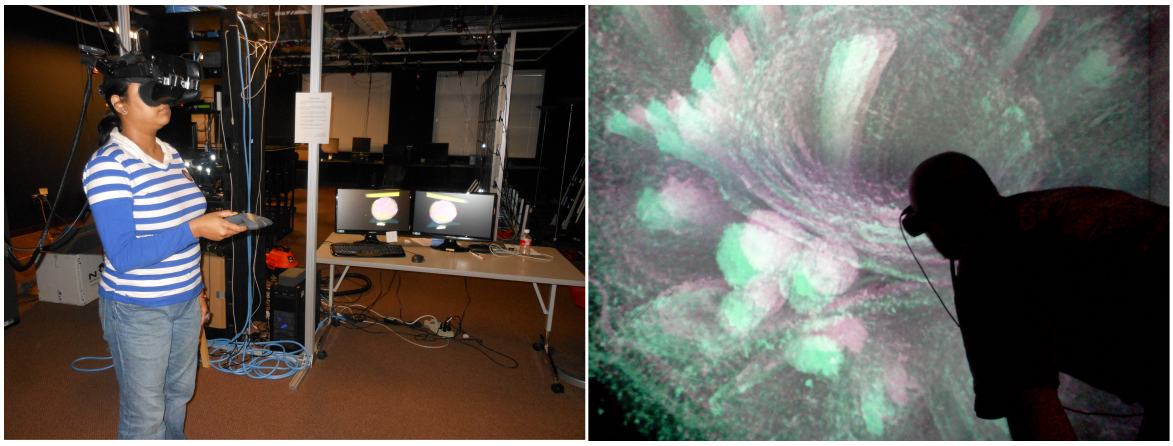
Validation of the MR Simulation Approach for Evaluating the Effects of Immersion on Visual Analysis of Volume Data

Bireswar Laha, Doug A. Bowman, and James D. Schiffbauer



(a) A participant wearing the NVis SX111 HMD (b) A participant inside the Viscube (4-sided CAVE)

Fig. 1. Participants using two different Mixed Reality (MR) simulator platforms in our studies.

Abstract—In our research agenda to study the effects of immersion (level of fidelity) on various tasks in virtual reality (VR) systems, we have found that the most generalizable findings come not from direct comparisons of different technologies, but from controlled simulations of those technologies. We call this the mixed reality (MR) simulation approach. However, the validity of MR simulation, especially when different simulator platforms are used, can be questioned. In this paper, we report the results of an experiment examining the effects of field of regard (FOR) and head tracking on the analysis of volume visualized micro-CT datasets, and compare them with those from a previous study. The original study used a CAVE-like display as the MR simulator platform, while the present study used a high-end head-mounted display (HMD). Out of the 24 combinations of system characteristics and tasks tested on the two platforms, we found that the results produced by the two different MR simulators were similar in 20 cases. However, only one of the significant effects found in the original experiment for quantitative tasks was reproduced in the present study. Our observations provide evidence both for and against the validity of MR simulation, and give insight into the differences caused by different MR simulator platforms. The present experiment also examined new conditions not present in the original study, and produced new significant results, which confirm and extend previous existing knowledge on the effects of FOR and head tracking. We provide design guidelines for choosing display systems that can improve the effectiveness of volume visualization applications.

Index Terms—MR Simulator, immersion, micro-CT, volume visualization, virtual reality, 3D visualization, HMD, virtual environments

1 Introduction

As Slater explains [28], we can consider *immersion* as the objective level of sensory fidelity produced by a VR system. This is different from the concept of presence, which is a user's subjective psychological response to a VR system, or the sense of being there. Immersion is related to display and interaction fidelity [17], and different VR systems vary widely in their levels of immersion. Since VR systems with high levels of immersion can be costly and complex, decision makers need evidence for the benefits of immersion if they are to choose such a system. For researchers, understanding the effects of immersion (realism) is one of the fundamental questions in the field.

- *Bireswar Laha is with the Center for Human-Computer Interaction and the Department of Computer Science, Virginia Tech. E-mail: blaha@vt.edu.*
- *Doug A. Bowman is with the Center for Human-Computer Interaction and the Department of Computer Science, Virginia Tech, Blacksburg, VA. E-mail: bowman@vt.edu.*
- *James D. Schiffbauer is with the Department of Geological Sciences, University of Missouri, Columbia, MO. E-mail: schiffbauerj@missouri.edu.*

Manuscript received 13 September 2012; accepted 10 January 2013; posted online 16 March 2013; mailed on 16 May 2013.

For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

To learn about the effects of individual components of immersion, researchers run controlled empirical studies [4]. It is very difficult and sometimes impractical to maintain experimental control, however, when comparing multiple VR systems, which may vary in many ways (e.g., FOV, stereoscopy, head tracking). We have previously claimed that more control can be achieved using the mixed reality (MR) simulation approach, where a high-immersion VR system (the MR simulator platform) is used to simulate lower-immersion VR systems to produce the conditions for a controlled experiment [5]. Based on the MR continuum [20], MR simulation encompasses both VR as well as AR simulation [15, 16]. The MR simulation approach allows us to simulate any VR or AR system based on selected levels of the various components of immersion. Thus, it promises to provide generalizable results no matter what hardware platform is used as the MR simulator.

However, some may question the real-world applicability and validity of the results of MR simulation studies. In particular, there is little evidence that experiments run on different MR simulator platforms will produce equivalent results. As we build up evidence in favor of or against the validity of the MR simulation approach, we need to connect and argue about the findings from both VR and AR simulation studies together [15, 16], under the term MR simulation.

In one of our prior MR simulation studies, which used a CAVE-like system as the simulator platform [14], we evaluated the effects of field of regard (FOR), stereo, and head tracking on visual analysis tasks with scientific volume datasets. To validate the results of the prior study and to examine the validity of the MR simulation approach, in this paper we present a new study that replicates some of the conditions from the prior experiment but uses a very different VR system (a head-mounted display, or HMD) as the MR simulator platform. We also added new conditions in the current study to understand more about the effects of FOR on volume data analysis.

Our results provide evidence both for and against the validity of the MR simulation approach. Absolute task performance was similar in the two studies, but many of the statistically significant results from the original experiment were not replicated in the current study. We also report significant findings from our current study that confirm and extend our previous knowledge on how different levels of FOR interact with head tracking for task performance with volume data. Based on the findings from the two experiments, we present improved guidelines for designing immersive systems that maximize the effectiveness of volume visualization applications.

2 Related Work

VR researchers have been running empirical studies to evaluate the effects of immersive environments for analyzing scientific datasets, and various other tasks. One of the first such studies was run by Zhang et al. [31], reporting significant benefits of the CAVE over a desktop display for interpretation of volume visualized diffusion tensor magnetic resonance imaging (DT-MRI) datasets of brain tumor surgery. Gruchalla [9] reported significant benefits of using a CAVE over a desktop monitor for an oil well path-editing task. Schuchardt et al. [25] showed significant benefits of higher levels of immersion for accuracy and task performance of spatially complex and detailed search tasks in a 3D visualization of underground cave structures. Prabhat et al. [23] compared desktop, fishtank VR, and the CAVE in an empirical study, and found significant benefits of more immersive environments for analyzing volume visualized confocal datasets. Whereas all of the studies above found benefits of higher levels of immersion, Demiralp et al. [6], found significant benefits of fishtank VR, with a lower level of immersion, over a CAVE for an abstract visual search task.

These results, although intriguing, lacked generality because they directly compared actual VR systems. Multiple components of immersion varied simultaneously between conditions. In this way, these studies failed to establish which components of immersion (or combination of components) caused the significant results.

Other researchers have found effects of individual components of immersion, mostly through controlled experimentation using the MR simulation approach [4], in which a system with high levels of immersion, like a six-sided CAVE, can simulate systems with lower levels of different components of immersion. Through such experiments, we know the significant effects of specific components of immersion for search and comparison tasks [21, 22], single-user object manipulation [18], path tracing tasks [2], understanding complex geometric models [30], and graph visualization [29]. Our previous study also reported several effects of three components of immersion for analyzing volume visualized micro-CT datasets [14].

As we gather evidence for the effects of different components of immersion for analyzing scientific and volume datasets, we also need evidence supporting the validity of the MR simulation approach [5]. Prior studies have provided some such evidence by replicating an experiment from the literature and demonstrating that the effects of the simulator's latency were independent of other effects [15], and by comparing results from an experiment with a simulated AR system to those from an actual AR system [16]. In this paper, we extend this prior work by presenting an experiment designed to obtain evidence that results from two different MR simulator platforms are similar.

3 Experiment

We designed a controlled experiment to reproduce most of the conditions from our previous experiment and also to find more granular results on the interaction effects of FOR and head tracking for analyzing volume datasets.

3.1 Goals and Hypotheses

Our primary goal in this study is to understand if our prior results on task performance with visual analysis of volume data [14] still hold when the experimental conditions are recreated with a different MR simulator platform. Thus, our first research question is:

1. *Are there differences in the findings for various experimental conditions when different MR simulator platforms are used to run the experiment?*

Our earlier study [14] used a four-sided CAVE as the simulator platform; we decided to use a high-end HMD with important differences from the CAVE platform in the current study.

In our previous study [14], we had two levels of FOR (*high* or 270 degrees and *low* or 90 degrees). In several cases, we found significant interactions between FOR and head tracking (HT), with FOR high/HT on and FOR low/HT off proving to be better than the other two combinations. We also found several significant individual effects of FOR. With only two levels of FOR, however, the effects of the highest possible level (360 degrees) and of moderate levels (e.g., 180 degrees) were unknown. This leads to our next research question:

2. *What are the individual effects of FOR and its interaction with HT on visual analysis tasks with volume datasets?*

In this study, we chose to have four levels of FOR (90, 180, 270, and 360), and two levels of HT (on and off).

In response to these research questions, we hypothesized the following:

1. *There are no differences between the findings of an MR simulation experiment run on a CAVE and those from an experiment run on an HMD.*

In theory, since the level of immersion is an objective description of a VR system [28], the effects produced by any MR simulator platform, which is characterised by particular levels of immersion components, should be comparable. Thus, if we simulate the experimental conditions as closely as we can, using the different MR simulator platforms, then we should get similar results. However, other differences between the platforms, such as FOV, weight, accommodation distance, and the presence or absence of seams on the display, could potentially affect the results. The primary differences between the platforms are shown in Table 1. We hypothesize that the effects will come primarily from the variables being studied, and not from these differences in the platforms.

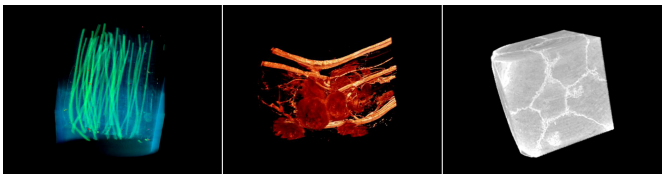
2. *The combination of the highest level of FOR with HT on will produce the best results, followed by the combination of the lowest level of FOR with HT off.*

We hypothesize that the trends from our previous experiment [14] will continue when new levels of FOR are considered.

3.2 Datasets

Computed Tomography (CT) performed at the microscopic (10^{-6}) level, or micro-CT, produces 3D internal imaging of objects, and is useful in various disciplines such as biology, palaeontology, and medicine. Traditionally, researchers have used desktop displays to visualize and analyze micro-CT data in volumetric format. As good visualization is essential for the analysis of such datasets, scientists have shown great interest in evaluating VR platforms for analyzing their datasets [14].

We worked with domain scientists to identify three datasets actively used in their work. The first one is a 3D Scaffold dataset (Fig. 2-a) used in bone regeneration studies [27]. The scaffold mimics the structure of a cortical bone and contains bundles of poly-L-lactide fibers on polyglycolide cores. The individual bundles mimic the osteon, a structural unit of the bone.



(a) 3D Scaffold dataset (b) Mouse Limb dataset (c) Fossil dataset

Fig. 2. The micro-CT datasets used in our studies.

The second dataset was a mouse limb [26], imaged at the major knee joint of the mouse (Fig. 2-b). The visualization also showed the major blood vessels, the soft tissues, and the surrounding musculature in that part of the mouse.

The third dataset was a fossil (Fig. 2-c), dated to 600 million years ago, known as *Parapandorina raphospissa*. This fossil has been interpreted as a potential early animal embryo from the Doushantuo phosphorites of South China [24]. The visualization that participants viewed was of an incomplete fractured specimen.

3.3 Apparatus

3.3.1 Hardware and software

We used the NVis SX111 head mounted display (HMD) as our MR simulator platform (Fig. 3). It offers a FOV of 102° by 64°, with 1280x1024 pixels per eye. Head movements were tracked by a wired head tracker of an Intersense IS-900 tracking system, which also provided a wireless wand device with a joystick and five buttons. A participant using the system is shown in Fig. 1-a. A participant using the MR simulator from our previous experiment (a four-sided CAVE-like system) can be seen in Fig. 1-b. Table 1 shows differences between the CAVE and HMD we used.

We used DIVERSE [11] to get data from the head tracker and the wand from the IS900 system. The open source 3D Visualizer [3] gave us a platform for interactive volume rendering, with stereo capabilities for the two screens of the HMD. We used a customized version of VRUI [12] for the specific 3D interaction needs of our experiment, as described in the following section.

Table 1: Primary Differences between the two MR Simulator Platforms

Factor	CAVE	HMD
Horizontal FOV	90° (with stereo glasses)	102°
Resolution	1920×1920 per wall	1280x1024 per eye
Weight worn on head	85 grams	1.3 kilograms
Stereoscopic display technology	Infitec stereo (passive; rear projected)	Separate displays for each eye
Accommodation distance	Approx. 1.5 m	Infinity
Seams between displays	Visible seams	None
Occlusion of body and surrounding environment	No occlusion	Full occlusion

3.3.2 User interface

To translate the viewpoint, the user could press the joystick forward to travel in the direction the wand was pointing, or press the joystick backward to travel in the opposite direction. Pressing the joystick to the left or right would cause the dataset to rotate about an axis perpendicular to the plane of the wand.

The user could also grab the dataset by holding down a button on the wand, after which the user's hand could be used to directly manipulate the position and orientation of the dataset. Another button press activated a cutting plane feature, which allowed the user to use hand movements to slice the dataset along any arbitrary 3D plane, revealing inner features of the volume data. These interactions were identical to those used in the prior experiment [14].

To correctly simulate the head tracking conditions from our previous study (where we used a four-sided CAVE-like environment [14]), we disabled positional head tracking for conditions where head-tracking was off. Positional head tracking affects the rendering of the volume based on the position of the head tracker. The rotational head tracking was enabled, even in the non-head tracking condition, because in a CAVE-like setting without head tracking, head rotations still allow the user to see views of the dataset in different directions. In head-tracked conditions, both positional and rotational head-tracking were enabled.

Fig. 3. NVis SX111 Head Mounted Display.

3.4 Tasks

We used the same set of tasks from our previous study [14], but with a few key modifications (in appendix). Tasks were either quantitative (counting features) or qualitative (describing characteristics). In the quantitative tasks, participants gave their answer verbally, and the experimenter recorded the response on paper.

For qualitative tasks (deviating from the previous experiment in which the participants answered verbally), the participants were shown a choice of five response options, from which they marked the most appropriate one. We chose to do this to have more objectivity to the analyses of the results, as previously we found that open-ended responses to descriptive tasks often produced a wide array of responses, some of which were difficult to interpret and grade objectively.

In our previous project, we had worked with micro-CT researchers to identify tasks that are of actual importance to their research, to make sure that any benefit of immersion that we found could be used by them and others in their community. Since we planned to run the studies with novice participants (to avoid confounds based on prior knowledge level) we ensured that the tasks were of real technical importance to experts but at the same time not so cryptic so as to confuse novices. We assumed a basic knowledge of what blood vessels, cells, bones, and other simple biological structures look like.

To train the participants, we had three quantitative and three qualitative tasks with a training dataset (Fig. 2-a). The tasks for the two main datasets with the suggested strategies and new multiple response options for the qualitative tasks are shown in the appendix.

The tasks in each dataset were different. But as before, we categorized them in abstract task categories (see Table 3) and counterbalanced the order of the datasets so that each combination of dataset and experimental condition was studied.

3.5 Design

This controlled experiment was primarily designed as a follow up to our previous experiment [14]. In this experiment we wanted to closely study the effects of two independent variables, FOR, and HT, keeping all other factors constant. FOR had four levels: 360, 270, 180, and 90 degrees. At level 'x', the user could view x° of the virtual environment by rotating her head about the vertical axis. HT had two levels: on and off. At the 'on' level, both rotational and positional HT was enabled. At the 'off' level, only rotational HT was enabled. The different conditions with the levels of FOR and HT are shown in Table 2.

For producing the four levels of FOR, we created two virtual black walls. The black walls extended infinitely in the vertical direction. Horizontally, they merged four inches behind the head position, and formed a horizontal angle corresponding to the FOR. The walls moved (changing position, but not orientation) with the user's head. While this is not exactly like the different levels of FOR in a CAVE-like display, it ensured that the user could not move his head through the walls. Conditions with 360-degree FOR had no black walls.

Table 2. Conditions Experienced by the Eight Groups in the Experiment

Group#	First Condition (Mouse-Limb)		Second Condition (Fossil)	
	FOR	HT	FOR	HT
1	360	On	360	Off
2	360	Off	360	On
3	270	On	270	Off
4	270	Off	270	On
5	180	On	180	Off
6	180	Off	180	On
7	90	On	90	Off
8	90	Off	90	On

With four levels of one variable and two levels of the other, we had eight possible conditions. We chose to vary FOR between subjects and HT within subjects, as in the previous experiment. This allowed us to study whether individuals used different strategies to explore the datasets with and without HT. Although all participants experienced both levels of HT, we consider those who experienced HT on first to be a separate group from those who experienced HT off first, since the two datasets were not comparable in terms of complexity. All participants first performed tasks with the mouse limb dataset followed by the fossil dataset (Table 2).

As before, the dependent variables were the amount of time taken for each task and the responses of the participants to each task, recorded and graded offline by the experimenter using the grading rubric. We also recorded participants' responses for the difficulty levels of each task, and their subjective levels of confidence in their answers for each task, both on seven-point scales.

3.6 Participants

We recruited 65 voluntary unpaid participants for our study, all of whom reported no prior experience in analyzing volume visualized micro-CT datasets. Most of the participants were recruited through a university wide recruitment system, and got awarded two credits in a psychology course for their participation. Four of them were pilot participants. We dismissed 13 participants based on below-threshold scores on a spatial ability test [7]. This gave us a total of 48 participants, distributed uniformly in eight study groups (six participants per group), with comparable spatial ability scores in each group. The overall average spatial ability of the participants in this study (8.35, max score 20) was lower than that (10.95, max score 20) in the previous study [14]. In this study, 26 males and 22 females participated, all undergraduate or graduate students. They were 18 years to 41 years old, with an average age of 21 years.

3.7 Procedure

Our study was approved by the Institutional Review Board of our university. Before beginning the study the participants signed a standard informed consent, informing them of their right to withdraw at any point during the study. Next, participants filled out a background questionnaire capturing information on their demographics, experience with VR, and experience analyzing CT and micro-CT datasets. Following that, they took the spatial ability test [7] discussed above. The participants were then given a brief background of the purpose of the study, introduced to the hardware, and trained with the various 3D interactions they were about to use.

The tasks in the training dataset (Fig. 2-a) trained the participants in the different expert strategies that domain scientists use. The training introduced the participants to the various interactions with the HMD and the wand, how to analyze a volume visualized micro-CT dataset using that system, and how to complete quantitative and qualitative tasks. The participants trained on the same condition in which they would experience the first (mouse limb) dataset. The participants also completed three rotation tasks, about the three orthogonal axes, with the joystick of the wand, to make sure they could comfortably use the rotations when needed, without outside assistance. The training took around 15-20 minutes.

After the training, participants were asked to take a short break. Then they started analyzing the mouse limb dataset (Fig. 2-b). The participants were asked to be as accurate as possible in their responses. They were informed that there was a maximum amount of time for each task. For the quantitative tasks, they were asked to let the experimenter know when they were ready to answer. The experimenter recorded the time using a stopwatch. For the qualitative tasks, they were asked to analyze the dataset for the entire available time, after which they were shown five options. After every task completion, the experimenter also recorded the perceived level of difficulty and confidence level in two separate seven-point scales. The details of the tasks are in the appendix.

The participants then rested for a short while, and again underwent training in the condition they would use to analyze the fossil dataset (Fig. 2-c) in. They performed seven tasks with the fossil, in the same manner as the mouse limb, and the experimenter recorded their responses in the response sheets.

As in the previous study, if the participants digressed too much from the expected strategy for a particular task, we guided them towards the correct expert strategies. The appendix lists the main strategies for each task identified by the domain experts. We thus tried to emulate expert strategies as closely as possible.

Finally, the participants completed a post-questionnaire capturing their opinions for both the head-tracked and non-head-tracked conditions on seven-point scales for: comfort level, ease of getting the desired view and exploring the dataset in general, and ease of understanding the features of a dataset and doing different tasks with the dataset. For both levels of HT, participants also rated the effectiveness of three visual analysis strategies: changing the viewpoint by rotating or grabbing the dataset with the wand, slicing the dataset with the wand, and physically walking around the dataset to look from different viewpoints.

The datasets in each condition were rendered at the same initial position and orientation in front of the participants. Each question was read out loud to the participants, using consistent wording.

4 Results

In this section, we first present the significant results from our recent controlled experiment. We then present the comparison of the results of our current study with those from our previous study with the CAVE-like system. We first compared the significant results from the two studies. We then compared the grade metric in all the conditions from both experiments in which all the components of immersion were at the same level.

In the current study, task time was analyzed as a numeric continuous variable, while the other measures (grade, difficulty, and confidence) were considered to be numeric ordinal variables. To understand the significant main effects and the two factor interaction effects of our independent variables (FOR and HT), we ran a two-way analysis of variance for the time metric, and an Ordinal Logistic Regression based on a Chi-square statistic for all other metrics.

For the sake of brevity, we shall use the task numbers as defined in the appendix; e.g., 'M1' will denote the first task with the mouse limb dataset, and 'F4' will denote the fourth task with the fossil dataset. We used our previous classification of the tasks in the abstract categories as shown in Table 3. Table 3 also shows the relative weights of the tasks (totalling 1.0 for each dataset) determined by domain scientists, based on the perceived relative importance of the tasks to their own research. We used these weights

to calculate the weighted totals ∑M and ∑F for the mouse limb and fossil datasets respectively. ∑M and ∑F helped us to evaluate the overall effects on the tasks with a particular dataset.

Table 3. Relative Weights of Tasks and Abstract Task Categories

Mouse task#	Task Type	Weights	Fossil Task#	Task Type	Weights
M1	Simple search	0.25	F1	General description	0.15
M2	General description	0.15	F2	Internal feature search	0.25
M3	Visually complex search	0.3	F3	General description	0.05
M4	Spatially complex search	0.3	F4	Visually complex search	0.25
			F5	Visually complex search	0.1
			F6	General description	0.15
			F7	Simple search	0.05

4.1 Significant results from current experiment

Here we report all the significant main effects and interaction effects of the independent variables FOR and HT in our present experiment with the HMD system. For the interaction effects, we also present graphs to compare them with those from the previous experiment.

4.1.1 Grades (accuracy in task performance)

The significant main effects of FOR and HT on the grades received by the participants are shown in Table 4. We found significant interactions of FOR and HT on the grades received by the participants in two cases. The first case is the effect on M4 grade ($\chi^2_{df=3}$=10.371, p=0.016) and is shown in Fig. 4. The left graph is from the original experiment with a CAVE-like system; the right graph is from the present experiment. Overall, the M4 grades show similar trends in both experiments. Grades were better with higher FOR when HT was on, but were better with lower FOR when HT was off. The mean and variances of the grades are also comparable. Additionally, in the HMD experiment, we learned that the grades reached the lowest level at FOR 270 with HT off, and didn't change significantly from FOR 270 to 360 with HT on.

Table 4. Significant Main Effects on Grades

Task: source	χ^2	DF	p-value	Note (higher grade is better)
F1: FOR	7.983	3	0.046	270>360>180>90
F3: FOR	11.849	3	0.008	180>360=90>270
F4: HT	8.342	1	0.004	on > off
∑F: HT	9.967	1	0.001	on > off

The second significant interaction of FOR and HT on F4 grades ($\chi^2_{df=3}$=8.672, p=0.034) is shown in Fig. 5. Again, the left graph is from the previous experiment with a CAVE-like system; the right graph is from the current experiment. The F4 grades in the two graphs are comparable at FOR 270 and 90 for both HT on and off. In the HMD experiment, the best results were achieved with FOR 360 and HT on, and the worst results with FOR 180 and HT off. We found that three of the six participants in group five (with FOR 180 and HT off condition in the fossil dataset) failed the task completely. As a result the data point probably became an outlier in our study.

4.1.2 Task completion time

We found a significant main effect of FOR on F4 time (F(3, 40) = 5.4773, p=0.003, power=0.9149) in the HMD experiment. A post-hoc t-test (t=2.021) indicated that the task performance at FOR 360, and FOR 270 was significantly faster than that at FOR 90, and FOR 180, with the fastest mean time achieved with FOR 360.

4.1.3 Subjective metrics

There were no significant main or interaction effects of FOR or HT for the perceived difficulty metric in the current experiment.

For perceived confidence levels of F6, we found a significant main effect of FOR ($\chi^2_{df=3}$=8.394; p=0.0385), and of HT ($\chi^2_{df=1}$=4.58; p=0.0323). Confidence levels were highest with FOR 180, decreased with FOR 90 and FOR 270, and were lowest with FOR 360. Confidence levels were higher with HT on.

There were significant interaction effects between FOR and HT for three tasks: F4, F5, and F6. Table 5 shows the χ^2 and p-values of the interaction effects and the means of the different conditions.

From the table, we see that the participants consistently had higher confidence levels for three conditions: FOR 90 with HT off, FOR 360 with HT on, and FOR 180 with HT on, and consistently had the lowest confidence for the condition FOR 360 with HT off.

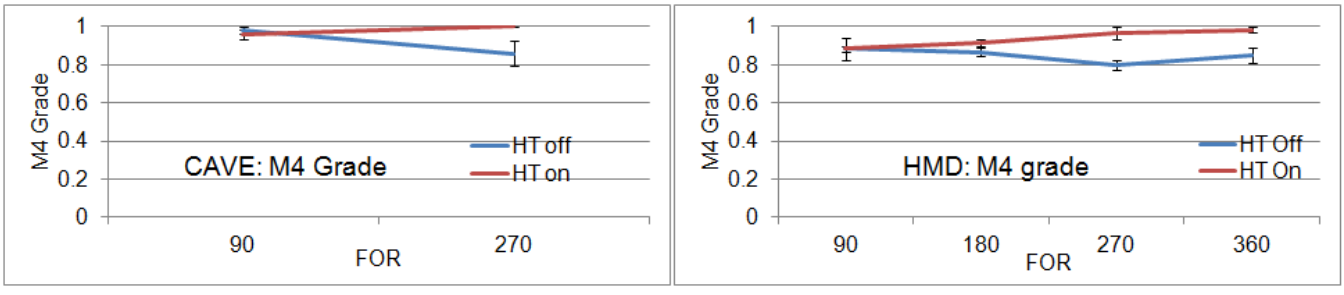


Fig. 4. Interaction between FOR and HT for M4 grade.

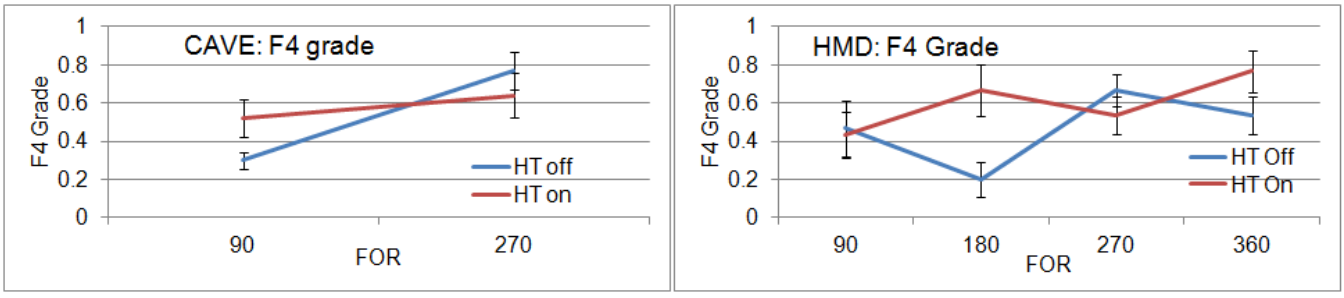


Fig. 5. Interaction between FOR and HT for F4 grade.

Table 5. Significant Interaction Effects of FOR and HT on Confidence

Task: source	χ^2	DF	p-value	Mean values in descending order (higher is better)	
F4: FOR & HT	10.552	3	0.014	FOR 90 HT off	5.3
				FOR 360 HT on	5.2
				FOR 180 HT on	4.7
				FOR 180 HT off	4.3
				FOR 270 HT off	4.2
				FOR 90 HT on	4.2
				FOR 270 HT on	3.7
				FOR 360 HT off	3.2
F5: FOR & HT	8.858	3	0.031	FOR 90 HT off	6
				FOR 180 HT on	5
				FOR 360 HT on	4.8
				FOR 180 HT off	4.5
				FOR 90 HT on	4.5
				FOR 270 HT on	4.3
				FOR 270 HT off	4
				FOR 360 HT off	3.5
F6: FOR & HT	9.689	3	0.021	FOR 180 HT on	6.2
				FOR 90 HT off	5.8
				FOR 360 HT on	5.3
				FOR 180 HT off	5.2
				FOR 270 HT off	5.2
				FOR 90 HT on	5
				FOR 270 HT on	4.3
				FOR 360 HT off	3.5

4.1.4 Post-questionnaire results

In our post-questionnaire, we captured subjective ratings of user's perception, as in our previous experiment, on a seven-point scale. The users felt that head tracking significantly improved ($\chi^2_{df=1}$=11.784; p=0.0006) the ease of getting the view they wanted, and exploring the dataset in general. In the previous experiment with CAVE as the MR simulator platform [14], we also found the same result ($\chi^2_{df=1}$ = 3.854, p = 0.0496).

4.1.5 Effects of spatial ability and gender of the participants

We ran pairwise correlation analyses (nonparametric Spearman's ϱ) between the spatial abilities, and gender of participants and the different metrics in our study. We found a few significant correlations with spatial ability: M2 grade (ϱ=0.42; p=0.003), M3 time (ϱ=0.3146; p=0.0294), F1 grade (ϱ=0.3693; p=0.0098), and F4 time (ϱ=0.2882; p=0.047). Accuracy was higher but performance was slower with higher spatial ability for these tasks.

Significant correlations of gender were with F2 time (ϱ=-0.3131; p=0.03), and M4 difficulty (ϱ=0.4573; p=0.001), Females were faster in F2, but found M4 more difficult than males.

4.2 Comparison of MR simulator platforms

We designed our current experiment so that the immersion conditions had significant overlap with those from a previous study [14], which was run on a CAVE-like system. The 90- and 270-degree FOR conditions in our current study matched with the low and high FOR conditions, respectively, in the previous study. Both studies had the HT on and off conditions. We also had conditions with stereo on in the previous study, and all the conditions in our current study had stereo on. Thus, we had two similar levels each of two independent variables in both experiment, giving us four independent conditions to compare.

In the current study, for the qualitative tasks, we introduced a multiple choice response system, which was different from the previous study. Further, we asked the participants to analyze the dataset for a fixed amount of time for every qualitative task. Thus, we could not directly compare the results of the two studies for the qualitative tasks. The quantitative tasks remained the same from the previous study. There were three quantitative tasks in each of the two datasets in our study, giving us six independent tasks for comparison.

With four comparable immersion levels, and six independent tasks to compare, we had 24 sets of comparable results from both experiments that we could use to provide evidence for the validity of the MR simulation approach.

4.2.1 Comparison of significant effects

Table 6 compares the significant effects found in the experiments using the two different MR simulation platforms, for the quantitative tasks only, organized by metric (columns) and variable (rows). The cell with significant results found in both experiments is shaded. Note that the results in the table are slightly different from the results reported in our prior paper [14], because we ran a new two-factor analysis of the earlier data using only the conditions with stereo on, and because FOR is treated here as an ordinal, rather than nominal variable. In addition, these effects are different than those reported above, because this analysis for the current experiment used only the data from the FOR 90 and FOR 270 conditions. Only one of the seven significant effects from the CAVE experiment was reproduced in the HMD experiment. Three new significant effects were seen in the HMD experiment that did not occur in the CAVE experiment.

Out of 65 non-significant effects in the CAVE experiment, we found 62 in the HMD experiment. Overall, then, out of the 72 significant and non-significant results (6 tasks x 3 effect types x 4 metrics), we found 63 in the current study. The similarity between the results suggests that the first hypothesis has some validity. We discuss possible causes for the differences in significant effects in section 5.1.

Table 6. Significant Effects from the two Experiments

	Grade		Time		Difficulty		Confidence	
	CAVE	HMD	CAVE	HMD	CAVE	HMD	CAVE	HMD
FOR	F4			F4				F2
HT	F4	M4			M4		F4	
FOR* HT	M4	M4					F4	
	F4							

4.2.2 Contingency Analysis results

Out of the four metrics in our study, grade was the most important, as it represented accuracy in task performance. We ran contingency analyses to find the relation between the two sets of grades in each comparable condition from the two studies, as the grades were considered ordinal data [1].

Table 7. P-Values for Likelihood Ratio and Pearson Chi-Square

Tasks	FOR 270		FOR 90	
	HT ON	HT OFF	HT ON	HT OFF
M1	0.2956 0.402	0.411 0.4974	0.1308 0.2307	0.2028 0.2909
M3	0.1446 0.2592	0.0044* 0.0266*	0.0995 0.2399	0.2264 0.387
M4	0.2551 0.3384	0.0356* 0.0981	0.4278 0.5169	0.0076* 0.0219*
F2	0.2542 0.3519	0.2264 0.387	0.3883 0.52	0.448 0.569
F4	0.5543 0.6879	0.2328 0.3820	0.3268 0.5102	0.1847 0.3447
F7	0.6179 0.621	0.1308 0.2307	0.2475 0.3611	0.0233* 0.0601

In Table 7 above, the upper value in each cell is the likelihood ratio, and the lower value is the p-value for the Pearson Chi-Square statistic. P-values less than 0.05 (marked with an asterisk) lead us to reject the null hypothesis *H_0: The ordinal values in the two series come from the same distribution*. In other words, insignificant p-values indicate that we did find a difference in grades between the two experiments. Out of the 24 different cases, no significant difference was found in 20 cases, while four cases, all from tasks with HT off, had a significant difference.

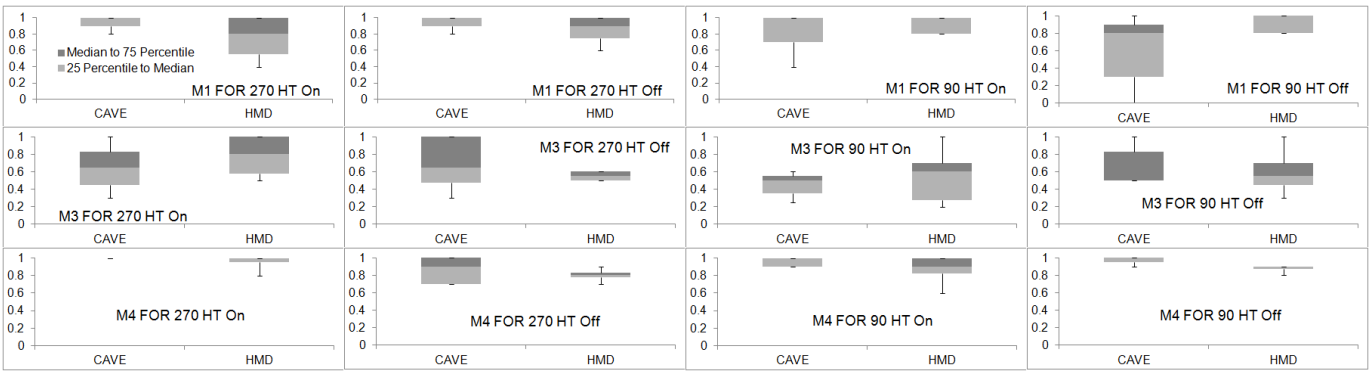


Fig. 6. Box plot comparison of grades of mouse limb dataset from the CAVE experiment and the recent HMD experiment.

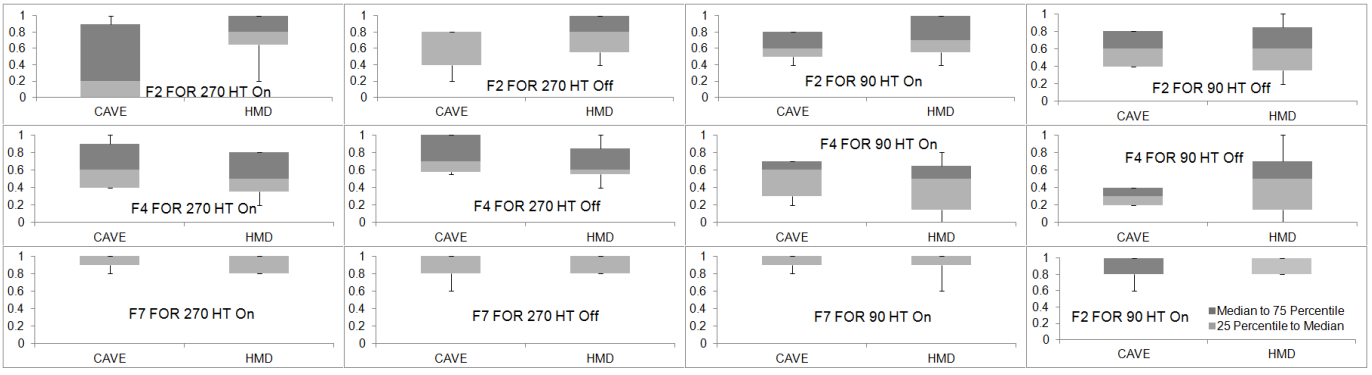


Fig. 7. Box plot comparison of grades of fossil dataset from the CAVE experiment and the recent HMD experiment.

4.2.3 Mann-Whitney test results

We also ran a Mann-Whitney test to compare the medians of the grades from the two experiments. This is a non-parametric test parallel to comparing the means of two independent populations [1]. Since the grade metric was categorical data, comparing medians is more appropriate rather than comparing means.

The p-values in the Table 8 are those for the Chi-square test statistic, the null hypothesis H_0 being that *the group means or medians are in the same location across groups*. In other words, insignificant p-values (>0.05) indicate that we did find a difference between the two MR simulators. Of the 24 different cases, no significant difference was found in 23 cases. The only significant difference was found for task M4 with FOR 90 and HT off. We also report the power of each test in Table 8. Lower power indicates higher error (type II) in claiming that the two means are the same.

The box plots in Fig. 6 compare the grades of the 12 cases from the mouse limb dataset, and those in Fig. 7 compare the grades of the 12 cases from the fossil dataset. A discussion on how the graphs compared is in section 5.

Table 8. P-Values for Chi-Square Statistic in Mann-Whitney Test

		FOR 270		FOR 90	
Tasks		HT ON	HT OFF	HT ON	HT OFF
M1	Prob>ChiSq	0.1073	0.2844	0.8157	0.077
	Power	0.3412	0.1748	0.0691	0.3661
M3	Prob>ChiSq	0.6363	0.0926	0.3993	0.7728
	Power	0.1403	0.2427	0.0906	0.0609
M4	Prob>ChiSq	0.3613	0.5069	0.3129	0.0105*
	Power	0.1281	0.1253	0.1655	0.9140
F2	Prob>ChiSq	0.1844	0.3865	0.5069	1.0
	Power	0.2944	0.1188	0.0999	0.05
F4	Prob>ChiSq	0.508	0.509	0.5778	0.3513
	Power	0.0967	0.112	0.0778	0.1502
F7	Prob>ChiSq	0.6374	0.8157	1.0	0.3061
	Power	0.0691	0.0522	0.0598	0.0862

5 Discussion

5.1 MR Simulation Validity

Looking back at our research questions for this study (section 3.1), we have found some evidence for, and some evidence against, our first hypothesis that MR simulation produces valid results even with very different simulator platforms (see Table 1).

It should be noted that our studies have limited scope, examining only two components of immersion, a few task types, and a particular application domain (volume visualization). It is not possible to definitively prove the validity of MR simulation with these experiments.

In favor of validity are the results comparing the grades for the quantitative tasks in the two experiments. Of the 24 different combinations of system characteristics and tasks that we compared from the two experiments, we found that the measured grades were not significantly different in 20 cases (83.3%) with a contingency analysis and 23 cases (95.8%) with a Mann-Whitney test. This is evidence that, overall, our two MR simulator experiments produced consistent results for task performance in individual conditions, even though the two MR simulator platforms were quite different.

The box plots shown in Fig. 6 and Fig. 7 give a quick and comprehensive visual comparison of the grades obtained by the participants in the two studies. In a few cases, we notice a big variance in the grades from one of the MR simulator platforms, such as for task F2 with FOR 270 and HT on. It is hard to make any judgments in these cases. But in many cases we can judge that the plots are very similar. These plots are visual evidence of the similarity between the grades obtained from the two MR simulator platforms, to complement the statistical analyses.

Half of the tasks from the two datasets were qualitative in nature, which changed from the previous study to become multiple choice questions. As verbal reports would require more cognition to form the answer, it might have been easier and/or quicker for the participants to answer these in the HMD experiment. We thus traded

off the chance to compare the results of these tasks between the MR simulator platforms, for more objectivity in the results of this study.

On the other hand, we found that most (six out of seven) of the significant effects found in the original experiment were not reproduced in the current study (section 4.2.1). This argues against the validity of MR simulation experiments run on simulator platforms that are highly different than the actual target platforms. It is important to understand what might have caused these differences. Table 1 listed the differences in the MR simulator platforms we used. We also noted that our experiments had a relatively small number of participants in each condition, differed in the male to female participant ratio, and used participants with different mean spatial ability scores.

We have no direct evidence to tell us which of the differences in the simulator platforms or participants might have caused the differences in results between the two studies, but we can speculate. We believe that participant spatial ability may have played an important role, especially in the lower-immersion conditions, but spatial ability alone is probably not enough to explain the differences. Based on observation and experience with the two systems, we believe that occlusion, weight, and stereo display technology may have been the most important technological differences between the platforms.

The total occlusion of the real world (including the participant's body) in the HMD platform means that users are more tentative about turning and walking, and are more likely to feel disoriented. Prior MR simulation studies measuring presence, and anxiety using a virtual 'pit' have found that the ability to see the user's body [8, 10] or a lack of it [10, 19] could affect presence, and thereby movement of users in the environment.

The weight and general encumbrances of the HMD platform also contribute to these phenomena. Many of the users felt that the HMD was heavy, and were concerned about its tethering. In particular, in the non-head tracking conditions, the users in the HMD experiment might not have moved as much as the participants in the CAVE experiment, who could see their whole body, as well as lacked the tethering, and the weight of the HMD on their shoulders. We wonder whether having a lighter and wireless HMD simulating the various conditions in our MR simulation would have produced results more similar to those from the CAVE experiment.

The Infitec stereo technology used in the CAVE study is a passive stereo approach that (anecdotally) results in color perception problems in some users. Stereo effects may be more pronounced or easier to view in the HMD platform, perhaps reducing the need to rely on head tracking.

We intentionally chose two MR simulator platforms that were very different, in order to test how far we could go without compromising the validity of the results. It seems that the differences in our CAVE and HMD simulator platforms were too pronounced, so future studies should use MR simulators that have fewer differences from the target platform.

5.2 Effects of FOR

For our second research question regarding the effects of FOR, we have significant results from this study supporting our second hypothesis. We found that HT was more effective with higher FOR (360 and 270). The combination of high FOR with HT on was significantly better than the 90 FOR with HT off for M4 grade, which was a spatially complex search task. Additionally, we observed that grades improved considerably from FOR 180 to FOR 270 with HT on for M4. These results suggest that for complex search tasks, the combination of high FOR with head tracking provides the best conditions for accurate visual analysis, probably because it allows users to physically walk around the dataset, rather than rotating it, to view it from different directions.

To take advantage of an increased FOR, a user needs to look from different distances in different directions. In the current study, we found in most tasks that the participants moved much less than in the study in the CAVE. Thus, the only significant effects of FOR that we found were on tasks F1 and F3, both of which were of general description type (see Table 3).

We also observed that participants' confidence levels were consistently higher with conditions similar to desktop displays (FOR 90 HT off) or to the real world (FOR 360 HT on) than other conditions (section 4.1.3). This is consistent with the results of other studies that have found benefits of the most familiar conditions [17].

5.3 Implications for design

From our two studies, we can extend our previous guidelines for designing VR systems for improving effectiveness for visual analysis of volume data [14]:

- VR systems with FOR 270 degrees or above with head tracking are useful for spatially complex search tasks with volume datasets.
- VR systems with fewer encumbrances might produce more significant benefits of higher immersion for visual task analysis with volume datasets.

For researchers studying the effects of immersion on visual task analysis with volume data, we recommend the use of the *MR simulation approach* for creating more generalizable results, but the MR simulation platform used in these studies should be similar to the target platform for the application.

6 Conclusions and Future Work

Our goal in this research was to examine the validity of MR simulation by comparing the results from two similar experiments using quite different MR simulator platforms (CAVE and HMD). We have presented a variety of evidence both for and against the validity of the MR simulation approach for empirical studies with volume datasets. We believe that the differences we found in the results of the two studies are primarily due to the highly different simulator platforms we chose to compare, and that studies with more similar platforms will result in more-repeatable results.

A secondary goal of this work was to learn more about the effects of FOR on visual analysis tasks with volume datasets. We observed that beneficial effects of FOR depend on physical movement. VR hardware such as HMDs that have greater weight, less visibility of the real world, and other encumbrances can cause reduced movement and might lessen the positive effects of FOR on visual analysis tasks. Even with the HMD platform, however, we still found at least one task category where high levels of FOR combined with head tracking resulted in improved task performance.

Our work can be extended in several ways. Although we have studied the validity of MR simulation based on two simulator platforms, additional evidence from studies of different platforms, tasks, metrics, and application domains is needed.

A generic task taxonomy for visual tasks performed by scientists on volume datasets would be very helpful in designing controlled experiments and generalizing their results across application domains. We have made some progress in this direction [13], but more work is needed to ensure a comprehensive and useful list of generic visual analysis tasks.

Appendix

Tasks with the mouse limb dataset:

M1. *You are allowed a maximum time of 1 minute for this task. How many soft tissues like these could you count in the dataset?* (**Simple search**) – A soft fluffy structure was shown near the surface, which was a soft tissue. The task was to visually search for similar structures in the dataset.

Strategy: Rotate the dataset completely about any axis.

M2. For your next task, you are allowed a maximum time of 1 minute. This is a bone of the mouse limb. Please study the inner core of the bone for a minute. After that I will ask you to choose an answer to describe it. Your time starts now.

Response options - A. Network of fibers/strands. B.Fluffy like cotton. C. Spongy. D. Single block; no texture. E. Collection of many small pieces.

(**General description**) – A particular bone was shown. The task was to describe the structure of the bone marrow.

Strategy: Look at the bone marrow from different angles. Might also use the slice tool to look at various cross-sections.

M3. For your next task, you are allowed a maximum time of 2 minutes. Here is an example of a distinct bone segment. Count the other distinct bone segments in the sample. If a bone branches into two or more parts, only count it once. Also describe the overall structure formed by the bone segments – what letter of the alphabet does the structure resemble? (**Visually complex search**) – In the previous task, the participant worked with a bone. This task was to visually search for bone structures present in the entire dataset, and connect them together to form a letter of the alphabet.

Strategy: Rotate and also slice and look from various angles.

M4. For your last task, you are allowed a maximum time of 3 minutes and 30 seconds. Now let's say, this is the top and this is the bottom of the structure. Please find and count the number of distinct blood vessels which are visible from the top and bottom, but cannot be seen from any side-view. (**Spatially complex search**) – The task was to look separately from the top and bottom of the dataset and search for blood vessels present at each end. There were two blood vessels visible from the top and inside the dataset, and one visible at the bottom.

Strategy: Rotate to the top and bottom. Slice and look from different angles, and also look inside the dataset.

Tasks with the fossil dataset:

F1. For your first task, you are allowed a maximum time of 1 minute. Describe in your own words the 3D structure that you see in front of you.

Response options:

a. Half of a nearly spherical object, with distinct internal chambers. Outside resembles a soccer ball.
b. A completely spherical object, with a cloud-like internal structure. Outside resembles a basketball.
c. A spherical object cut in half, with a grainy texture inside. Outside resembles a soccer ball.
d. A completely spherical object with no distinct internal structure.
e. Half of an oblong object, like an American football, with distinct internal chambers.

(**General description**) – The task was to describe the structure of the entire dataset.

Strategy: Look from different angles and describe.

F2. For your next task, you are allowed a maximum time of 3 minutes. As you can see, the Parapandorina specimen consists of multiple bounded compartments, each of which is potentially a cell. You can see these compartments forming in 3D as you use the cutting plane to slice through the dataset in a single direction. Please count the number of compartments that you can identify. (**Internal feature search**) – A cell structure was shown in 3D inside the fossil volume. The task was to count all the cells that the participant can identify in the entire volume.

Strategy: Slice through the dataset, spatially constructing the cells in various layers of the volume; look from different angles.

F3. For your next task, you are allowed a maximum time of 1 minute 30 seconds. Describe the shape of the borders and the joints between the cells. Are the borders and joints between different cells comparable or different in shape?

Response options:

	Thickness	Texture/Close Inspection	Brightness	Junctions
A	Comparable	Variable; predominant sheet-like	Darker than surrounding material	X-shaped
B	Variable	Similar fabric; clotted, similar to bunches of grapes	Brighter than surrounding	X-shaped
C	Variable	Similar fabric; clotted, similar to bunches of grapes	Brighter than surrounding	Y-shaped
d	Comparable	Variable; clotted, similar to bunches of grapes	Darker than surrounding material	Y-shaped
e	Variable	Similar fabric; predominant sheet-like	Brighter than surrounding	X-shaped

(**General description**) – A border is the line separating two cells, and a joint is where more than two cells join. This task is about describing the shape of the borders and joints in the volume, comparing the borders to each other and joints to each other.

Strategy: Slice, rotate and look from different angles.

F4. For your next task, you are allowed a maximum time of 3 minutes. Within each cell there may or may not be intracellular bodies. In how many of the cells can you identify intracellular bodies? How many per cell? (**Visually complex search**) – An intracellular body was shown in 3D. The task was to scan through the entire volume and search for intracellular bodies in all the cells that the participant had identified.

Strategy: Slice through the entire volume, look from different angles, identify structures in cells and reconstruct them in 3D.

F5. For your next task, you are allowed a maximum time of 1 minute and 30 seconds. Please identify the position of the intracellular bodies within the cells (e.g. – near the boundary or centrally located).

Response options:

a. Always floating within the cell not attached to boundaries.
b. Attached to the outer surface of the fossil structure.
c. Contained within cell boundaries only, no part of the intracellular structure juts into the interior of the cell.
d. When close to the cell boundaries, they always merge with the boundary. Otherwise, they are completely detached and floating within the cell.
e. Always touching cell boundaries.

(**Visually complex search**) – The location of intracellular bodies inside a cell in 3D could be close to a boundary or more towards the center. The task was to search for gaps between the intracellular bodies and the boundaries in all directions in every cell and report.

Strategy: Slice; look from different angles through the volume.

F6. For your next task, you are allowed a maximum time of 1 minute and 30 seconds. Describe the shape of the intracellular bodies. Are they of comparable or differing shape?

Response options - A. Cubic, solid, and variable in size. B. Pill-shaped (spherocylindrical), solid, and variable in size. C. Spherical or moon-shaped, typically hollow rather than solid, and comparable in size. D. Spherical or moon-shaped, typically solid rather than hollow, and comparable in size. E. Pill-shaped, typically hollow rather than solid, and comparable in size.

(**General description**) – The task was about describing the structures of the intracellular bodies in the cells.

Strategy: Slice; look from different angles through the volume.

F7. For your last task, you are allowed a maximum time of 1 minute and 30 seconds. This crack is called a fracture. How many cells does this fracture cut through? (**Simple search**) – The fracture near the bottom right corner (looking from the cut surface of the fossil) was shown. The task was to search for cells through which the fracture cut.

Strategy: Look from different angles; slice if necessary.

ACKNOWLEDGMENTS

The authors wish to thank Kriti Sensharma (Virginia Tech) for sharing his insights from medical biology domain. Thanks to Patrick Shinpaugh (Virginia Tech), and Oliver Kreylos (University of California, Davis) for their valuable technical support in this project. Thanks to the anonymous reviewers for their insightful comments.

REFERENCES

[1] A. Agresti, *Analysis of Ordinal Categorical Data*: John Wiley & Sons, 2010.

[2] W. Barfield, C. Hendrix, and K. Bystrom, "Visualizing the structure of virtual objects using head tracked stereoscopic displays," in *IEEE Virtual Reality Annual International Symposium*, 1997, pp. 114-120.

[3] M. I. Billen, O. Kreylos, B. Hamann, M.A. Jadamec, L.H. Kellogg, O. Staadt, and D.Y. Sumner, "A geoscience perspective on immersive 3D gridded data visualization," *Computers & Geosciences,* vol. 34, pp. 1056-1072, 2008.

[4] D. A. Bowman, and R. P. McMahan, "Virtual Reality: How Much Immersion Is Enough?," *Computer,* vol. 40, pp. 36-43, 2007.

[5] 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*, 2012.

[6] C. Demiralp, C.D. Jackson, D.B. Karelitz, S. Zhang, and D.H. Laidlaw, "CAVE and Fishtank Virtual-Reality Displays: A Qualitative and Quantitative Comparison," *IEEE Transactions on Visualization and Computer Graphics,* vol. 12, pp. 323-330, 2006.

[7] R. B. Ekstrom, J.W. French, and H.H, Harman, "Cognitive factors: Their identification and replication," *Multivariate Behavioral Research Monographs,* 1979.

[8] M. Gandy, R. Catrambone, B. MacIntyre, C. Alvarez, E. Eiriksdottir, M. Hilimire, B. Davidson, A. C. McLaughlin, "Experiences with an AR evaluation test bed: Presence, performance, and physiological measurement," in *2010 9th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, ed, 2010, pp. 127 -136.

[9] K. Gruchalla, "Immersive well-path editing: investigating the added value of immersion," in *Proceedings of the IEEE Virtual Reality*, 2004, pp. 157–164.

[10] M. C. Juan and D. Pérez, "Comparison of the Levels of Presence and Anxiety in an Acrophobic Environment Viewed via HMD or CAVE," *Presence: Teleoperators and Virtual Environments,* vol. 18, pp. 232-248, 2009.

[11] J. Kelso, S.G. Satterfield, L.E. Arsenault, P.M. Ketchan, and R.D. Kriz, "DIVERSE: A Framework for Building Extensible and Reconfigurable Device-Independent Virtual Environments and Distributed Asynchronous Simulations," *Presence: Teleoperators and Virtual Environments,* vol. 12, pp. 19-36, 2003.

[12] O. Kreylos, "Environment-Independent VR Development," in *Advances in Visual Computing*. vol. 5358, G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, F. Porikli, J. Peters, J. Klosowski, L. Arns, Y. K. Chun, T. Rhyne, and L. Monroe, Ed., ed: Springer Berlin / Heidelberg, 2008, pp. 901-912.

[13] B. Laha, D. A. Bowman, "Identifying the Benefits of Immersion in Virtual Reality for Volume Data Visualization," presented at the Immersive Visualization Revisited Workshop of the IEEE VR conference, 2012.

[14] B. Laha, K. Sensharma, J. D. Schiffbauer, D. Bowman, "Effects of Immersion on Visual Analysis of Volume Data," *IEEE Transactions on Visualization and Computer Graphics,* vol. 18, pp. 597-606, 2012.

[15] C. Lee, S. Bonebrake, T. Höllerer, D. A. Bowman, "The Role of Latency in the Validity of AR Simulation," in *Proceedings of IEEE Virtual Reality*, 2010, pp. 11-18.

[16] C. Lee, S. Gauglitz, T. Hollerer, D. A. Bowman, "Examining the equivalence of simulated and real AR on a visual following and identification task," in *IEEE Virtual Reality Conference*, ed. Los Alamos, CA, USA: IEEE Computer Society, 2012, pp. 77-78.

[17] R. P. McMahan, D. Bowman, D. Zielinksi, R. Brady, "Evaluating Display Fidelity and Interaction Fidelity in a Virtual Reality Game," *IEEE Transactions on Visualization and Computer Graphics,* vol. 18, pp. 626-633, 2012.

[18] R. P. McMahan, D. Gorton, J. Gresock, W. McConnell, and D.A. Bowman, "Separating the effects of level of immersion and 3D interaction techniques," in *Proceedings of the ACM symposium on Virtual reality software and technology*, 2006, pp. 108–111.

[19] M. Meehan, B. Insko, M. Whitton, F. Brooks, "Physiological measures of presence in stressful virtual environments," *ACM Trans. Graph.,* vol. 21, pp. 645–652, 2002.

[20] P. Milgram, and F. Kishino, "A Taxonomy of Mixed Reality Visual Displays," *IECE Transactions on Information and Systems,* vol. 18, pp. 1321-1329, 1994.

[21] T. Ni, D. A. Bowman, J. Chen, "Increased display size and resolution improve task performance in Information-Rich Virtual Environments," in *Proceedings of Graphics Interface*, 2006, pp. 139-146.

[22] N. F. Polys, S. Kim, and D.A. Bowman, "Effects of information layout, screen size, and field of view on user performance in information-rich virtual environments," *Computer Animation and Virtual Worlds,* vol. 18, pp. 19-38, 2007.

[23] Prabhat, A. Forsberg, M. Katzourin, K. Wharton, and M. Slater, "A Comparative Study of Desktop, Fishtank, and Cave Systems for the Exploration of Volume Rendered Confocal Data Sets," *IEEE Transactions on Visualization and Computer Graphics,* vol. 14, pp. 551-563, 2008.

[24] J. D. Schiffbauer, S. Xiao, K. Sensharma, and G. Wang, "The origin of intracellular structures in Ediacaran metazoan embryos," *Geology,* 2012.

[25] P. Schuchardt and D. A. Bowman, "The benefits of immersion for spatial understanding of complex underground cave systems," in *Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, 2007, pp. 121–124.

[26] K. Sensharma, D.M. Vasilescu, A.S.K. Puliyakote, E.A. Hoffman, T. Andric, W.J. Freeman, C. Markert, J.D. Schiffbauer, S. Xiao, H. Yu, and G. Wang, "Novel Biomedical and Biological Applications using Lab-based Multi-scale CT System," presented at the Annual Meeting of the Biomedical Engineering Society, 2011.

[27] K. Sensharma, T. Andric, W.J. Freeman, C.L. Wyatt, and G. Wang, "Micro-CT for osteon-like scaffolds," presented at the 11th Annual Conference of the North Carolina Tissue Engineering and Regenerative Medicine Society, 2009.

[28] M. Slater, "A note on presence terminology," *Presence,* vol. 3, 2003.

[29] C. Ware, and G. Franck, "Evaluating stereo and motion cues for visualizing information nets in three dimensions," *ACM Transactions on Graphics,* vol. 15, pp. 121–140, 1996.

[30] S. Zhang, C. Demiralp, and D. H. Laidlaw, "Visualizing Diffusion Tensor MR Images Using Streamtubes and Streamsurfaces," *IEEE Transactions on Visualization and Computer Graphics,* vol. 9, pp. 454-462, 2003.

[31] S. Zhang, C. Demiralp, D. F. Keefe, M. DaSilva, D. H. Laidlaw, B. D. Greenberg, P. J. Basser, C. Pierpaoli, E. A. Chiocca, and T. S. Deisboeck, "An Immersive Virtual Environment for DT-MRI Volume Visualization Applications: A Case Study," in *Proceedings of IEEE Visualization*, 2001, pp. 437-584.