Analysis of Subject Behavior in a Virtual Reality User Study

Jürgen P. Schulze[1], Andrew S. Forsberg[1], Mel Slater[2]
[1] Department of Computer Science, Brown University, USA
[2] Department of Computer Science, University College London, UK
[1] {schulze,asf}@cs.brown.edu, [2] m.slater@cs.ucl.ac.uk

Abstract

This paper presents the findings of new analysis of data collected from a prior user study comparing a CAVE-like environment and a Fish Tank VR setup. In particular, our earlier study focused on subject performance across multiple conditions and this study focuses on demographics and subject behavior during trials. We found some unexpected relationships between subject background and performance, as well as interesting details of the marking process with regard to timing and posture. We present novel ways to analyze the large amounts of data gathered in virtual reality user studies.

1 Introduction

User studies are an important part of virtual reality research because they provide insight into how well a subject pool can use systems and techniques to solve real problems. We designed and ran a user study in order to examine differences in performance of a task associated with different levels of immersive system. In this study the task was derived from a biological one, where scientists visualize a complex structure normally rendered as 3D volume data, and have to count the occurrence of specific features. In order to avoid the problem of the 3D volume data rendering at different speeds on different machines, we abstracted out of this a generic task in which subjects were asked to place the tip of 3D cone-shaped markers inside spheres that represented cells from a biological application. Subjects did this task in five variations of a four-walled CAVE-like environment and a 'Fish Tank' Virtual Reality setup (a CRT screen with active stereo and six degree of freedom head- and hand-tracking). Our initial analysis was aimed at discovering which environment subjects marked spheres fastest and most accurately in, as well as which environment subjects preferred.

In an earlier work [1], data we collected led to four significant findings: a) in the Cave the subjects preferred medium sized or large spheres over small spheres, b) when only a few of the targets had been pre-marked, larger spheres were marked faster than small spheres, c) large spheres are marked most accurately, and d) the single Cave-wall display is not comparable to the fish tank virtual reality environment when the spheres are small. Additionally, occlusion and larger field of view inhibited performance in the Cave more than at the fish tank when the task was dominated by visual search.

Below we present new findings after further analysis of the data collected previously.

2 The Experiment

We have implemented a software application named VOX (for VOlume eXplorer) that can allow users to mark 3D spheres in virtual reality. Our goal was to find out how well this task could be performed in different types of virtual environments and at different dataset scales. This system was motivated by our collaboration with developmental biologists at Brown University with whom we built VOX for immersively viewing data sets from confocal microscopes.

Counting cells or cellular components is standard practice in studying many biological processes, such as assessing the proliferation of a tissue and determining the size of cells. The density of certain components within a particular volume is often compared in control and experimental samples. Immunohistochemical techniques that use antibodies, tagged (or labeled) with a fluorochrome or other molecule that fluoresces under particular wavelengths of light, allow biologists to highlight (or recognize) the structures of interest within a tissue preparation. The digitized data is collected with a laser scanning confocal microscope that generates a static volume data set.

Counting cells requires uniquely identifying and tallying the cells within a volume; here the main challenges are isolating individual cells and (due to the large number of cells) avoiding double-counting. We selected this task because it did not require specialized knowledge of biology but still involved visualizing and interacting with the biological data.

In immersive VR, the user is fully immersed in a three-dimensional world that is completely computer-generated. The user sees a stereo image of the dataset through stereo glasses, and the user's head is tracked so that the image can be interactively rendered for the user's viewpoint. Direct 3D interaction with objects is also possible. Thus, we expected immersive VR would be a good technology for the cell-counting task because head-tracked stereo viewing would ease both seeing the dense cluster of cells and annotating already counted cells. However, there are many variations of immersive VR systems and we could not predict which combination of system attributes (e.g., field of view, scale of data) would be "best". This study was aimed at investigating these issues.

Our visualization system runs in a Fish Tank VR environment and also in our Cave, which is a CAVE-like virtual reality system. The Cave can be configured to show images only on its front wall, making it a single-wall display.

In our study users used cone-shaped icons to mark spheres while the computer tracked the total number of markers placed. In the Cave, we scaled the data set to be $0.30m^3$, $0.91m^3$, and $2.13m^3$; at the single wall and the fish tank we used only the $0.30m^3$ data set because only that size would completely fit on the smaller screen. Below we refer to the five conditions as follows: CS = "Cave small" $0.30m^3$ condition, CM = "Cave medium" $0.91m^3$ condition, CL = "Cave large" $2.13m^3$ condition, SW = "Single wall" Cave condition, and FT = Fish Tank condition.

Because we were interested in the rate of marking, it was not necessary for subjects to mark the entire data set. However, we expected that marking rates might vary over time—in particular, as more cells were marked the task might become dominated by visual search for the next unmarked cell. Time constraints also limited the number and length of conditions we could use. Consequently, instead of having the subjects mark all 250 spheres in a data set, we divided the marking task in two halves: we started with an unmarked data set (UM) and gave the subjects two minutes to mark as many spheres as they could. Then we interrupted the experiment, removed the markers the subject had placed, and loaded spheres in the same configuration, but with 210 of them pre-marked (PM). The subject had another two minutes to mark the remaining 40 spheres.

3 Analysis

Below we present our additional analysis of the data collected during the user study. In particular, we report our findings on a demographic analysis, an analysis of tracker data, an analysis of the marking process, and further insights into why the $0.30m^3$ data set was marked slower than the larger ones.

3.1 Description of variables

Response Variables (dependent variables). The dependent variable is the count of the number of correctly marked spheres.

Independent Variables. The condition was the only independent variable (CS, CM, CL, SW, FT).

Explanatory Variables. Age, type of degree, and gender were not controlled for. Additionally, in our analysis we measured the number of seconds each user spent, respectively, tumbling and moving the data set in UM and PM.

3.2 Marking accuracy and rate

The data shows several interesting results with respect to marking accuracy and rate. A marker was "accurately placed" if its tip was inside a sphere and it was considered "inaccurately placed" if its tip was outside of the sphere. We expected that if a subject placed markers slowly then they would place them accurately. However, the data shows that there was no correlation between marker accuracy and rate. Before the experiment we had instructed all subjects to focus both on marker accuracy and rate, so perhaps they actually followed this guideline, and the differences are caused by their general ability to use our system, which might be influenced by experience with other computer programs, especially games.

3.3 Relationships between marking rate and independent and explanatory variables

Our analysis also showed a relationship between the response variable (the number of markers placed) and the independent (condition) and explanatory variables (age, degree type, gender, time spent rotating data set). Since our response variable is a count, it would have a Poisson distribution. The null hypothesis is that it is unrelated to any of the independent or explanatory variables. We use log-linear regression analysis from the generalized linear model [2].

Many user study analyses of virtual reality applications have been published in the past [3-10], but few of them went beyond the direct and exclusive analysis of the response variables.

3.3.1 Unmarked condition

Condition is significant. When no spheres had been pre-marked we found the condition is significant (Chi-squared to delete from the model is 89.17 on 4 d.f.). Subjects counted a significantly higher number of spheres in CM and CL than CS, but they were not significantly different from each other. SW is not significantly different from CS. Subjects counted significantly more spheres in FT than CS, but less than CM or CL (these statements are taking into account the other terms in the regression analysis).

We think that the fact that the spheres were bigger made them easier to mark, despite the greater distance between the spheres. It is our hypothesis that at some level of scale larger than CL, the marking rate will start to drop because the user spends significantly more time navigating between spheres. More sample points in the condition domain would be needed to determine the "optimum" level of scale in the Cave.

Degree is significant. Non-Computer Science concentrators marked a significantly lower number of spheres than Computer Science concentrators. (Chi-squared for deletion from model is 29.78 on 1 d.f.).

We think that this result may be influenced by the subjects' amount of experience with similar, game-like tasks, but only a further user study that collects that information can prove this hypothesis.

Time spent rotating model. Subjects that spent more time tumbling the data set with the trackball marked significantly fewer spheres than those who spent less time tumbling the data set with the trackball. (Chi-squared for deletion from model is 7.411 on 1 d.f.).

The explanation for this result could be that those subjects that spent less time tumbling the data set had more time to place markers.

3.3.2 Pre-marked condition

Condition is significant. Subjects did not mark significantly different numbers of spheres in CS, CL, and SW. CM and FT were associated with higher sphere counts, but were not significantly different from one another. Chi-squared for removal is 14.4 on 4 d.f.

Pre-marked trackball tumbling is significant. Subjects that spent more time tumbling the data set with the trackball marked significantly fewer spheres than those who spent less time tumbling the dataset. Chi-squared for removal is 13.29 on 1 d.f. (very highly significant). Notice that this result is aligned with the corresponding result for the unmarked condition.

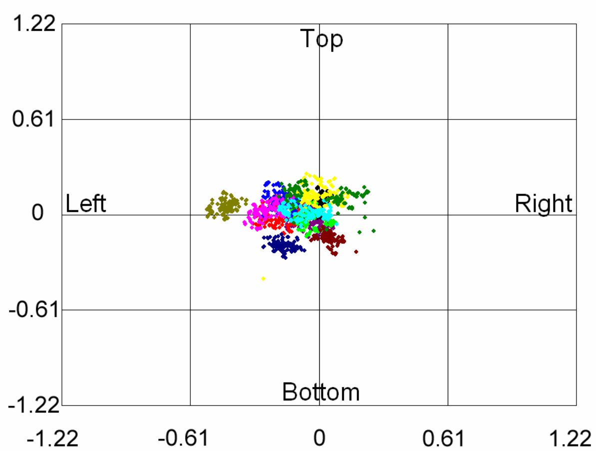

Figure 1: Wand position in CS-UM as seen from the rear of the Cave. The Cave is 2.44m wide and 2.44m high, the origin of the coordinate system is in its center. The coordinate axes indicate position in meters.

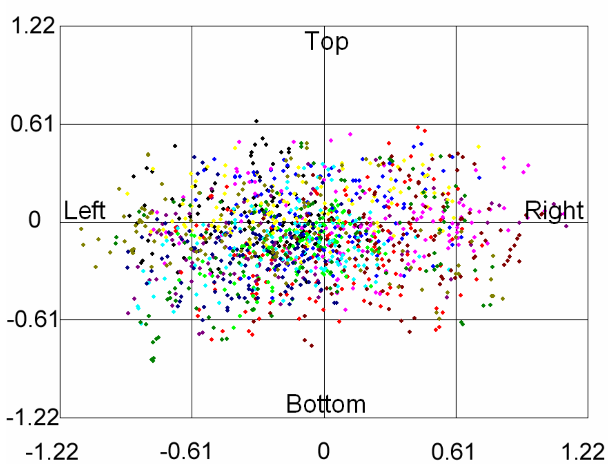

Figure 2: Wand position in CL-UM, as seen from the rear of the Cave.

3.4 Analysis of tracker data

In each experiment we stored position and orientation of head and hand once every second. Then we graphed the results with scatter plots where each dot represents a tracker sample. We use colors to distinguish subjects.

Figure 1 shows the result for condition CS-UM. The dots are projected on a plane parallel to the front wall of our four-walled Cave. Most users moved the wand relatively little.

Figure 2 shows the corresponding graph for CL. It is obvious that the users moved around a lot to reach spheres. Instead of reaching out for spheres they could have navigated the data set and bring the spheres to them, but they preferred to take advantage of the space in the Cave.

Figure 3 shows the head tracker data for CL-UM. The head moved much less than the hand. This graph is the one with the most head movement in our study, and thus it indicates that for our marking task most head tracking happened between about 1.2m and 1.8m from the ground.

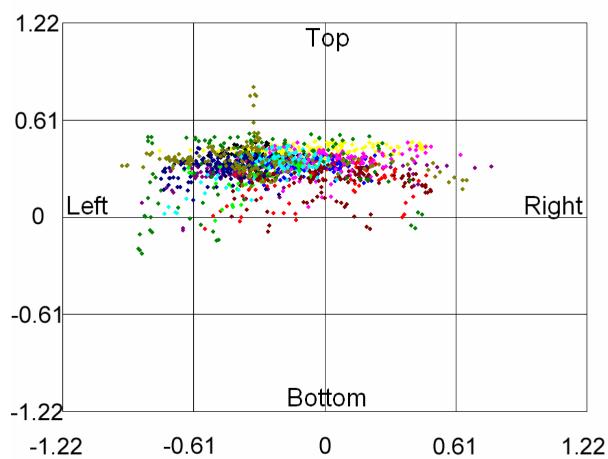

Figure 3: Head position in CL-UM, as seen from the rear of the Cave.

3.5 Marking rate development

In Section 3.3 we analyzed how the marking rate as a whole (i.e., total number of markers placed in each trial) is related to other variables of the experiment. In this section we focus on the marking process itself, looking at the times when individual markers were placed.

In our studies subjects marked spheres with two initial conditions: UM and PM. We chose these two tasks because they simulated the beginning of the marking process: when it is easy to pick out unmarked spheres and the marking rate is determined by the speed the person can move the hand to place a marker. Towards the end of a real marking process, it gets harder to find non-marked spheres and placing a marker is not as critical any more. We expected the curve to be exponential because we assumed that it is more and more difficult to place markers the more are placed.

Figure 4 shows the result for the large scale data set (CL) in the Cave. The graph consists of two parts: the left shows the marking process in the UM condition, the right shows PM. As in the scatter plots, colors distinguish users. The graph shows that all users' marking rates decreased over time. There is a significant difference in the slope of the curves in UM and PM, indicating that we are missing data between UM and PM to know how the complete graph looks, one that we would get if subjects were given enough time to mark all spheres. However, even with the existing data we can see that the curve is not strictly exponential. We hypothesize that this is because subjects typically tumble the data set and then work on a previously unmarked cluster of spheres, in which marking is fast at the beginning and slower towards the end, before the user tumbles the data set again. In the graph, when there is a gap between two marker placement events, it can be assumed that users were navigating by tumbling or moving the data set.

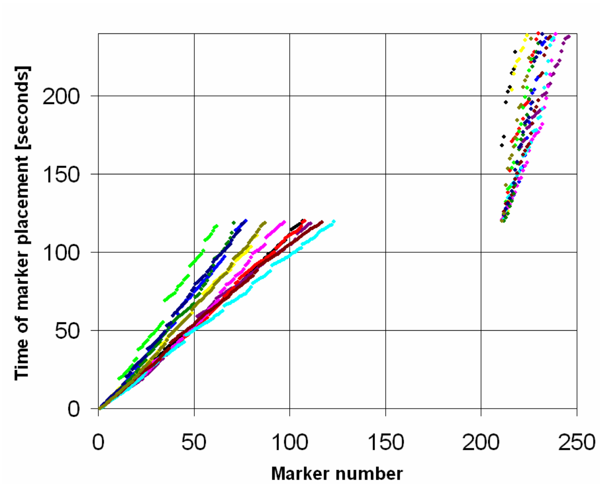

Figure 4: Marker placement times for CL.

To learn more about the rate at which subjects marked, we looked at the time that passed between any two consecutive marker placements. Figure 5 shows the corresponding graph for CL. The middle 90% of these times are between 0.7 and 2.7 seconds for UM and between 0.9 and 13.3 seconds for PM (for this analysis we ignored the slowest and fastest 5% of the times to reduce the number of outliers). The 90% boundaries for the other conditions are listed in Table 1.

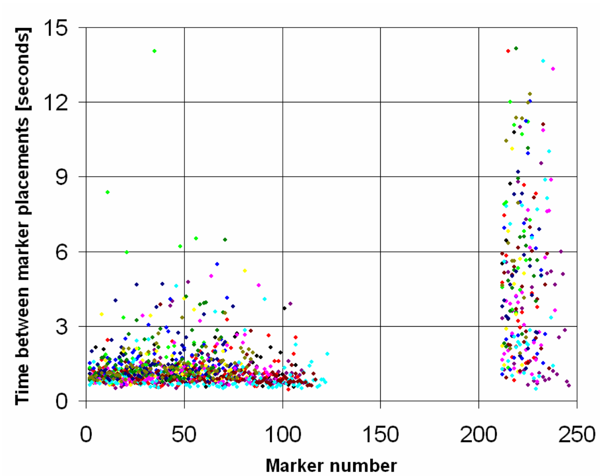

Figure 5: Marker placement rate for CL.

Table 1: 90% of the times between marker placements were between these numbers of seconds.

	CS	CM	CL	SW	FT
UM	0.8..3.1	0.7..2.3	0.7..2.7	0.8..3.0	0.7..3.4
PM	1.2..10.5	1.1..10.9	0.9..13.3	1.3..11.5	1.1..11.4

Condition CL-PM has the greatest range of times between marker placements. This may indicate that large spheres are easy to mark when they are in reach, but when none are in reach navigation takes longer than under other conditions.

3.6 Head and hand posture

A hypothesis we had on user posture was that subjects would hold the wand at the same average distance from the head, independent from the data set's level of scale, because each subject had their own "natural" distance. We were surprised that the data suggests the opposite. Table 2 lists the average distance of the wand from the midpoint of the eyes in the three Cave conditions. The table shows that with larger scale data sets the average distance from eyes to wand increased. The difference between UM and PM within each condition is not significant. We do not list SW because it is comparable with CS, and we do not list FT because it is not comparable since the user sits in front of the computer as compared to standing in the Cave.

Table 2: Average (and standard deviations) distance of wand from midpoint between eyes, given in meter.

	CS	CM	CL
UM	0.38 (0.05)	0.53 (0.10)	0.59 (0.10)
PM	0.39 (0.06)	0.53 (0.10)	0.58 (0.08)

We analyzed the same tracker data as above, wand position relative to head, for the angle at which the wand was held relative to the head direction (see Table 3). We found that there is no significant difference between conditions CS, CM, and CL. However, there is a difference between UM and PM. In PM the average angle is consistently higher than in UM, and it has a higher standard deviation. We think this result can be explained by the subjects having to look around more in PM compared to UM, because they have to search for unmarked spheres.

Table 3: Average (and standard deviations) angle between vector from eyes to wand and head direction, given in degrees.

	CS	CM	CL
UM	29.2 (7.3)	23.2 (9.6)	23.1 (10.6)
PM	32.7 (10.3)	34.4 (15.6)	38.4 (19.1)

4 Discussion

We think that one of the most important results of our user study is that the working volume plays a significant role in a highly interactive task in virtual environments. The more the working volume resembles the natural, comfortable range of human motion, the more efficient a task can be performed. In our user study those levels of scale of the data set which took advantage of the bigger working volume in the Cave, compared to the Fish Tank, resulted in higher marking rates. We think that for even larger spheres than we tested, at some point the marking rate will become smaller again, and it would be interesting to find out when this happens. A trivial “worst case” scenario to support this hypothesis is that if each sphere is about the size of the Cave, then it would take a considerably larger amount of interactions to navigate to unmarked spheres and increase occlusion, which would outweigh the benefit of visually easy-to-target spheres. Our result in Section 3.4 supports this hypothesis, because it shows that the subjects took advantage of most of the available space in the Cave to mark the spheres in condition CL. Bigger spheres than in CL would mean that more navigation will be required, which, according to our interpretation of the results in Section 3.3, would result in a smaller marking rate.

5 Conclusions

We presented novel ways to analyze the large and detailed amounts of data that can be gathered in user studies involving virtual environments. Our analysis is not meant to be complete, but we think it is likely that other interesting and meaningful relationships can be found in our data.

In the future we want to refine our data recording mechanisms to store information about the user performance in virtual environments on top of low-level tracker events, for instance the relationship between features in the data set and the viewing direction. We would also like to record user behavior in greater detail than just tracking head and hand. Furthermore, we would like to test more levels of scale to refine the number of samples we have for this condition.

6 Acknowledgements

This work was partially supported by DOE award DE-FG02-03ER25582, LLNL Research Subcontract No. B527302, and NSF (CCS-0086065). The user study reported here has been approved by the Institutional Review Board (IRB) of Brown University.

References

[1] Schulze, J.P., Forsberg, A.S., Kleppe, A., Zeleznik, R.C., Laidlaw, D.H. (2005). Characterizing the Effect of Level of Immersion on a 3D Marking Task. *Proceedings of HCI International*, Las Vegas.

[2] McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. London; New York: Chapman and Hall.

[3] Arns, L., Cook, D., & Cruz-Neira, C. (1999). The Benefits of Statistical Visualization in an Immersive Environment. *Proceedings of the IEEE Virtual Reality Conference,* Houston, TX, pp. 88-95.

[4] Bowman, D.A., Datey, A., Ryu, Y.S., Farooq, U., & Vasnaik, O. (2002). Empirical Comparison of Human Behavior and Performance with Different Display Devices for Virtual Environments. *Proceedings of Human Factors and Ergonomics Society Annual Meeting*, Baltimore, pp. 2134–2138.

[5] Demiralp, C., Laidlaw, D.H., Jackson, C., Keefe, D., & Zhang, S. (2003). Subjective Usefulness of CAVE and Fish Tank VR Display Systems for a Scientific Visualization Application. *IEEE Visualization '03 Poster Abstract,* Seattle, WA.

[6] Gruchalla, K. (2004). Immersive Well-Path Editing: Investigating the Added Value of Immersion. *Proceedings of the IEEE Virtual Reality Conference*, Chicago, IL, pp. 157-164.

[7] Pausch, R., Proffitt, D., & Williams, G. (1997). Quantifying Immersion in Virtual Reality. *ACM SIGGRAPH 97 Proceedings*, pp. 13–18.

[8] Robertson, G., Czerwinski, M., & van Dantzich, M. (1997). Immersion in Desktop Virtual Reality. *Proceedings of the Symposium on User Interface Software and Technology (UIST)*, Banff, Alberta, Canada, pp. 11-19.

[9] Slater, M., Usoh, M., Linakis, V., & Kooper, R. (1996). Immersion, Presence and Performance in Virtual Environments: An Experiment with Tri-Dimensional Chess. *Proceedings of ACM VRST*, pp. 163–172.

[10] Swan II, J.E., Gabbard, J.L., Hix, D., Schulman, R.S., & Kim, K.P. (2003). A Comparative Study of User Performance in a Map-Based Virtual Environment. *IEEE Virtual Reality '03 Proceedings,* Los Angeles, CA, pp. 259–266.

[11] Swartz, K., Thakkar, U., Hix, D., & Brady, R. (1999). Evaluating the Usability of Crumbs: a Case Study of VE Usability Engineering. *Proceedings of the 3rd International Immersive Projection Technologies Workshop*, May '99, Springer-Verlag, pp. 243–252.