and levels of tracking artifacts. Our system supports interactive creation and placement of augmented content, which greatly facilitated design and preparation of the task environment for experiments.

Using this system, we designed an augmented scene around a stereo panorama of a major courtyard within the Luxor Temple complex (see Figure 1). The annotations of the scene consist of a chart, a map, and a photograph, which are connected to the statues of the temple via extensive 3D links. We intentionally designed the annotations to cover big portions of the field of regard, even exceeding the human visual field. This kind of large-scale augmentation is typical of several types of augmented reality applications that are commonly used to motivate the usefulness of AR and stir the public's imagination. A recent HoloLens demo involved large-scale annotations, such as virtual movie screens filling entire wall segments, and close-by inspection of life-size objects such as a virtual motorcycle. A mobile AR theme park could present larger-than-life dinosaurs or virtual skyscrapers in the user's field of view. Or, related to our test scenario, the full reconstruction of ancient ruins could be presented at an actual archaeological site, including large amounts of associated cultural history data.

In order for such visions to be truly effective, an AR system with a wide field of view would likely be desirable. Anecdotal evidence of user reactions to recent AR demonstrations involving a *limited* FOV reveals user dissatisfaction with the broken illusion of virtual objects being real and with having to turn one's head to scan the entire augmented scene [20].

But what is the actual impact on user performance in representative AR tasks? We designed a controlled experiment with two different fields of view (full-surround field of view, only limited by the 108×82 degree stereo shutter glasses versus a smaller 45×30 degrees field of view), with or without mild tracking artifacts. Participants were asked to perform information seeking tasks as part of a cultural history tourist scenario.

We first conducted a pilot experiment with 9 participants, and then a main experiment with 33 participants. We found that FOV had a significant impact on task completion time in both the pilot experiment and main experiment. While the pilot experiment indicated a possible effect of tracking artifacts, we did not observe this in the main experiment. The main experiment had a group of more diverse participants in terms of age, gender, discipline of study and education level. A deeper look into the results revealed that older people tend to be more affected by tracking artifacts than young people. In general, both field of view and tracking artifacts had a significant impact on head movement speed.

2 Related Work

Conducting AR user studies is challenging. With AR technology still far from a standardized technology, display devices suffer from low resolution, low field of view, latency, improper occlusion, and other perceptual problems [25]; the capabilities of devices differ a lot from one to another, and many AR immersion parameters are currently not achievable, especially in the mobile AR domain [28].

Usability evaluations in AR have been performed regarding questions of perception, performance and collaboration [37] using a range of evaluation techniques [12, 19]. In our experiment, we employ objective measurements on task time, correctness and head movement speed, subjective measurements from pre- and post-study questionnaires and informal feedback from participants.

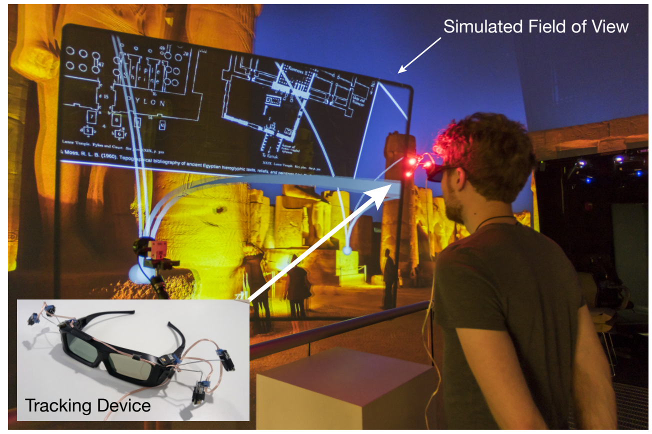


Figure 2: Simulating AR with a small field of view. We let the viewport follow the user's head movement. The user only see a small part of the annotations. The border of the viewport simulates the border of the display of an AR device.

There are currently no commercial solutions yet for truly wide FOV mobile AR, even though it has been the focus of both academic and industrial research [17, 26, 13]. Likewise, low latency and high-precision tracking is still a challenging problem, particularly for mobile AR. Impressive progress has been made at some cost of hardware bulk and expense [29].

2.1 Simulation of Novel Technologies

It is fairly standard practice to first simulate a novel display technology before prototyping and implementing it. For instance, State et al. [36] built a simulator for their HMD display. Arthur et al. [3] simulated several head-worn display concepts for NASA. Mixed reality simulation [32, 9] uses virtual reality to simulate virtual or augmented reality experiences. It can be done using a head mounted display, or in a CAVE or similar high-end surround-view VR environment.

The validity of AR simulation is an important topic for investigation. Lee et al. [23] studied the role of simulator latency in the validity of AR simulation. They found that for a path tracing task simulator latency does not interact with artificial (simulated) latency and that there is instead an additive effect on task performance. All significant findings from an experiment from the research literature were replicated in simulation. In a later experiment [24], Lee et al. studied the effects of visual realism on search tasks and found that task performance did not significantly vary among different visual realism settings.

2.2 Virtual and Augmented Reality Experiments

It is also important to consider the results of previous VR experiments on immersion factors [8], as they can provide valuable hints for AR experiments. There are many VR studies involving wide FOVs. Arthur [4] studied the effects of FOV on task performance and participant comfort. Covelli et al. [11] experimented FOV in pilot performance in flight training scenarios. Jones et al. [14] found that in medium-field VR, peripheral vision is an important source of information for the calibration of movements. Ragan et al. [33] examined the effects of varying FOV in terms of a visual scanning task as part of virtual reality training. Higher field of view led to better training performance. The results generally indicate that higher

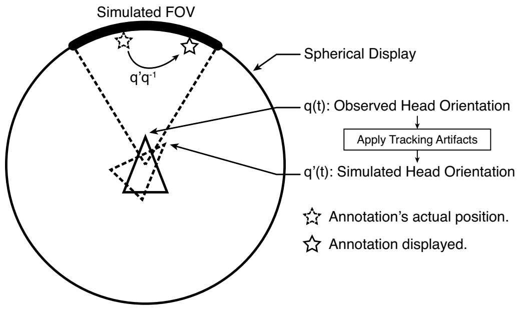


Figure 3: Simulating an augmented reality device with virtual reality in a full-surround display.

FOV leads to better performance. However, Knapp et al. [18] found that limited FOV of HMDs is not the cause of distance underestimation in VR. Ball et al. [5] showed that the opportunity for physical navigation is more important than increased FOV for display walls.

Tracking artifacts including latency and jitter are also of research interest in virtual reality. Ventura et al. [39] studied the effects of tracker reliability and field of view on a target following task using an HMD as the simulator. This study examined the same overall types of immersion parameters as our work here, but wide-FOV AR was not considered, as the maximum FOV was limited by the simulator HMD, and tracking artifacts were limited to interruptions of varying length. A low FOV proved detrimental to user performance, as did prolonged sensor dropout periods. Buker et al. [10] studied the effect of latency on a see-through HMD and determined that simulator sickness could be reduced by predictive tracking compensation.

Our understanding of FOV in AR is still limited [16]. Researchers have explored the effect of field of view with actual AR prototype devices. Van Nguyen et al. [38] developed a wide FOV AR display. Using a similar setup, Kishishita et al. found that a distribution of annotations in the peripheral vision decreases target discovery rate [16], and task-performance differences between in-view and in-situ annotations become smaller as the FOV approaches 100 degrees and beyond [15]. Although they explored wide FOVs, their annotations are relatively small and point-based, compared to what we used in our experiment.

Our work, relating to the findings from the literature, constitutes the first time that the effect of constraints in FOV and tracking artifacts were examined for AR scenes involving annotations spanning a wide field of view, even exceeding the human visual field.

3 Study System

We have designed and implemented a system that allows us to interactively create content in a full-surround display environment (as illustrated in Figure 2 and Figure 4). We used the system to create the annotations for our experiment.

Our system utilizes Ren et al.'s iVisDesigner [34], a data-driven vector graphics editing software, in which users are not only able to create graphical primitives but can also bind them to data items and use data attributes to control graphical attributes. Each 2D visualization designed in iVisDesigner is rendered as a flat rectangle in the 3D virtual space. Users can use a tablet-based interface to freely arrange them. 3D linking between the visualizations and the background scene is also supported. We employed an Illuminated Steamlines [40] approach to render the links, and also highlighted the endpoints of the links for better visibility. The light source can be either fixed (e.g., at the center) or moved around the user at a constant speed.

In a 3D VR setting like for our experiment, 2D visualizations can be displayed as overlays on top of an environment such as a stereo panorama. Links can be drawn from visualizations to 3D locations in the virtual environment as well. In the context of AR, this can be considered as a simulation where the background environment simulates the real world and the visualizations are what the users would see through an AR device. Simulation of AR experiences is useful for conducting user studies investigating user interfaces or display parameters when an actual AR device is not yet available.

3.1 Simulating Augmented Reality Parameters

We designed an AR simulator for use with the system. It can simulate two important parameters for AR devices: field of view and tracking artifacts.

3.1.1 Simulating Field of View

Unlike head-mounted AR/VR devices, our display is fixed, surrounding a user standing in the middle of it. Therefore, in order to simulate a small augmentation field of view, we have to know the head orientation with respect to the display. Head tracking was performed by a PhaseSpace Impulse X2 system [30] which uses active optical tracking. We installed 6 PhaseSpace LEDs on a pair of stereo glasses, which is worn by the participant during our experiment (see Figure 2). We carefully placed the LEDs such as not to interfere with the user's maximum field of view. From the head tracking information, we can easily derive which part of the annotations is to be shown to the user and which part is to be hidden given a desired field of view.

We display a black rectangle frame at the border of the field of view to inform the user that the augmentations end there. Figure 2 shows an example of a limited field of view.

3.1.2 Simulating Tracking Artifacts

Simulating tracking artifacts is a bit more involved than simulating field of view. Since our surround display is fixed, we have to simulate the latency and jitter of augmentations as if a display were worn by the user based on their head movement. Let the true head orientation be denoted by the quaternion $q(t)$; the head orientation estimated by the simulated AR device be denoted by the quaternion $q'(t)$, then the displacement of the augmented content shown on the screen is given by $q'(t)q(t)^*$, as illustrated in Figure 3[1].

When simulating an AR device with perfect tracking, we have $q(t) = q'(t)$, thus $q'(t)q(t)^* = 1$, so there is no displacement of the augmentations on the full-surround display.

Some latency is introduced by the PhaseSpace tracking system. Therefore, the latency we simulate is delayed by the PhaseSpace system latency. However, when simulating perfect tracking, since there is no displacement of the augmented content at any time, the PhaseSpace latency has no effect on augmentations (only on the exact placement of the FOV frame in case of a limited FOV) and augmentations are rock-solid. In addition, previous work on the validity of AR simulation [23] suggests that simulator latency is an additive effect on artificial latency, thus we can consider them together. We measured the end-to-end latency of the system (the sum of tracking system latency, network latency and rendering latency) using a high-speed camera at 100fps. The result is 80 ± 20 ms.

3.2 Limitations

While the PhaseSpace system tracks 6 degree of freedom head pose, for our study we decided to assume a single viewing position in the center of our spherical display. This is because the stereo panorama that acts as the real world cannot respond to position changes. Hence, the scene will be distorted if users move away

[1] Here we use quaternions to represent rotations. Readers unfamiliar with quaternions can treat them as 3×3 rotation matrices, and treat $q(t)^*$ as matrix inversion.

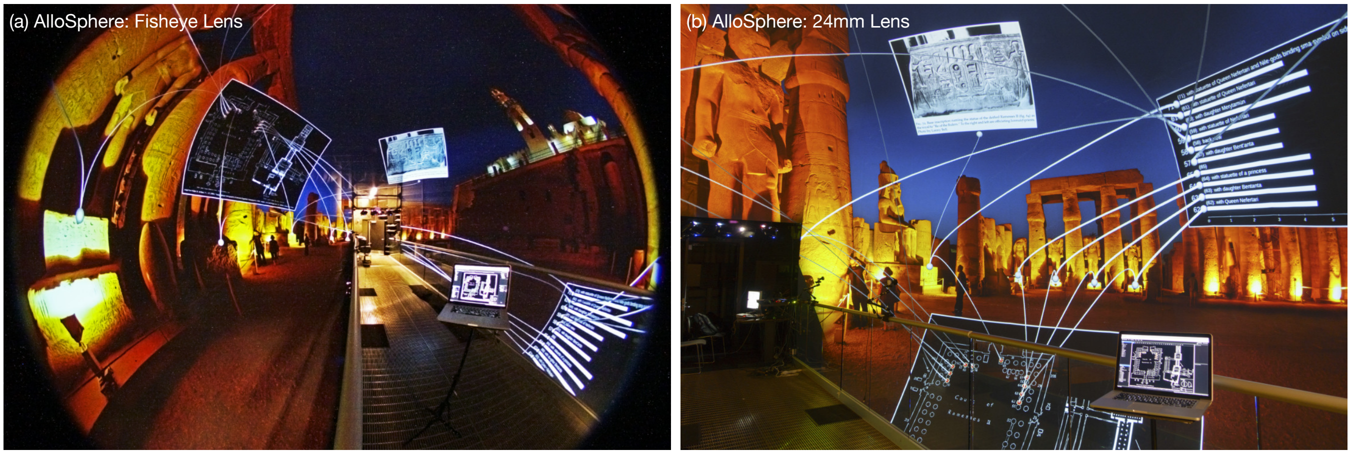


Figure 4: Annotations on a stereo panorama taken by Tom DeFanti et al. at the Courtyard of Ramses II, part of the temple of Luxor in Egypt. (a) Full-surround version in the AlloSphere, captured using a fisheye lens. (b) Captured using a 24mm lens.

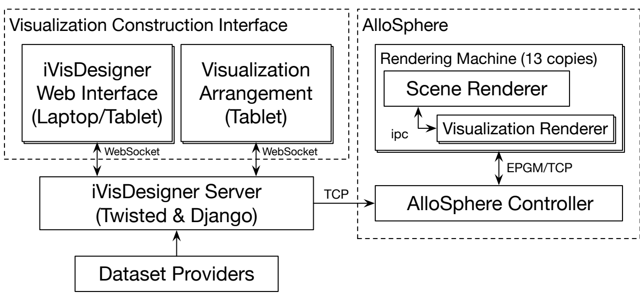


Figure 5: The architecture of our content creation system.

from the center. However, the surround display, with a 10m diameter has a fairly big sweet spot, so distortions are very limited on small movements. This setup is however not suitable for tasks that require walking. In our experiment, we placed all the annotations around the participant, and instructed the participant to stay close to the center of the sphere.

Likewise, with this setup, it is hard to display objects that are very close to the participant. This is because we cannot guarantee that the user's head is at the exact center of the sphere, and position inaccuracies are more pronounced when the object is very close to the user. Hence, in our experiment, we avoided annotations that are too close to the participant.

The FULL FOV condition we simulated is partially occluded by the frame of the 3D shutter glasses, which are measured to have an approximately 108×82 degree FOV. Users can still see part of the augmented scene in full human FOV underneath and beyond the frame, but the view is more limited than by most vision-correcting eyewear.

The simulation of the real world is currently presented as a still stereo panorama. No motion effects were employed for real-world simulation in this study.

3.3 Implementation Details

The software architecture of our system is shown in Figure 5. iVisDesigner consists of a web-based interface for visualization construction and a web server using Twisted and Django. In our content creation system, the web server also acts as a gateway to the display system infrastructure.

The display system currently has 13 rendering servers, each connected to two stereo projectors driving the surround display. Our system performs the synchronization among rendering servers to ensure that each projector shows the right content at the right time.

The augmentation rendering system is mainly written in JavaScript, running under Node.js with custom bindings to OpenGL and the Skia graphics library. This allows us to use the rendering and state synchronization code from iVisDesigner.

Additionally, we have created a testbed program for normal desktop computers, which allows us to view the visualizations for debugging purposes or demonstrate our system to others without being inside the virtual environment. It can render the 3D scene in either perspective or equirectangular projection mode, with either monocular, anaglyph, or dual viewport stereo modes.

4 Experiment Design

In this section, we present and discuss the design of our experiment. Our dependent variables were task time and answer correctness, and we also recorded all user head motion for further analysis.

4.1 Environment

For the 'physical world" portion of our experiment, we employed a high-resolution OmniStereo panorama taken at the Luxor Temple in Egypt by DeFanti et al. [1, 35]. This scene serves as the "real-world backdrop" for wide-spanning annotations that link real-world objects (i.e., the statues in the scene) to charts. As shown in Figure 1, we created three charts in the scene and linked the statues to the charts with 3D lines. The charts contain descriptions of each statue (e.g., "Ramses II with daughter Bent'anta"), associations between books and statues, as well as statue numbers and labels. Participants can follow the lines from a chart to a statue or vice versa, thus allowing them to follow instructions such as "Locate the statue with daughter Bent'anta." or answer more complex questions, such as "How many headless statues are associated with this book?," referencing an art history book listed in the chart.

The content of the scene was designed around archaeological facts. The map of the temple, the descriptions of the statues and the names of the books came from *Topographical bibliography of ancient Egyptian hieroglyphic texts, reliefs, and paintings* [31]. The photograph was from *Temples of Ancient Egypt* [2]. We altered the associations between the statues and the books in order to design tasks with similar difficulties.

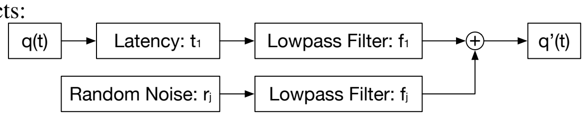
4.2 Factors

Our experiment had a three-factor within subject design.

Field of View We used two conditions for the FieldOfView factor: FULL and 45×30. The FULL field of view covers the whole field of regard (constrained by the shutter glasses), where users have both central and peripheral vision. The 45×30 field of view covers a field of view of 45 degrees wide and 30 degrees high.

Tracking Artifacts For the TrackingArtifacts factor we used two conditions: NONE and MILD tracking artifacts. We experimented with perfect tracking without any latency and jitter, and MILD tracking artifacts with a small amount of latency and jitter akin to those in state-of-the-art AR devices.

The following diagram shows how we simulate tracking artifacts:



We used parameter values of $t_1 = 0, f_1 = 10, f_j = 0.1, r_j = 0.1$. These parameters yield mild tracking artifacts, determined by AR experts to be representative for high-performance AR devices. We settled on this representation after extensive experimentation. Note that even though t_1 is set to 0, the lowpass filter is introducing a realistic latency effect. Refer to the supplemental video for an impression of the visual effects of these simulated tracking artifacts.

Tasks Groups We designed the tasks as five task groups of four similar questions each. There are two reasons we considered the TaskGroup a factor. First, as the scene only contains 12 statues, it is hard to design questions that are independent to each other such that participant will not remember which line links to which statue. Therefore, we grouped the questions into groups of four. Second, considering the task group a factor provides more statistical power since some task groups are harder than others.

4.3 Tasks

The task groups were designed as follows:

1. F1EC: Follow 1 link, easy chart question. Example: "Where is statue No. 62? Locate the statue in the scene." This task involves looking at the chart, finding the number "62", and following the link to the actual statue. Users answer the question by looking at the statue in question and pressing a button on a mobile device.

2. F1M: Follow 1 link, map question. Example: "Which statue is tagged with letter A? Locate the statue in the scene." This task involves finding the letter "A" on the map and following the link to the actual statue.

3. F1MC: Follow 1 link, moderate chart question. Example: "Which statue has the smallest number of associated books? Locate the statue in the scene." This task involves looking at the chart and finding the statue with the smallest number of associated books, and following the link to the actual statue.

4. F2MC: Follow 2 links, moderate chart question. Example: "Locate all the statue(s) that are described in the book A Century of Excavation." In this task, participants have to find the named book in a list, follow association links to the list of statues, and then follow the links to the actual statues.

5. FB1SC: Follow 1 link back to chart, simple chart question. Example: "What is the number of THIS (ask your supervisor) statue?". This task involves following the link from the statue to the chart and finding the number of the statue.

We also interspersed eight extra tasks that ask for other information from the scene, as we wanted to let the participant experience more aspects of the scene not limited to the five task groups. These eight tasks varied in difficulty, e.g., "Is the book *Thebes* associated with at least one statue with a head?", "How many books are associated with at least one headless statue?". We included these eight tasks in the experiment in order to make the questions more diverse and reduce the chance that participants remember which line links to which statue.

In order to control for the learning effect and the difference between questions inside each task group, we randomly permuted the four combinations of FieldOfView and TrackingArtifacts within each task group, and then shuffled the order of the questions together with the eight extra tasks for each participant. Therefore, each participant went through all the questions, but experienced a different order of questions, and each question was tested on different combinations of FieldOfView and TrackingArtifacts.

4.4 Apparatus

Our apparatus was a large-scale full-surround immersive VR projection system consisting of a three-story high full-surround display environment driven by 26 active stereo projectors. The scene was rendered with 13 rendering machines, each equipped with two NVIDIA Quadro K5000s.

In order to minimize the efforts to read the questions and provide the answers, we employed a hand-held mobile device (Samsung S4 smart phone with custom software) for participants to perform the tasks. The mobile device was wirelessly connected to our system so that tasks and questions could be automatically presented to the user.

4.5 Pilot Experiment

We recruited 9 participants working in or close to our lab for a pilot experiment before we started the main experiment. These participants were familiar with the AlloSphere (e.g., they had experienced it multiple times before, with significantly different content).

During the pilot run, we monitored the participants and refined the experiment design. One important issue turned out to be lack of familiarity with the Luxor Temple environment. After observing people getting confused about where to look, we decided to include a short introduction to the environment, and let participants look around and understand that there were statues, charts and links between them. We also refined the wording of the questions to reduce confusion. Moreover, we decided to let the participants read out the questions before starting to seek the answer, which helped the experimenter control the pace of the experiment and identify potential issues.

In addition, we collected feedback about the length of the experiment and whether or not the participant experienced any dizzyness during the experiment. The whole experiment process took around 20 to 40 minutes.

We analyzed the data from the pilot study with a three-factor ANOVA analysis and found significant effects on both FieldOfView and TrackingArtifacts, see Figure 6 for the ANOVA table and confidence intervals. Given the interesting results from the pilot study, we formed our hypotheses, and conducted the main study with more participants.

4.6 Participants and Procedure

After refining and formalizing our study procedure, we recruited 33 participants for the main experiment.

The study was conducted on a one-by-one basis. First, we introduced the purpose of the study to the participant, and collected background information in a pre-study questionnaire.

Then, we explained the setup and calibrated tracking. Before starting the actual tasks, we let the participant get familiar with the

	F value	$Pr(>F)$	
TaskGroup	54.809	< 0.0001	***
FieldOfView	5.623	0.0191	*
TrackingArtifacts	5.885	0.0166	*
TaskGroup:FieldOfView	0.342	0.8491	
TaskGroup:TrackingArtifacts	0.320	0.8644	
FieldOfView:TrackingArtifacts	2.293	0.1323	

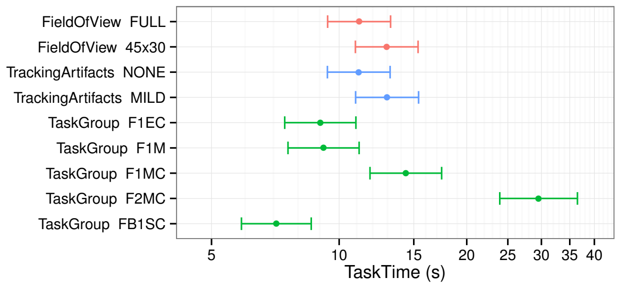


Figure 6: Pilot Study LogTaskTime: (Top) Analysis of Variance Table of type III with Satterthwaite approximation for degrees of freedom. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. (Bottom) Least Squares Means with 95% Confidence Intervals. Values are converted from the log scale to seconds.

scene by introducing the Luxor Temple scene and pointing out the statues, visualizations and links between them. The participant was then asked to perform two simple warm-up tasks to get familiar with the mobile device for reading and answering questions. During these initial tasks, we monitored the participant and answered his or her questions.

Once the actual tasks began, the experimenter sat nearby, minimizing any disturbance of the participant. For each task, we first let the participant read the question out loud for the experimenter to follow along. Once the participant understood what to do for the question, s/he pressed the "Start" button on the mobile device to start task timing. Scene augmentations only appeared after the start button was pressed. Once the participant arrived at their answer, s/he pressed another prominent button and then recorded his or her answer. For tasks that asked users to "Locate" particular statues, the participant recorded the answer by pressing a button while looking at the respective statue.

During the experiment, we took note of any misunderstandings of the questions (reported by the participant or observed by the experimenter), technical issues (e.g., tracking system accidentally stopped working), or exceptions during the study (e.g., a fire drill causing a 15 minute interruption of one participant).

After completing the experiment, we had the participant fill in a post-study questionnaire. Each participant was compensated with $10 for the study, which altogether lasted approximately 45 minutes to an hour per participant.

5 Results

In this section, we present our analysis of the study results, including statistical analysis of task time and correctness, and descriptive analysis of head movement data and questionnaire responses.

5.1 Data Processing

To do a meaningful analysis, it is necessary to reject outliers and remove problematic runs. We removed three participants: one who experienced an interruption of the experiment by a fire drill; one who accidentally closed the mobile application that resulted in data loss and disturbance of the order of the questions; and one who was not able to follow the instructions throughout the study. After the removal of three participants, we had 30 participants (10 female and 20 male) as part of our analysis. The average age was 25, ranging from 19 to 35 years old.

	F value	$Pr(>F)$	
TaskGroup	259.738	< 0.0001	***
FieldOfView	14.062	0.0002	***
TrackingArtifacts	1.261	0.2621	
TaskGroup:FieldOfView	0.817	0.5148	
TaskGroup:TrackingArtifacts	1.658	0.1585	
FieldOfView:TrackingArtifacts	1.693	0.1938	

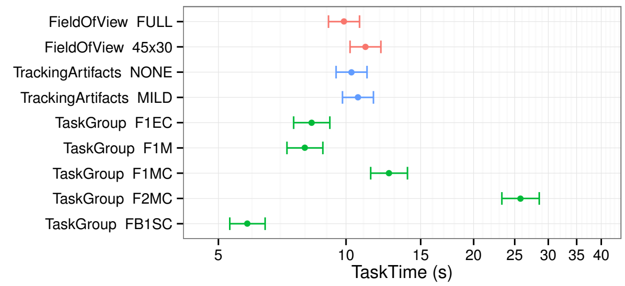


Figure 7: Main Study LogTaskTime: (Top) Analysis of Variance Table of type III with Satterthwaite approximation for degrees of freedom. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. (Bottom) Least Squares Means with 95% Confidence Intervals. Values are converted from the log scale to seconds.

Table 1: Follow-up analysis considering Gender and AgeGroup as factors. Both have a significant impact on TaskTime, and there is a significant interaction effect between TrackingArtifacts and AgeGroup. Only significant effects are shown.

	F value	$Pr(>F)$	
TaskGroup	214.914	< 0.0001	***
FieldOfView	10.910	0.0010	**
Gender	3.093	0.0903	.
TaskGroup:AgeGroup	2.219	0.0659	.
TrackingArtifacts:AgeGroup	3.655	0.0565	.

We looked up the experimenter's record of each participant and removed all the problematic tasks we recorded, including: 1) the participant pressing the start button without understanding the question; 2) the participant accidentally pressing the "ready to answer" button without having completed the task; 3) minor system issues, such as tracking interruptions requiring a restart of the tracker; 4) other interruptions, such as the participant being unsure about the task and querying the experimenter during the timing period.

We only considered tasks that were correctly answered within 60 seconds, which is a reasonable filter, since tasks that took too long might have been the result of misunderstanding of the question, or issues of the system that were not reported by the participant.

After all filtering, there were 571 tasks (total is 600 = 30 participants times 5 task groups times 4 tasks per group) left for analysis.

5.2 Task Time

Our experiment was set up as a three-factor within-subject design. Since we removed a set of data points by the procedure stated above, there were missing values with an irregular pattern. We first examined the distribution of TaskTime under different combinations of factors, and found that it was skewed towards zero. The log-normal distribution being a reasonable model of task time, we transformed the TaskTime into LogTaskTime using the natural logarithm. The distribution of LogTaskTime approximated a normal distribution, leading us to ANOVA analyses on LogTaskTime. We

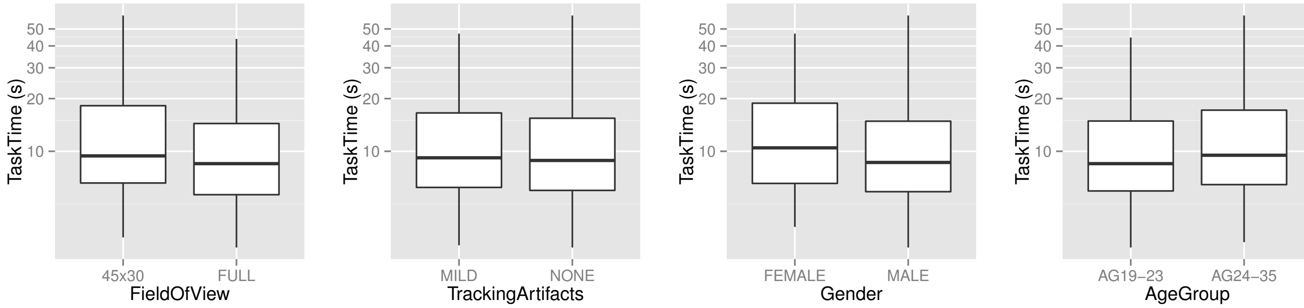


Figure 8: We found a significant effect of FieldOfView on TaskTime. However, with our number of participants, we did not observe a significant effect of TrackingArtifacts. The box plots used in this paper employ the standard 25%, 50%, 75% version with items beyond 1.5*IQR distance suspected outliers.

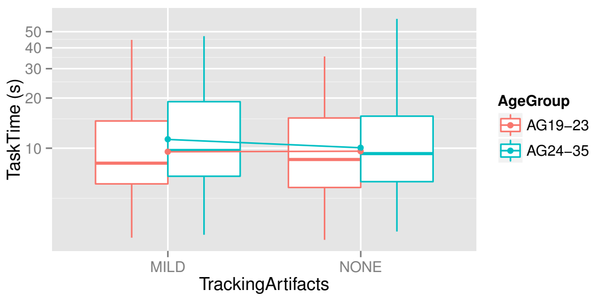


Figure 9: Interaction effect between TrackingArtifacts and AgeGroup. Older people tend to be more affected by tracking artifacts.

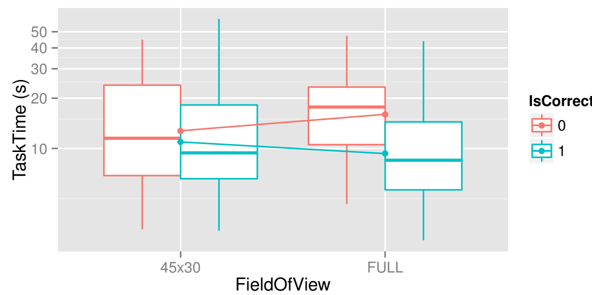


Figure 10: TaskTime grouped by correctness under different FOVs.

employed a Linear Mixed Model for the analysis (Mixed Models are more suitable than repeated measures ANOVA since it better deals with missing values). The analysis was performed in R with the lme4 [6] and lmerTest [22] packages. The results of this three-factor analysis are shown in Figure 7.

FieldOfView and TaskGroup had a significant effect on task completion time. Users performed faster on the FULL FOV condition. In terms of task group, FB1SC required the lesat time, followed by easy chart questions and map questions. Moderate chart questions took considerably more time, and looking for two statues required around twice the time. We did not find a significant interaction effect between FieldOfView and TaskGroup.

We did not find TrackingArtifacts having a significant effect (see Figure 8 for FieldOfView and TrackingArtifacts). With the reminder that our pilot experiment indicated significance on both factors, it is interesting that we only found significance on FieldOfView in the main study.

Our background data from the pre-study questionnaires revealed that the main study had more diverse participants in terms of age, gender, discipline of study, and education level. By comparing the estimated means and confidence intervals showin in Figure 6 and Figure 7, we found that participants in the pilot study generally performed more slowly than the main study, which can be explained by the difference in age and gender distribution. To perform a deeper analysis, we grouped the main study data by Gender (male and female) and AgeGroup (19-23 and 24-35). The charts shown in Figure 8 indicate the possibility that AgeGroup or Gender have an impact on task completion time. We ran a five factor analysis in addition to the standard three factor analysis in order to have a deeper look into the data. It is important to note that we do not consider the significant results found here as robust as the ones from the three factor analysis, since our experiment was not originally designed this way and the five-factor data was not fully balanced. The result of the five-factor analysis is shown in Table 1. We found that there is an interesting borderline interaction effect ($p = 0.565$) between AgeGroup and TrackingArtifacts (see Figure 9), which can be interpreted as older people being more affected by tracking artifacts than young people, which is an intriguing observation to be followed up by future experiments.

5.3 Correctness

The overall correctness was 92.5%. Individually, it was 90.6% and 94.4% for the FULL and 45×30 FOV conditions, 92.3% and 92.6% for NONE and MILD tracking artifact conditions, and 89.0%, 94.0%, 92.6%, 86.4% and 100.0% for F1EC, F1M, F1MC, F2MC and FB1SC task groups conditions respectively. For individual task groups, the worst case was 86.4% on F2MC, which is the task group that asks for two statues, required following two levels of links, and took the most time. Clearly, participants were able to understand the tasks and find the correct information most of the time.

It is interesting to note that with the 45×30 degrees field of view the correctness was 94.4%, which is higher than the full field of view's 90.6%. Fisher's Exact Test between the FULL and 45×30 conditions shows a p-value of 0.0573. Task time distribution in FULL and NONE FOVs by correctness is shown in Figure 10, where we found that the task time difference between correct and incorrect attempts in the FULL FOV condition is much larger than the small FOV condition (we did not find a similar difference pat-

	F value	$Pr(>F)$	
TaskGroup	30.5529	< 0.0001	***
FieldOfView	3.5801	0.0591	.
TrackingArtifacts	4.8348	0.0284	*
TaskGroup:FieldOfView	0.4137	0.7988	
TaskGroup:TrackingArtifacts	0.2862	0.8869	
FieldOfView:TrackingArtifacts	1.2266	0.2686	

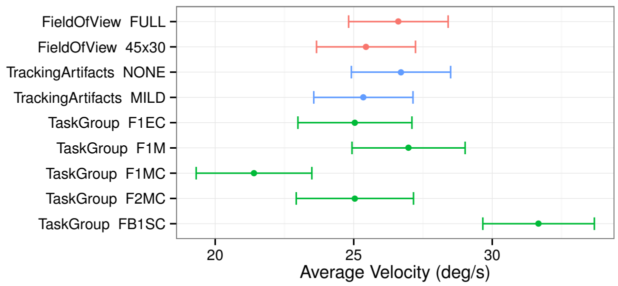


Figure 11: Main Study AverageVelocity: (Top) Analysis of Variance Table of type III with Satterthwaite approximation for degrees of freedom. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. (Bottom) Least Squares Means with 95% Confidence Intervals.

tern for tracking artifacts). This is not what we expected. Some participants reported via informal feedback that they were more careful with the small field of view, and the black frame of the viewport helped them concentrate. We hypothesize that in the full condition, when people get confused or lost, they tend to take more time and get the answer wrong more often than the small condition. An interesting follow-up question here is what would happen if we did not have the black frame, which is left for further investigation.

5.4 Head Movement

We recorded head movement data while participants were performing the tasks. After the experiment, we extracted the trajectories of the center point of the user's field of view, as shown in Figure 12. By comparing the trajectories between full FOV and constrained FOV, we found that participants looked up more under the constrained FOV condition. This suggests that people tend to avoid looking up whenever possible (i.e., with full FOV they can use eye movement to look up and find the information, whereas in constrained FOV they have to turn their heads up). In addition, the trajectories of constrained FOV are more condensed than those of full FOV, which suggests that a small field of view can help people focus better (which was also reported by a few participants after our study). This may also help explain the difference in correctness.

Head movement speed is another interesting measure to analyze. We calculated the average angular velocities based on the trajectories. With a three-factor analysis on AverageVelocity using the Linear Mixed Model (as shown in Figure 11), we found that TrackingArtifacts had a significant effect ($p = 0.0284$). Gender and age also play an important role here. We found that female participants move their heads slower than males (a five-factor analysis yielded $p = 0.0151$), and younger people move their heads faster than older people ($p = 0.0059$). See Figure 13 for boxplots.

We looked through the trajectories from each question, and observed that there are different behaviors. For instance, some people double checked their answers by following the lines multiple times, while some followed the line only once.

5.5 Questionnaire Responses

We did not alert the participants to the fact that during the study they would be experiencing different field of views and tracking artifacts. In the post-study questionnaire, we asked about whether field of view and tracking artifacts made a difference to them. 53% of the participants reported that field of view made a difference on their performance, while 63% of them reported that tracking artifacts made a difference. It is interesting to note that when asked by the experimenter[2] after the study, 94% of the participants (15 of 16 who reported) reported that they observed the change of tracking artifacts; however, only 67% (12 of 18 who reported) observed the change of field of view. It is noteworthy that the relatively constrained 45×30 FOV, which had clear impact on task performance, was apparently not noticed by one third of the participants.

Participants found the lines directly above their heads hard to follow. This aligns with the fact that in the full FOV condition participants appeared to avoid to some extent looking up, whereas, with a constrained FOV, users had to look up more in order to reveal the information.

We observed different strategies in answering the questions. For example, the question "How many books are associated with at least one headless statue", which requires a lot of cognitive effort, can be answered in at least two different ways. Arguably there is a difference in time and accuracy between strategies, and different people chose different strategies.

6 Discussion and Future Work

In this section, we further discuss the results of our experiment and derive directions for further investigation.

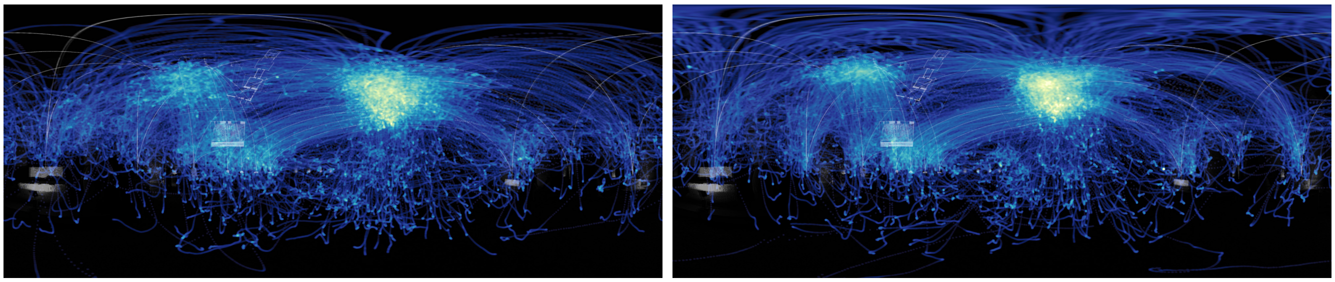
In our experiment, we investigated a set of information seeking tasks in an archaeological tourism application scenario, with wide-field-of-view AR annotations connecting objects far apart in the field of regard.

We found that a difference in field of view, i.e., field of regard constrained by 108×82 degrees field of view shutter glasses versus 45×30 degrees of limited field of view for the augmentations had a significant effect on task completion time. In future experiments, we hope to determine what field of view threshold is sufficient for these and other kinds of tasks. Full field of view would likely be best, but it might be the case that a relatively large field of view (e.g., 90×60 degrees) will work just as well. What does the reward function for FOV increases look like for the range from even smaller FOV frames (e.g., 30×17.5 degrees) to when it might approximate the performance afforded by a full FOV? We would like to remove the 108×82 degrees field of view constraint imposed by the stereo glasses, and experiment with multiple FOVs ranging from full human FOV to even smaller ones (e.g., 30×17.5 degrees) in future studies.

The field of view differences might have affected the way people performed the tasks. With a small FOV, users have to turn their heads more upwards in order to see content with higher elevation, which was reported as cumbersome by some people. On the other hand, several participants reported that a small field of view helped them focus on the annotations, partially because there was a black frame indicating the boundary of the field of view. The black frame might help the brain to establish object permanence [20]. If there was not a black frame, the signal might be that annotations outside the field of view cease to exist. Therefore, the effect of such a frame may warrant further investigations: Should an AR device indicate the boundary of its display or not, and if yes, how?

Unlike in our pilot study, we did not observe that TrackingArtifacts caused a significant effect on task completion time. We explain this by the pilot experiment having a group of participants with low demographic and background variance: most of them were Masters or PhD students around our lab, in the same age group, and only one of them was female. Our main study, on the other hand, had a more diverse population, both in terms of age and gender, as well as discipline of study and education level.

[2] These questions were asked by the experimenters after they observed this phenomenon, only a subset of participants reported this.



Trajectories of FULL field of view. Trajectories of 45×30 field of view.

Figure 12: Comparison between trajectories from full FOV and 45×30 degrees FOV conditions. With constrained FOV, the trajectories on the map are more condensed and a bit more upwards, and there are more trajectories that enter the top area. The altitude distributions show that the amount of time the participants rotate their heads above 60 degrees was longer with the constrained FOV.

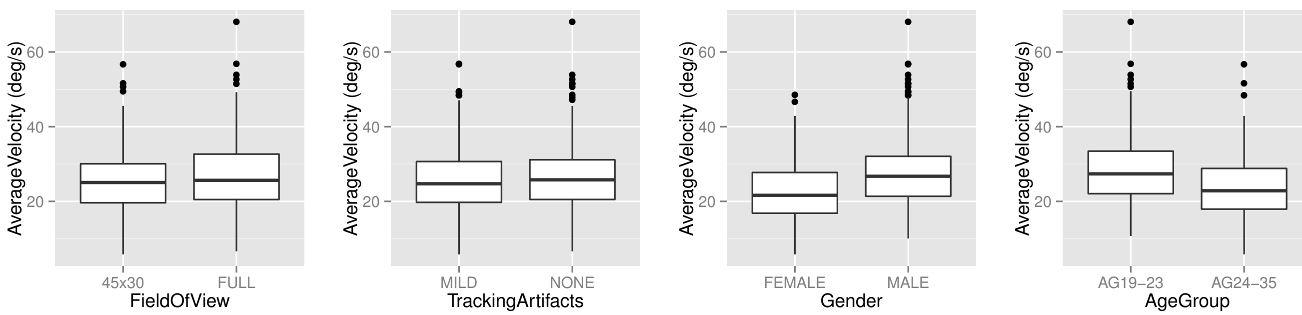


Figure 13: Average angular velocity of head movement.

Although TrackingArtifacts itself did not show a significant effect, when looking deeper into the data, we found an interesting interaction between AgeGroup and TrackingArtifacts, i.e., older people being more negatively affected by tracking artifacts than younger people. This also aligns with our pilot study, where the participants were mostly over 25 years old and performed comparatively worse than the main study population. With the design of our experiment, we are not able to claim such a finding as undeniable fact, but we think it is worth exploring in future experiments. Given the results of the pilot experiment, we can speculate that TrackingArtifacts may have an effect on task performance for increased numbers of participants (beyond the scope of this work). In future research, we would like to conduct further experiments with more participants to potentially verify our assumption.

In terms of correctness, our results indicate that participants performed worse and incorrect attempts took more time in FULL FOVs. We are not able to claim this as a fact as the number of incorrect attempts is still relatively low overall (27 and 16 for FULL and 45×30 respectively). Studying this in depth is left for the future.

Our annotations consisted of a visualization, a map and a photograph linked to statues. We have tested five task groups, with to-be-expected results: easier tasks took less time and exhibited better correctness, tasks asking for two statues required around twice the time for one. It would be interesting to see the results with other types of annotations, such as virtual 3D objects, looking inside of real-world objects using AR X-ray vision, etc. For each type of annotation and task, what field of view is adequate and sufficient? What is the impact of tracking artifacts?

In our experiment, the whole scene including annotations was static. Our system already supports interactive exploration and placement of annotation data. The impact of dynamic and interactive content on AR task performance is another future work item. How well can people cope with limited field of view and tracking artifacts in a dynamic scenario? Research in psychology suggests that peripheral vision is adept at motion detection [27]; therefore, people might perform such tasks better with full FOV AR systems. One might argue that tracking artifacts could become more disconcerting when annotating, moving, and/or changing objects.

7 Conclusion

In this paper, we presented a first detailed experiment on wide-field-of-view augmented reality with annotations connecting objects that are far apart in the field of regard.

We found that people complete tasks faster with full field of view than with a constrained field of view (45×30 degrees).

While in our pilot study we observed a significant effect of (carefully crafted simulated) tracking artifacts, we did not observe it in the main study. However, we surmise an interaction effect between age and tracking artifacts, i.e., older people being more negatively affected by tracking artifacts. We have discussed our results and explanation of findings, as well as topics for future experiments. We hope that this will lead to deeper investigations regarding the effects of field of view and tracking artifacts in AR in general, and wide-field-of-view AR in particular.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions. This work was partially supported by ONR grants N00014-14-1-0133 and N00014-13-1-0872, as well as NSF grant IIS-0747520.

References

[1] R. A. Ainsworth, D. J. Sandin, J. P. Schulze, A. Prudhomme, T. A. DeFanti, and M. Srinivasan. Acquisition of stereo panoramas for display in VR environments, 2011.

[2] D. Arnold and B. E. Shafer. *Temples of ancient Egypt*. Cornell University Press, 1997.

[3] J. Arthur, L. Prinzel, K. Shelton, L. J. Kramer, S. P. Williams, R. E. Bailey, and R. M. Norman. Design and testing of an unlimited field-of-regard synthetic vision head-worn display for commercial aircraft surface operations. In *Proceedings of SPIE, Enhanced and Synthetic Vision*, Apr. 2007. Paper 12.

[4] K. W. Arthur. *Effects of field of view on performance with head-mounted displays*. PhD thesis, 1 2000.

[5] R. Ball and C. North. The effects of peripheral vision and physical navigation on large scale visualization. In *Proceedings of Graphics Interface 2008*, GI '08, pages 9–16, Toronto, Ont., Canada, Canada, 2008. Canadian Information Processing Society.

[6] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, C. Eigen, and L. Rcpp. Package lme4, 2015.

[7] M. Billinghurst, A. Clark, and G. Lee. A survey of augmented reality. *Foundations and Trends in Human-Computer Interaction*, 8(2-3):73–272, 2015.

[8] D. A. Bowman and R. P. McMahan. Virtual reality: How much immersion is enough? *Computer*, 40(7):36–43, 7 2007.

[9] D. A. Bowman, C. Stinson, E. D. Ragan, S. Scerbo, T. Höllerer, C. Lee, R. P. McMahan, and R. Kopper. Evaluating effectiveness in virtual environments with MR simulation. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, pages 12075–1 – 12075–11, Orlando, FL, 2012.

[10] T. J. Buker, D. A. Vincenzi, and J. E. Deaton. The effect of apparent latency on simulator sickness while using a see-through helmet-mounted display: reducing apparent latency with predictive compensation. *Human Factors*, 54(2):235–49, 4 2012.

[11] J. M. Covelli, J. P. Rolland, M. Proctor, J. P. Kincaid, and P. Hancock. Field of view effects on pilot performance in flight. *The International Journal of Aviation Psychology*, 20(2):197–219, 2010.

[12] A. Dünser, R. Grasset, and M. Billinghurst. A survey of evaluation techniques used in augmented reality studies. In *ACM SIGGRAPH ASIA 2008 courses on - SIGGRAPH Asia '08*, pages 1–27. ACM Press, 1 2008.

[13] Innovega Inc. iOptik platform: A new architecture. http://innovega-inc.com/new-architecture.php, accessed September 2015.

[14] J. A. Jones, J. E. Swan, II, G. Singh, and S. R. Ellis. Peripheral visual information and its effect on distance judgments in virtual and augmented environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, APGV '11, pages 29–36, New York, NY, USA, 2011. ACM.

[15] N. Kishishita, K. Kiyokawa, J. Orlosky, T. Mashita, H. Takemura, and E. Kruijff. Analysing the effects of a wide field of view augmented reality display on search performance in divided attention tasks. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, pages 177–186, Sept 2014.

[16] N. Kishishita, J. Orlosky, T. Mashita, K. Kiyokawa, and H. Takemura. Poster: Investigation on the peripheral visual field for information display with real and virtual wide field-of-view see-through hmds. In *3D User Interfaces (3DUI), 2013 IEEE Symposium on*, pages 143–144, March 2013.

[17] K. Kiyokawa. A wide field-of-view head mounted projective display using hyperbolic half-silvered mirrors. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 1–4. IEEE, 11 2007.

[18] J. M. Knapp and J. M. Loomis. Limited field of view of head-mounted displays is not the cause of distance underestimation in virtual environments. *Presence: Teleoper. Virtual Environ.*, 13(5):572–577, Oct. 2004.

[19] N. Kostaras and M. Xenos. Usability evaluation of augmented reality systems. *Intelligent Decision Technologies*, 6(2):139–149, 6 2012.

[20] O. Kreylos. Hololens and field of view in augmented reality. http://doc-ok.org/?p=1274, accessed September 2015.

[21] J. Kuchera-Morin, M. Wright, G. Wakefield, C. Roberts, D. Adderton, B. Sajadi, T. Höllerer, and A. Majumder. Immersive full-surround multi-user system design. *Computers & Graphics*, 40(0):10–21, May 2014.

[22] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen. Package lmertest.

[23] C. Lee, S. Bonebrake, D. A. Bowman, and T. Höllerer. The role of latency in the validity of AR simulation. In *2010 IEEE Virtual Reality Conference (VR)*, pages 11–18. IEEE, 2010.

[24] C. Lee, A. G. Rincon, G. Meyer, T. Höllerer, and D. A. Bowman. The effects of visual realism on search tasks in mixed reality simulation. *IEEE Computer Graphics and Applications*, 19:547–556, 2013.

[25] M. Livingston. Evaluating human factors in augmented reality systems. *IEEE Computer Graphics and Applications*, 25(6):6–9, 11 2005.

[26] A. Maimone, D. Lanman, K. Rathinavel, K. Keller, D. Luebke, and H. Fuchs. Pinlight displays. *ACM Transactions on Graphics*, 33(4):1–11, 7 2014.

[27] S. P. McKee and K. Nakayama. The detection of motion in the peripheral visual field. *Vision Research*, 24(1):25–32, 1984.

[28] T. Olsson, T. Kärkkäinen, E. Lagerstam, and L. Ventä-Olkkonen. User evaluation of mobile augmented reality scenarios. *Journal of Ambient Intelligence and Smart Environments*, 4(1):29–47, 1 2012.

[29] T. Oskiper, M. Sizintsev, V. Branzoi, S. Samarasekera, and R. Kumar. Augmented reality binoculars. *IEEE Transactions on Visualization and Computer Graphics*, 21(5):611–623, 5 2015.

[30] PhaseSpace Impulse X2 motion tracking system. http://www.phasespace.com/impulse-motion-capture.html, accessed September 2015.

[31] B. Porter and R. L. B. Moss. *Topographical bibliography of ancient Egyptian hieroglyphic texts, reliefs, and paintings*, volume 2. Oxford, Clarendon Press, 1972.

[32] E. Ragan, C. Wilkes, D. A. Bowman, and T. Höllerer. Simulation of augmented reality systems in purely virtual environments. In *Virtual Reality Conference, 2009. VR 2009. IEEE*, pages 287–288, 2009.

[33] E. D. Ragan, D. A. Bowman, R. Kopper, C. Stinson, S. Scerbo, and R. P. McMahan. Effects of field of view and visual complexity on virtual reality training effectiveness for a visual scanning task. *IEEE Transactions on Visualization and Computer Graphics*, 21(7):794–807, 7 2015.

[34] D. Ren, T. Höllerer, and X. Yuan. iVisDesigner: Expressive Interactive Design of Information Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(99):2092–2101, 2014.

[35] N. G. Smith, S. Cutchin, R. Kooima, R. A. Ainsworth, D. J. Sandin, J. Schulze, A. Prudhomme, F. Kuester, T. E. Levy, and T. A. DeFanti. Cultural heritage omni-stereo panoramas for immersive cultural analytics – from the Nile to the Hijaz. In *8th Int'l Symposium on Image and Signal Processing and Analysis (ISPA)*, number Ispa, page 6, 2013.

[36] A. State, K. P. Keller, and H. Fuchs. Simulation-based design and rapid prototyping of a parallax-free, orthoscopic video see-through head-mounted display. In *Proceedings of the 4th IEEE/ACM International Symposium on Mixed and Augmented Reality*, ISMAR '05, pages 28–31, Washington, DC, USA, 2005. IEEE Computer Society.

[37] J. Swan and J. Gabbard. Survey of user-based experimentation in augmented reality. In *Proceedings of the 1st HCI International Conference on Virtual Reaity*, 2005.

[38] D. Van Nguyen, T. Mashita, K. Kiyokawa, and H. Takemura. Subjective image quality assessment of a wide-view head mounted projective display with a semi-transparent retro-reflective screen. In *Proceedings of the 21st International Conference on Artificial Reality and Telexistence (ICAT 2011)*, 2011.

[39] J. Ventura, M. Jang, T. Crain, T. Höllerer, and D. Bowman. Evaluating the effects of tracker reliability and field of view on a target following task in augmented reality. In *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*, pages 151–154, 2009.

[40] M. Zöckler, D. Stalling, and H.-C. Hege. Interactive Visualization of 3D-Vector Fields Using Illuminated Steam Lines. *Proceedings of IEEE Visualization 96*, pages 107–114, 1996.