The Effect of Stereoscopic Immersive Environments on Projection-based Multi-dimensional Data Visualization

Ronak Etemadpour*, Eric Monson† and Lars Linsen*
*Jacobs University, Bremen, Germany
†Duke University, Durham, NC, U.S.A.

***Abstract*—Multi-dimensional data impose a challenge for visual analyses. Commonly, dimensionality reduction techniques are used to project the multi-dimensional data into a 2D visual space. Poco et al. [9] showed that projection into a 3D visual space can increase the performance of common visual analysis tasks due to a higher projection precision. They also backed up their findings with a user study. However, when conducting the user study they displayed the 3D visual space on a 2D screen, which may impede the correct perception of the third dimension. In this paper, we present a study that investigates the effect of stereoscopic environments when used for the visual analysis of multi-dimensional data after projection into a 3D visual space. We conducted a controlled user study to compare correctness, timing, and confidence in segregation and precision tasks when performed in stereoscopic immersive environments and on a non-stereoscopic 2D screen. In terms of the stereoscopic immersive environments, we operated on and compared results obtained with two set-ups: a single screen and a six-sided highly immersive system, in both of which interaction was performed with a 3D input device. We investigated whether the stereoscopic immersive environments have an effect on user performance depending on the visual encodings. We used both 3D scatter plots and cluster visualizations in the form of enclosing surfaces or hulls for the visual analysis tasks.**

***Index Terms*—Stereoscopic Immersive Environment; High-dimensional data; 3D Projection.**

I. Introduction

Multidimensional data refer to data consisting of a set of objects each described by a vector of multiple attributes, where the number of attributes defines the number of dimensions. The data can be interpreted as a set of points in a multidimensional space. In analyses of multidimensional data, the goal is to gain insight into properties of the overall data distribution and find clusters that exhibit correlations among the dimensions or variables. Good perception of the clustering leads to an understanding of the overall distribution patterns and identification of unusual occurrences like outliers. The most common visualization techniques for multidimensional data are scatterplots and parallel coordinates. However, scatterplots have to be restricted to 2D or 3D visual spaces. If the dimensionality grows beyond three, multiple scatterplots, e.g., in form of a scatterplot matrix, may be used. However, clusters that do not align to the axes of the respective scatterplots can hardly be perceived. Thus, one often applies projection-based dimensionality reduction algorithms that map the high-dimensional attribute space to a 2D or 3D visual space. Many approaches to finding such projections exist, where the design goals include maintaining pairwise distances between points [15], maintaining distances within a cluster (i.e., keep one cluster together), or maintaining distances between clusters (i.e., keep two clusters separated) [14].

If algorithms prioritize cluster preservation, the idea is to project groups of similar points into segregated regions in the lower-dimensional space, which is commonly two-dimensional. Poco et al. [9] developed a 3D projection method by generalizing the Least Square Projection (LSP) technique from 2D to a 3D scheme, and showed that the cluster segregation is higher in the 3D visual space. First, they provided a quantitative analysis that confirmed that 3D projections lead to more precision compared to 2D projections. Second, they also conducted a user study to show that this additional theoretical gain also has an impact for practical analyses. The user study showed that users fulfill the tasks accurately and confidently in the 3D set-up, although all tasks were executed on a 2D screen, i.e., in the end the users were looking at 2D projections of the 3D projection. Through interaction users were able to profit from the additional information provided by the third dimension of the 3D projection. Still, the question arises whether performance would have been better had the users actually been using a 3D output device, i.e., a stereoscopic immersive environment.

This paper conducts a controlled user study to investigate users' performance when carrying out an interactive visual exploration of multidimensional data in a six-sided immersive virtual reality (VR) system. Users performed interactions in the VR environment that support the usual 3D transformations (i.e., zooming, translation, and rotation). The system set-up is described in Section III-E. We compare performance (accuracy) and speed in tasks, along with the users' confidence in their answers, when participants are using two different system configurations: a high fidelity (referred to as HD) system where all six walls of the VR are utilized, creating a maximally immersive environment; and a low fidelity (referred to as LD) system where only one wall is used, creating a less immersive environment. When referring to both set-ups without distinguishing between HD and LD, we refer to them as VR.

The goals of the paper are to evaluate the users' performance in the virtual environments when compared to the 2D

screen results reported by Poco et al., to evaluate the users' performance when comparing the HD and the LD set-ups, and to investigate the impact of the visual encoding. Poco et al. investigated cluster-based tasks by comparing 3D scatterplots, where clusters are encoded by color, to surface-based cluster encodings. They investigated four different surface types. Although there were some preferences for certain tasks, there was no generally best surface type. Consequently, we decided to carry out our study also using all four surface encodings (in addition to the color-coded scatterplots).

To allow for a direct comparison of our results to the ones reported by Poco et al., we had the users perform the same tasks. In addition, we identified a few extra tasks that we considered important in multidimensional data analysis, involving distance perception and cluster segregation. Section III describes the design of the study. While Poco et al. investigated one type of multidimensional data per study, we decided to use both types of datasets in a single study; namely document collection data and image collection data, which differ in terms of dimensionality and point density. Of course, when comparing to the results from Poco et al., we restrict our analyses to the tasks and datasets covered by their study. We formulated three hypotheses with respect to our three goals (VR vs. 2D comparison, HD vs. LD comparison, and comparison of visual encodings), see Section III-F. We compute accuracy of the user's answers with respect to a ground truth that is derived from the multidimensional datasets (i.e., before projection) and perform statistical analyses as described in Section IV. The results of our statistical analysis are detailed in Section V and the overall findings are summarized in Section VI.

II. Related Work

Many studies have confirmed that stereoscopic immersive environments create a unique feeling of presence for users. Some studies define presence as a measure of the effectiveness of a virtual environment. For example, Meehan et al. [6] found that heart rate and skin conductance changed when increasing the degree of presence. Greater presence evokes a greater psychological response, similar to that evoked in a real environment. Ware and Franck [16] showed the positive impact of 3D motion and stereoscopic viewing in understanding of abstract data. Other work disambiguates issues involved with physical and virtual navigation on large displays. Some of them confirm the trend of better performance and higher comprehension when mixing physical and virtual navigation using head tracking [10], [4]. A study by Ball et al. [3] concludes that for spatial visualizations, larger displays lead to more physical navigation and less virtual navigation. Researchers have also evaluated specific components of fidelity and improvements in user performance for spatial understanding tasks [18].

Classical information visualization, though, relies heavily on 2D visual representations. This is also true for most projection-based multidimensional data visualization methods. Several techniques specialize in representing groups of similar or dissimilar elements, e.g., the Nearest Neighbor Projection (NNP) [13] or the Least Square Projection (LSP) [7]. Many quantitative measures have been developed to estimate the quality of layouts produced by the various projection methods. For example, neighborhood hit [7] evaluates the capability of a projection to preserve the data neighborhoods found in the original space. The silhouette coefficient [11] evaluates the clustering capability. Other researchers like Tatu et al. [12] investigated user perception by conducting user studies on scatterplots. Users were asked to sort useful scatterplots among 18 instances. Albuquerque et al. [1] attempted to find a perception-based quality measure for scatterplots. A ranking function was used to estimate the value of the projections for a specific user task in a perceptual sense, based on the data from a psychophysical study. However, none of these estimates address user perception of 3D projections.

Some information visualization tools incorporate 3D data representations of multidimensional data. For example, Viz3d [2] creates a 3D representation of multidimensional data that is interactively manipulated by users to handle visual clutter and object occlusion. The 3D Grand Tour [17] generates a sequence of two-dimensional subspaces, moving continuously from one projection to the next, creating cluster-guided 3D data projections, and rendering the resulting visualizations in a CAVE virtual environment. They believe that this approach provides a natural metaphor to map the data onto a time-indexed family of 3D projections suitable for human exploration. Poco et al. [9] extended LSP to perform projections in 3D and created a strategy based on clustering and enclosing surfaces to interact with three-dimensional visual spaces. Quantitative metrics and user studies in that paper show that 3D projection enhanced cluster differentiation capability when compared to 2D displays. However, this higher correctness comes at the expense of spending more time for completing tasks. Still, users preferred to use the 3D system. Their results led us to extend this previous work to see whether an immersive 3D environment would improve users' ability to perform analytical tasks faster and/or more accurately.

III. Experimental Design of User Study

A. Data and Projection

Data with various characteristics should be used when performing analytical investigations of multidimensional projections. We used one dataset with lower dimensionality and higher densities, and a dataset with larger dimensionality and lower densities, described below. Pairwise distance plots (not shown) reveal that the distances are generally smaller in the lower-dimensional dataset. To allow for direct comparisons with Poco et al. [9], we used the same datasets: (1) A medical image dataset with 540 medical images (objects) obtained by magnetic resonance imaging. Each image is represented by 28 features (dimensions) including Fourier descriptors and energies derived from histograms, as well as mean intensity and standard deviation computed from the images themselves. (2) A document dataset with 681 scientific papers (objects), where each paper is described by the number of occurrences of 2,993 keywords (dimensions). While Poco et al. performed

the two parts of their study on one dataset each, we apply both parts of our study to both datasets. To produce clusters that are required for the tasks listed below, the datasets were clustered in the multidimensional space, i.e., before being projected, using an X-means approach [8]. The projection was performed using the 3D LSP technique [9].

B. Visual Encodings

For the visual encoding we compare colored 3D scatterplots, see Figure 1(a), with surfaces that capture each of the clusters. In the study by Poco et al. [9] four types of surface representations were used, which we reproduce here, see Figure 1(b)-(e). We define the following acronyms: (b) *ConvHull*: the convex hull of the points (of each cluster). (c) *PointsEncSurf*: an enclosing surface (of each cluster) that is isodistant to the cluster points. (d) *NonconvHull*: a non-convex hull (of each cluster) that is computed from a 3D Voronoi diagram of the cluster points. (e) *HullEncSurf*: an enclosing surface (of each cluster) that is isodistant to the non-convex hull above.

C. Set-up

The tasks have been divided in two different sessions with a five minute break in between. The first session was concerned with tasks on scatterplots, which involved tasks on individual objects as well as clusters, while the second session was concerned with tasks on different visual encodings, which involved only cluster-based tasks. During the first session each subject was trained in the VR conditions, which included a brief description of the system and an introduction to analyzing high-dimensional data. Then, the user study was conducted. Each subject was presented with one of the two fidelity conditions (HD or LD) and one of the two datasets (image or document data) based on their assigned study ID. Sitting outside of the VR system, the experimenter verbally delivered eight visualization-related questions that are described in Section III-D and eight confidence-related questions from the questionnaire corresponding to the presented dataset. A stopwatch was used to measure response times and the Qualtrics survey tool was used to record responses. Subjects were given 90 seconds to answer each visualization-related question and 15 seconds to answer each confidence-related question. Our user study involved 20 participants with different background. One user was a VR expert, while the others were regular computer users, but with no computer science background and no exposure to VR. The selection of participants is comparable to that of Poco et al. [9].

D. Tasks

The tasks given to the participants include typical analysis operations for multidimensional data. They involve individual objects and clusters and try to capture perceptions of distances/similarities, densities, and distributions.

Q1 Count the clusters.
Q2 Find closest cluster to a specific point.
Q3 Find closest cluster to a specific cluster.
Q4 Detect the densest cluster.
Q5 Find the most distant cluster pair.
Q6 Count the outliers.
Q7 Find the closest cluster to selected group of points.
Q8 Name all of the pairs of overlapping clusters.

All eight questions were used for the first session. For the second session we used only the cluster-based tasks, i.e., Q1, Q3, a repetition of Q3 with a different cluster, Q5, and Q8. Tasks Q1, Q2, Q4, and Q8 were taken from the first session of the study by Poco et al. [9], and tasks Q1, Q5, and Q8 from their second session. We added Q3 and Q7 to also account for nearest neighbor searches for clusters or groups and not just for individual objects (cf. Figure 2(a)), and we added Q6 to also consider outliers. In general, tasks Q1 and Q6 involve pattern identification, tasks Q2, Q3, and Q7 investigate neighborhood (similarity) relationships, tasks Q5 and Q8 are clutter estimation tasks, and task Q4 involves density estimations.

E. System

We evaluated both high level of fidelity (HD) displays and low level of fidelity (LD) displays to find the impact of these on user performance. We used a CAVE technology: a 3 x 3 x 3m fully immersive [4] system with 6 stereoscopic rear projected walls, head and hand tracking, and real time computer graphics. Participants wore liquid crystal shutter glasses (CrystalEyes 3, RealD) which synchronize with the projectors (Christie Digital Mirage S+2k DLP) through an infrared signal, presenting left/right eye images time sequentially. Head location and orientation were measured (Intersense IS-900 inertia/ultrasonic tracking system). User interaction was performed through a 3D mouse, called a wand (Intersense). A video of the explained interaction can be found online [1]. In HD mode, the user was surrounded by six stereoscopic VR display walls with full 360-degree field of regard (FOR). In the LD case, only one display wall was used, so the user could not move physically around the object due to the reduced FOR. Following McMahan et al. [5], we designed our experiment to reduce confounds when comparing HD and LD. The easiest way to bring data into the VR system was to use the Syzygy library, which uses OpenGL (C++ code compiled with MinGW under Windows). To transfer the geometry, we used Python and VTK (the Visualization ToolKit, vtk.org) to load saved PolyData (.vtp) files, glyph the points with spheres, color them according to their cluster, and export them as OBJ files. Figure 2(b) shows an example of the VR simulator display that looks forward into the 10-foot cubical space with the CAVE shown as a wireframe in the lower-right corner. Points or groups of points were highlighted using a white semi-transparent sphere (cf. Figure 2(c)).

F. Hypotheses

Our first analysis goal was to compare the virtual environments to the rendering on a 2D screen. We can state that VR creates a perception of presence and awareness of being

[1] http://vcgl.jacobs-university.de/wp-content/uploads/2013/05/Projection-Immersive.mp4

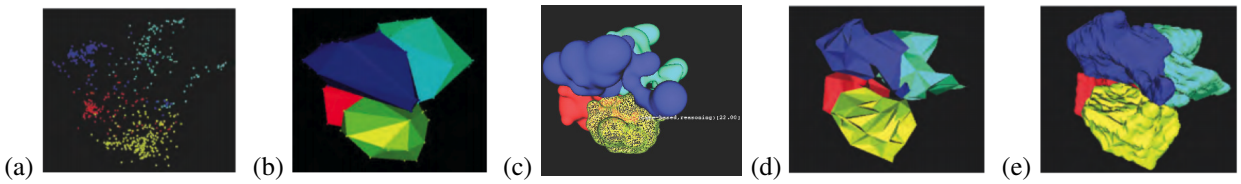

Fig. 1. Visual encoding of the clusters in 3D visual space using (a) Points, a 3D scatterplot with clusters encoded by color., (b) ConvHull, (c) PointsEncSurf, (d) NonconvHull, and (e) HullEncSurf.

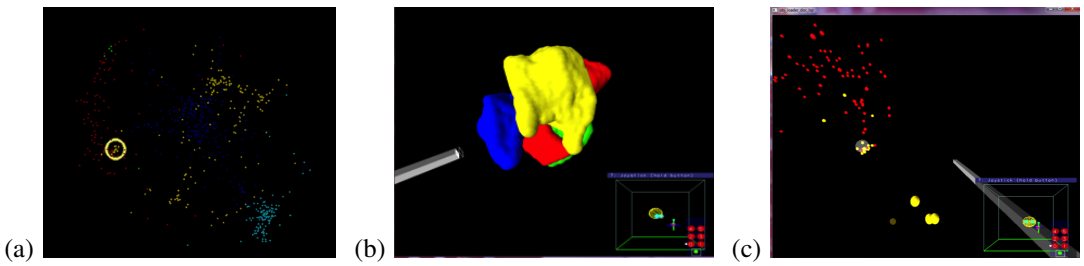

Fig. 2. (a) Example a 2D screen task: Q7 asks for neighborhoods of a group of selected points. (b) Simulator set-up: Cubical CAVE shown as wireframe in the lower-right corner, where the sphere indicates the head of the user. Interaction with HullEncSurf visual encoding. (c) Simulator set-up for Q7 with semi-transparent sphere indicating group of interest.

surrounded in an immersive environment. Hence, we believed that VR environments would improve the performance when compared to 2D screens. This general assumption is tested in detail in three hypotheses that address the different aspects of the user tasks.

Most intuitively, we felt that the stereoscopic viewing should allow for a better distance perception and formulated our hypothesis as:

H1a *Rendering 3D projections in a VR environment will improve performance on similarity tasks when compared to a 2D screen.*

Next, we felt that the improved depth perception in stereoscopic viewing should allow the users to better perform tasks that involve dealing with cluttered views. We formulated the hypothesis:

H1b *Rendering 3D projections in a VR environment will improve performance in the presence of clutter when compared to a 2D screen.*

Finally, we wanted to check whether the performance also improved in case of more involved, global analysis tasks such as the identification of patterns or detecting areas of highest densities. We formulated the hypothesis:

H1c *Rendering 3D projections in a VR environment will improve performance on global analysis tasks when compared to a 2D screen.*

Our second analysis goal was to compare different VR settings. In the maximally immersive HD environment, human perception and interactivity are quite natural, creating a sufficient belief that the environment is real. We believe that HD provides an improved capability for relation-seeking tasks of cluster segregation and pattern organization. Consequently, we formulated this hypothesis as:

H2 *Using an HD set-up will improve performance when compared to LD.*

Our third analysis task was to investigate the role of visual encoding. We compared point-based and surface-point renderings of clusters. Poco et al. [9] investigated the performance of four surface-based approaches compared to color-coded point clouds. The results did not reveal significant differences among approaches in terms of accuracy, except for one single task and one single dataset. Overall, the study concluded that point clouds are a viable alternative to surface renderings. This is surprising because of the difficulty of depth perception in 3D scatterplots. When using VR settings, we assume that the depth perception in 3D scatterplots is even better. Thus, we formulated our last hypothesis as:

H3 *Point- and surface-based cluster visualizations will work similarly well in VR.*

IV. Investigations and Statistical Methods

In order to apply statistical methods on the accuracy, we needed to compute various ground truths for the data. One was computed by estimating (pair-wise) distances (Euclidean for image data; cosine for document data) and densities via a minimum spanning tree in the high-dimensional space, i.e., before projection. The clusters were also computed before projection. Given the ground truths, we computed the errors in the participants' answers for each task.

For the tasks that required the subjects to estimate a number (Q1, Q4, Q5, and Q6), the error percentage was computed by:

$$e = \frac{|n_{true} - n_{answer}|}{n_{true}} \cdot 100,$$

where n_{true} is the estimated ground truth and n_{answer} is the reported answer. For Q2 and Q7, the error is either zero or one. To analyze the results for Q8, correctness was computed: the answer were 100% correct if users could identify all pairs of overlapping clusters accurately without naming any extra pairs

of clusters; otherwise, the number of incorrectly identified clusters (false positives and false negatives) reduced the score by the respective percentage.

For all analyses, we computed means and standard deviations of the accuracy (errors or correctness). We used the Shapiro-Wilk test, which is appropriate for small sample sizes (<50 participants). When the dependent variable was not normally distributed, the Mann-Whitney U Test was used to compare differences between two independent groups (e.g., participants who had seen an image in LD vs. participants who had seen the same image in HD). If the dependent variable had a normal distribution, an independent t-test was chosen. The Wilcoxon Signed Ranks Test was used for comparing within groups in the case of a non-parametric distribution, and the t-test in the case of a parametric distribution. In the second session, when comparing five different rendering methods, an ANOVA test was used in the case of a normal distribution, and a Tukey post-hoc test revealed pairwise significant comparisons. Furthermore, the Kruskal-Wallis test was used in the case of non-normal distributions for comparing the five different cluster visualization techniques.

V. Results

Here we present our detailed results. In all charts, values are mean quantities, while error bars show the standard error from the mean. Also, we use the convention that results for VR are encoded in green, 2D screen in purple, HD in red, and LD in blue. Moreover, the outcome of a pairwise significance test is indicated by horizontal lines under the bars using both position and color coding. So, for example, if in a pairwise HD-LD comparison HD had a significant better result, the color is red like the HD bars and is placed under an HD bar, whereas if LD was significantly better, the line would be blue and would be placed under the LD bar. Statistical significance is judged at p-value < 0.05.

A. Accuracy

We first investigate accuracy results in VR settings compared to 2D screens before comparing the two VR settings, HD and LD. In both comparisons, we first look into the 3D scatterplot tasks (Session 1), before delving into the issue of visual encoding (Session 2).

1) VR vs. 2D screen: **Session 1.** Since Poco et al. [9] only used the document dataset in their user study, which may be biased because of the characteristics of document data, we first investigated whether there were any significant differences in our results when comparing document data vs. image data. We actually observed that the differences in the findings were significant in 3 out of 4 tasks (Q1, Q2, and Q4). Hence, the nature of the data does play a role. Consequently, we had to restrict ourselves to the document data when comparing to the 2D screen results reported by Poco et al. Also, since Poco et al. only investigated tasks Q1, Q2, Q4, and Q8 for this session on 3D scatterplot analysis, we only compare the results for those tasks.

To test hypothesis H1a, we looked at the results of the similarity task, Q2. Users performed (statistically) significantly better (U = 52, p = 0.002) in VR when compared to the 2D screen, see Figure 3(b). The same holds true when comparing only HD against 2D or only LD against 2D. Thus, our results point toward the validity of hypothesis H1a.

To test hypothesis H1b, we investigated the clutter-estimation task, Q8. The results of a Mann-Whitney test showed that the accuracy in the LD condition was significantly higher than with the 2D display (U = 27, p = .025), as shown in Figure 3(d). According to this result, hypothesis H1b appears to be valid. However, when comparing HD and LD together against 2D, the improvement was not statistically significant, see Figure 3(c). Hence, we plan on conducting a follow-up study with more subjects to confirm the hypothesis.

To test hypothesis H1c, we investigated tasks Q1 and Q4, which involved a global investigation on the 3D scatterplot instead of looking into local distances or overlaps, see Figure 3(a). Apparently, task Q1 shows significant difference based on a Mann-Whitney U test (U = 18, p = .001) that evaluates whether the medians on a test variable differ significantly between two groups. The error in the number of clusters in HD and LD conditions had an average rank of 15.70, while the computer screen (2D display) had the average rank of 8.00. Note that the results of the test were not in the expected direction. Against our hypothesis, users made mean error=0 on a 2D screen, which resulted in a statistically significant lower mean error compared to VR. For the density task Q4, since there was a large spread in the answers, we normalized the estimated errors to the interval [0,1] by dividing them by the maximum error reported (0.01056). Although the mean rank error in the VR condition is lower (15.58) than on the 2D display (18.04), the results of a Mann-Whitney test show an insignificant difference for all pairwise comparisons (2D vs. LD, 2D vs. HD, and 2D vs. VR). Hence, hypothesis H1c needs to be rejected.

Session 2. For this session, we compared exactly the tasks that had been used in the second session evaluations by Poco et al. [9]. In particular, they performed their study only on the image data and only investigated tasks Q1, Q5, and Q8.

Considering Q1, a significantly larger mean error in VR when compared to the 2D display was found for the visual encoding using Points (U = 12.00, p = .004), NonconvexHull (U = 30.00, p = .025), and HullEncSurf (U = 24.00, p = .008), see Figure 4(a). So, we can conclude that the task of counting the clusters was not just better solved for visual encodings with Points but also for two of the surface-based visualization techniques in 2D displays. The surface-based methods helped to improve the performance in VR, but the 2D display did overall better on this task. For Q5, we did not find a significant difference between the 2D display and VR for any of the five visual encodings. However, when looking into the results of Q8, the ConvexHull visualization showed significantly larger correctness in VR when compared to the 2D display (U = 18.00, p = 0.014) as illustrated in Figure 4(b). One can state that the VR environment, at least,

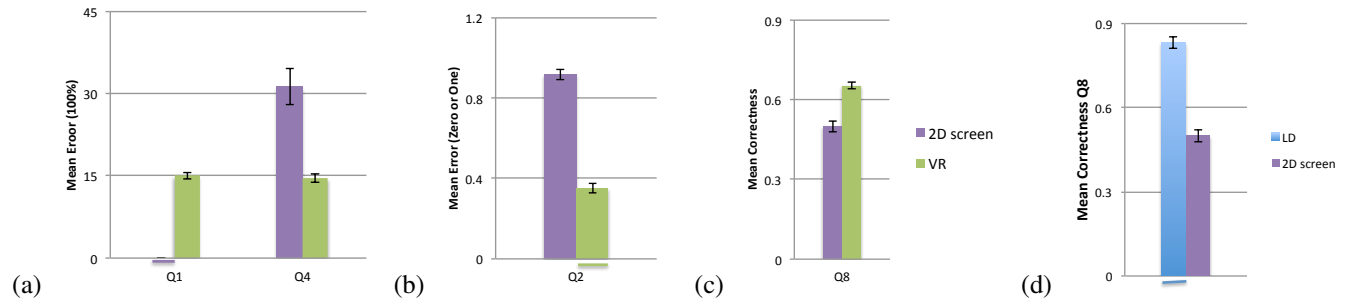

Fig. 3. Accuracy of VR vs. 2D screen for 3D scatterplots (Session 1). (a) Mean error percentage for cluster counting (Q1) and densest cluster determination (Q4). (b) Mean error values for identifying the similarity to an instance (Q2). (c) Mean correctness value for listing all pairwise overlaps (Q8). (d) Mean correctness for listing all pairwise overlaps (Q8) for only LD vs. 2D screen.

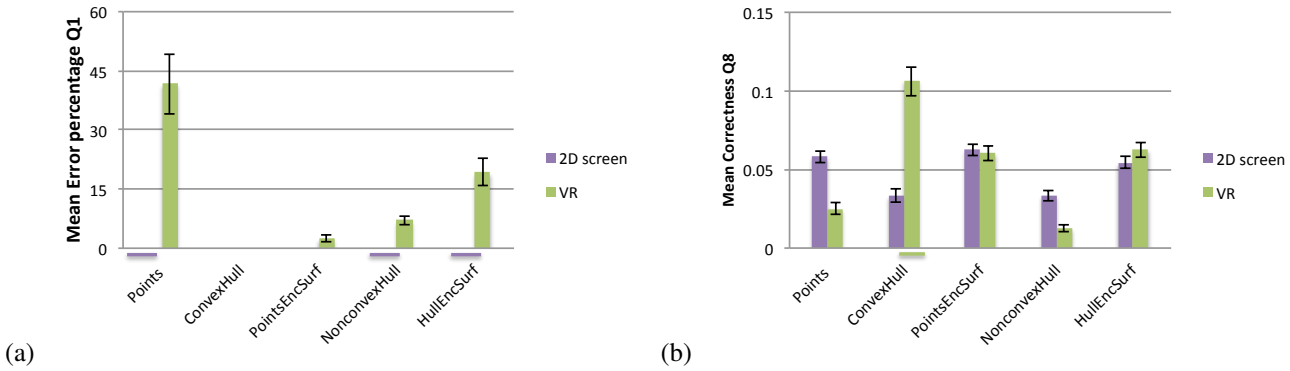

Fig. 4. Accuracy of VR vs. 2D screen for visual encodings (Session 2): (a) When considering Q1 (number of clusters), 2D display had significantly less mean error for Points, NonconvexHull, and HullEncSurf. (b) When considering Q8 (overlapping clusters), VR was significantly more accurate than 2D display for ConvexHull.

had a significant improvement for one of the configurations.

2) HD vs. LD: **Session 1.** Next, we compare the two VR settings, i.e., the accuracy of the answers for 3D scatterplots in HD and LD set-ups. Since here we are not comparing against the work by Poco et al., we can consider both data sets and all eight questions. Figure 5 summarizes the results. The bar charts in (a) show the mean error for the questions that generate percentage error, the bar charts in (b) show the error for questions that generate errors zero or one per answer, and the bar charts in (c) show mean correctness. It can be observed that HD for Q4 is ranked better when compared to LD. The Wilcoxon Signed Ranks test shows significantly less mean error for the HD condition ($Z = -1.996$, $P = 0.046$). Also, Q3 has significantly less mean error in HD ($Z = -2.449$, $P = 0.014$) when compared to LD. Investigations for the other tasks do not deliver any significant differences in LD vs. HD. It is worth noting that when only considering image data for Q2, there is significantly less mean error ($U = 30$, $P = 0.029$) in LD compared to HD, but it is statistically insignificant ($Z = -1$, $P = 0.317$) over the combination of both datasets. Hence, we conclude that HD could improve the performance of some of the visual analysis tasks, but not in general.

Session 2. For the analysis of the visual encodings we could again use both datasets and all cluster-related tasks, i.e., Q1, Q3, Q5, and Q8, where Q3 was repeated with another cluster selection. Looking into Q1, LD vs. HD comparisons did not reveal a significant difference among mean errors when investigating each visualization individually. However, within HD, a statistically significant difference between different visualizations ($P = 0.048$) was reported based on the median test. We chose the median test here, as it is more robust to violations of normality and homogeneity of variances in comparison to the Kruskal-Wallis. Points and ConvexHull show the most different mean errors as illustrated in Figure 6(a) and ConvexHull is significantly better than Points. In LD, however, the five visualizations do not have any significant differences in accuracy.

Since Q3 and repeated Q3 investigate the perception of similarity recognition and no significant differences among mean errors were found individually, we integrated them as a single measure for comparison. The results from Wilcoxon Signed Rank test ($Z = -1.414$, $P = 0.157$) did not reveal any significant difference for the whole dataset between HD and LD. For Q8, pairwise comparisons between HD and LD showed a significant difference ($t(3) = 4.970$, $P = 0.016$) for Points, which is quite interesting: As discussed before, the results for Points in Session 1 did not reveal any significance when dealing with 12 different clusters, but in Session 2 results have significance when involving 5 clusters for document data and 6 clusters for image data. However, no significant differences

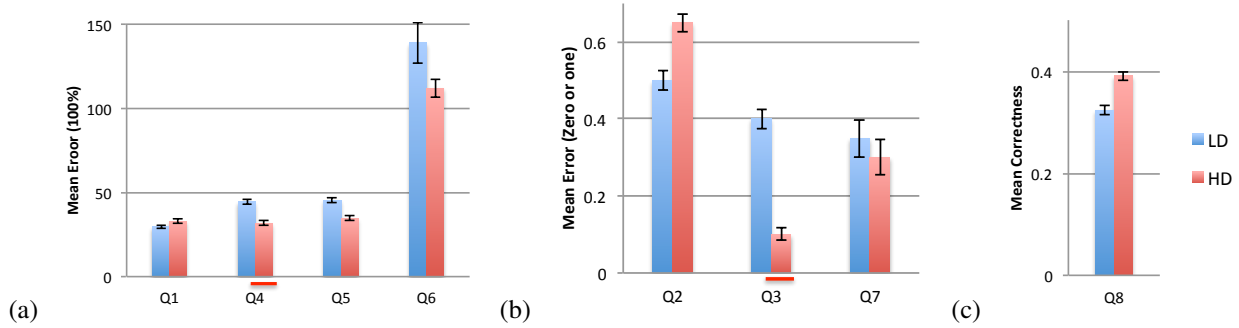

Fig. 5. Accuracy HD vs. LD for 3D scatterplots (Session 1). (a) Mean error values for Q1, Q4, Q5 and Q6. (b) Mean error values for Q2 and Q3 and Q7. (c) Mean correctness value for Q8.

were found among the five visualizations within HD and LD.

Furthermore, since the Q5 and Q8 results are complementary to each other, they have been integrated to create a single distance measure, and mean correctness has been analyzed. Applying a normality test showed a normal distribution, such that T-test paired samples were chosen to compare HD vs. LD conditions. With a T value = -0.123, we have 19 degrees of freedom and the sig value is 0.903. However, as shown in Figure 6(b), an ANOVA one-way test showed a statistically significant difference among visual encoding techniques ($F(4,15) = 5.055$, $p = .009$) in HD. A Tukey post-hoc test revealed that the accuracy was statistically significantly greater for HullEncSurf when compared to Points ($P = .004$). No significant differences were found among the other techniques.

B. Confidence

Confidence values are given in form of a 5-step Likert scale (5 being best, 1 being worst). When looking into the average confidence values for all tasks, all data sets, and all participants for the 3D scatterplots in Session 1, no statistically significant difference based on Mann-Whitney U test was observed ($U = 55.000$, $P = 0.737$) in VR vs. 2D screen. Likewise, the independent t-test delivered no significant difference in the confidence levels when comparing HD vs. LD.

When looking into the different visual encodings in Session 2, an ANOVA test showed no significant differences ($F(4,15) = 0.103$, $p = 0.980$) between HD and LD when grouping all visual encodings together. Similarly, Kruskal-Wallis showed no significant difference in VR vs. 2D screen ($H(2) = 3.585$, $P = 0.465$), see Figure 7. However, when looking into each visual encoding separately, statistical tests showed significant differences in VR vs. 2D display for ConvexHull ($P = 0.044$) and PointsEncSurf ($P = 0.027$), see Figure 7. Although participants had no major experience with the VR environment, in two out of five visual encodings they felt more confident when using it.

C. Time for task completion

Figure 8 shows the summary of the timings for task completion. Wilcoxon test is used to analyze the difference of timings for HD vs. LD for the 3D scatterplots (Session 2). No significant difference was revealed. Similarly, for Session 2 on visual encoding, an ANOVA test did not exhibit significant differences in HD vs. LD ($F(4,15) = 4.467$, $p = .109$). However, there are some significant differences between VR vs. 2D display when looking at each task of Session 1 (3D scatterplots) individually. Tests on the Q8 results show that time is significantly higher in the 2D display than in VR ($U = 17.000$, $P = 0.000$), while time in VR is higher compared to 2D display for Q2 ($U = 60.000$, $P = 0.019$). For the second session (visual encodings), fulfilling the tasks for Points ($P = 0.004$), PointEncSurf ($P = 0.029$), and HullEncSurf ($P = 0.031$) take longer in VR when compared to the 2D display. In summary, there is no general longer task completion time in VR, but certain tasks did take longer.

VI. Findings

In this section, we attempt to interpret the results relating back to the formulated hypotheses, draw conclusions about the findings, and provide guidelines.

Hypotheses H1a-H1c were concerned with comparing VR to 2D displays. While hypotheses H1a and H1b were accepted (with a caveat for H1b), hypothesis H1c needed to be rejected. Hence, we can conclude that distances between individual objects can be perceived better in VR, which led to an overall improved performance for local analysis tasks that focus on a specific part of the visuals. Global analysis tasks that require the user to comprehend the distribution of all points in the 3D scatterplot had no significantly better or worse performance in VR. It seems that the participants had trouble maintaining the overview or big picture when immersed in the VR environment. It may be a subject for a long-term study to find out whether this issue improves when getting more familiar with the environment. A small evaluation across the population in this study showed that mean error value for the one VR expert is significantly less (mean error of 0.092 with standard error of 0.008) compared to the rest of non experts' (mean error of 0.227 with standard error of 0.037). The missing training can also be observed when looking into the timings, where 2D displays more often outperformed VR than the other way round. On the other hand, participants seem to have a slight preference for the VR system, since 2D

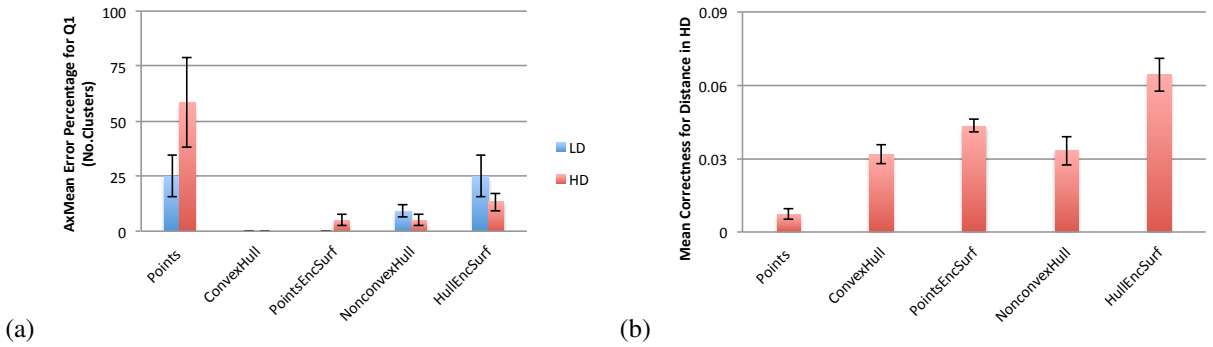

Fig. 6. (a) Accuracy in HD vs. LD screen for visual encodings (Session 2) when considering Q1: No significant difference when comparing HD vs. LD for each task individually. ConvexHull had significantly less mean error than Points in HD but not in LD. (b) Accuracy in HD for visual encodings (Session 2) when considering Q5 and Q8: HullEncSurf was significantly better than Points

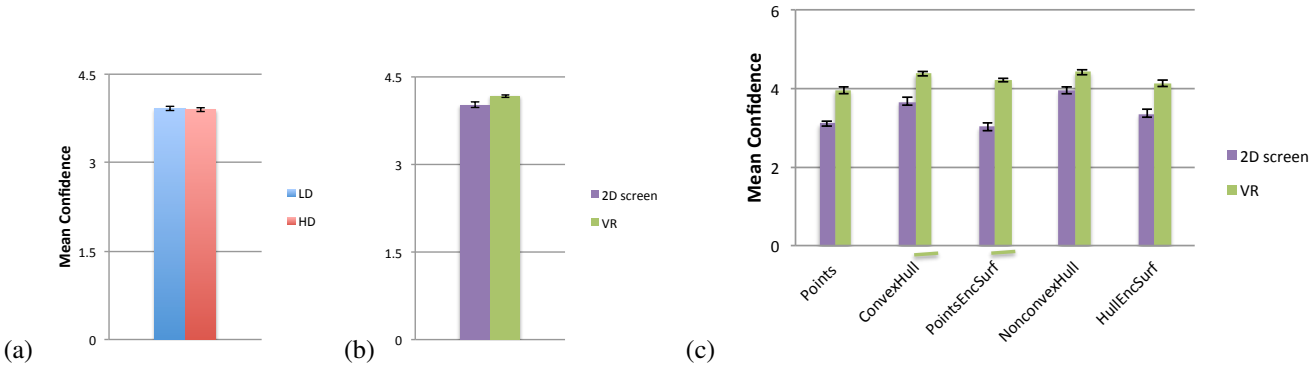

Fig. 7. Confidence for visual encodings (Session 2): (left) No significant difference between HD and LD. (middle) No significant difference between VR and 2D display. (right) VR was preferred for ConvexHull and PointsEncSurf .

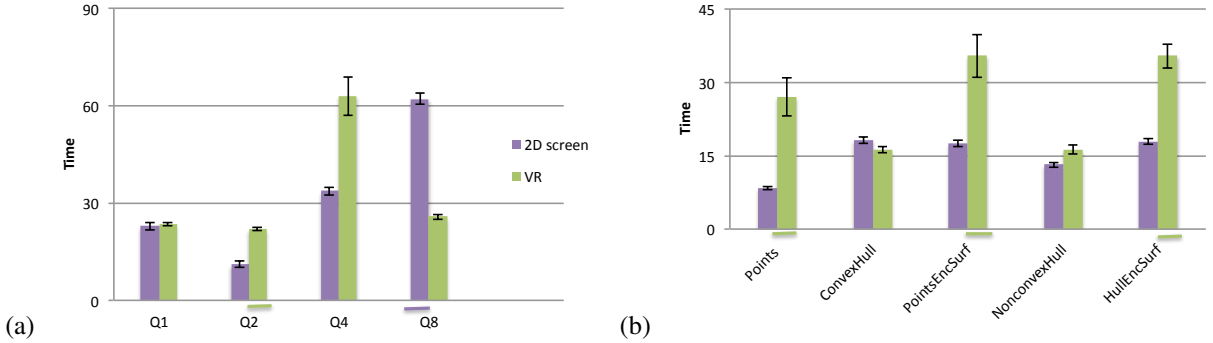

Fig. 8. Time for VR vs. 2D screen: (a) Session 1 on 3D scatterplots. (b) Session 2 on visual encoding.

displays never outperformed VR with respect to confidence.

Hypothesis H2 speculated that HD can outperform LD due to the larger FOR. Indeed, LD never had significantly higher accuracy than HD, but HD only had a few tasks where it had significantly improved accuracy over LD. Also, in time and confidence, there was no significant difference between the two settings. Still, we can conclude that HD can improve performance.

Hypothesis H3 assumed that performance when using the Points visual encoding would be boosted by the increased depth perception in VR. To our surprise this was not the case. While there were no significant differences between Points and the surface-based techniques in 2D displays, there was one surface-based technique that did significantly better than Points in HD. The performance of Points actually dropped in comparison to surface-based techniques when going to high-fidelity VR systems. Thus, this hypothesis could not be confirmed.

In addition, we want to note that the characteristics of the data set matters and we needed to separate image data and document data for our analyses when comparing to Poco et al. We also confirmed the findings in Poco et al. that the optimal

choice of the surface for visual encoding depends on the task, but in contrast to their findings, Points did not perform as well in our study (especially in HD).

VII. Conclusion

We have conducted a controlled user study to evaluate the role of perception when using a 3D projection for multidimensional data visualization. We investigated typical analysis tasks and data types in two stereoscopic immersive environments. We were able to identify groups of tasks where an immersive environment improved performance. This was true for local analysis tasks, but not for global ones. The results of our study show that the levels of display fidelity and interaction can also be a factor in determining performance on tasks involving spatial relationships in data. Surface-based visual encodings profit more from the VR environment than point-based renderings.

Acknowledgment

This work was supported by the research center on Visual Communication and Expertise (VisComX) at Jacobs University, Bremen, Germany as well as NSF CCF-0808847. We would like to thank Ryan McMahan, Rachael Brady, Victoria Szabo, and David J. Zielinski for their kind help in conducting this study at Duke University, Durham, USA. We also thank Rosane Minghim, Maria Cristina Ferreira de Oliveira, Jorge Poco, and Robson Carlos da Motta from Universidade de São Paulo, São Carlos, Brazil, for their efforts.

References

[1] Georgia Albuquerque, Martin Eisemann, and Marcus Magnor. Perception-based visual quality measures. In *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 13–20, October 2011.

[2] Almir Olivette Artero and Maria Cristina Ferreira de Oliveira. Viz3d: Effective exploratory visualization of large multidimensional data sets. In *SIBGRAPI 04: Proceedings of the Computer Graphics and Image Processing, XVII Brazilian Symposium*, pages 340–347, 2004.

[3] Robert Ball, Chris North, and Doug A. Bowman. Move to improve: Promoting physical navigation to increase user performance with large displays. In *Proceedings of SIGCHI conference on Human factors in computing systems*, pages 191–200, 2007.

[4] Carolina Cruz-Neira, Daniel J. Sandin, and Thomas A. DeFanti. Surround-screen projection-based virtual reality: The design and implementation of the cave. In *Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH, New York 1993*, pages 135–142, 1993.

[5] Ryan P. McMahan, Doug A. Bowman, David J. Zielinski, and Rachael B. Brady. Evaluating display fidelity and interaction fidelity in a virtual reality game. *IEEE Transactions on Visualization and Computer Graphics*, 18:626–6330, 2012.

[6] Michael Meehan, Brent Insko, Mary Whitton, and Frederick P. Brooks Jr. Physiological measures of presence in stressful virtual environments. In *Proc. ACM Siggraph, ACM Press*, pages 645–652, 2002.

[7] Fernando Vieira Paulovich, Luis Gustavo Nonato, Rosane Minghim, and Haim Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):564–575, 2008.

[8] Dan Pelleg and Andrew Moore. X-means: Extending k-means with efcient estimation of the number of clusters. In *Proceedings of the 17th. International Conference on Machine Learning, ICML '00*, pages 727–734, 2000.

[9] Jorge Poco, Ronak Etemadpour, Fernando V. Paulovich, Tran Van Long, Paul Rosenthal, Maria Cristina Ferreira de Oliveira, Lars Linsen, and Rosane Minghim. A framework for exploring multidimensional data with 3d projections. In *Computer Graphics Forum*, volume 30, pages 1111–1120, 2011.

[10] Dheva Raja, Doug A. Bowman, John Lucas, and Chris North. Exploring the benefits of immersion in abstract information visualization. In *Proceedings of Immersive Projection Technology Workshop*, May 2004.

[11] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley Longman, Boston, MA, USA, 2005.

[12] Andrada Tatu, Peter Bak, Enrico Bertini, Daniel A. Keim, and Jörn Schneidewind. Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '10)*, pages 49–56, 2010.

[13] Eduardo Tejada, Rosane Minghim, and Luis Gustavo Nonato. On improved projection techniques to support visual exploration of multidimensional data sets. information visualization. 2(4):218–231, 2003.

[14] Joshua B. Tenembaum, Vin de Silva, and John C. Langford. A global geometric faramework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[15] W. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika.*, 17(4):401–419., 1952.

[16] Colin Ware and Glenn Franck. Evaluating stereo and motion cues for visualizing information nets in three dimensions. *ACM Transactions on Graphics*, 15(2):121–140, 1996.

[17] Li Yang. 3d grand tour for multidimensional data and clusters. In *IDA '99 Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis*, pages 173–186, 1999.

[18] Yei-Yu Yeh and Louis D. Silverstein. Spatial judgements with monoscopic and stereoscopic presentation of perspective displays. *Human Factors*, 34(5):583–600, 1992.