Effects of Stereo and Screen Size on the Legibility of Three-Dimensional Streamtube Visualization

Jian Chen, *Member, IEEE*, Haipeng Cai, Alexander P. Auchus, and David H. Laidlaw, *Senior member, IEEE*

(a) FA   (b) TRACING   (c) NAMING

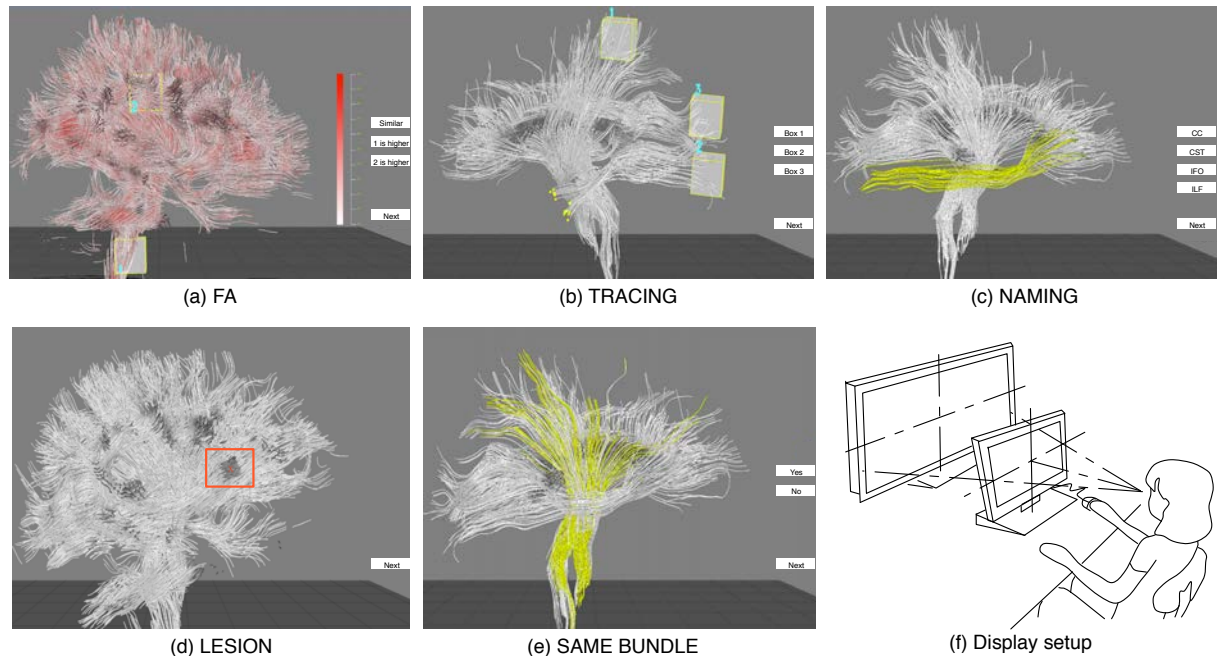(d) LESION   (e) SAME BUNDLE   (f) Display setup

Fig. 1. Five 3D DMRI visualization tasks on two display devices.

Abstract—We report the impact of display characteristics (stereo and size) on task performance in diffusion magnetic resonance imaging (DMRI) in a user study with 12 participants. The hypotheses were that (1) adding stereo and increasing display size would improve task accuracy and reduce completion time, and (2) the greater the complexity of a spatial task, the greater the benefits of an improved display. Thus we expected to see greater performance gains when detailed visual reasoning was required. Participants used dense streamtube visualizations to perform five representative tasks: (1) determine the higher average fractional anisotropy (FA) values between two regions, (2) find the endpoints of fiber tracts, (3) name a bundle, (4) mark a brain lesion, and (5) judge if tracts belong to the same bundle. Contrary to our hypotheses, we found the task completion time was not improved by the use of the larger display and that performance accuracy was hurt rather than helped by the introduction of stereo in our study with dense DMRI data. Bigger was not always better. Thus cautious should be taken when selecting displays for scientific visualization applications. We explored the results further using the body-scale unit and subjective size and stereo experiences.

Index Terms—Display characteristics, diffusion tensor MRI, virtual environment.

---◆---

1 INTRODUCTION

Recent advances in display technologies for scientific visualization have allowed once exotic and expensive techniques to become so inexpensive, accessible, and lightweight that they can be used directly in scientific research. Visualization researchers have increasingly been

- *Jian Chen is with University of Maryland Baltimore County, e-mail: jichen@umbc.edu.*
- *Haipeng Cai is with University of Southern Mississippi, e-mail: haipeng.cai@eagles.usm.edu.*
- *Alexander P. Auchus is with the University of Mississippi Medical Center, e-mail: aauchus@umc.edu.*
- *David H. Laidlaw is with Brown University, e-mail: dhl@cs.brown.edu.*

using advanced displays to build applications and toolkits in three dimensions (3D) using a variety of techniques [25]. Our collaborators in brain sciences are also excited about making use of such displays in data analysis (Fig. 1). However, the DMRI data are often highly dense, imposing greater visualization challenges to produce legible visualizations. Nevertheless, usability studies are necessary for visualizations to reach their full potential; guidelines are needed concerning the added value or appropriateness of alternative solutions that will lead to more fundamental insights into why a particular solution is effective [23].

Brain researchers are using DMRI techniques to study human brain structure in pathological conditions, such as stroke and Alzheimer's disease. DMRI is a MRI technique that measures the directional dependence of motion of water molecules in tissue. Experimental evidence has shown that water diffusion is anisotropic in organized tissues, such as white matter and muscle, and that reconstructing the orientation and curvature of white matter can provide detailed information about pathways. The curves (or fibers) are portrayed graphically

using streamline algorithms or glyphs such as hyperstreamlines initialized at seed points to show fiber tracts. Tracts following similar directions are called fiber bundles [38]. Displaying the fiber tracts as tubes is a popular way to visualize DMRI data. Given the advances in image capturing and processing techniques, we can display human brain features at millimeter scales, making the visualization highly dense, where a whole-brain tractography can have about ten thousand tubes within the volume of a human head.

Numerous studies have found benefits in using large stereoscopic displays. Most would agree that large displays lead to fundamentally different user experiences [7]. Some have reported benefits in task completion time when using large displays and have suggested that people tend to use more egocentric navigation strategies on large displays and hence improve their task performance [42]. This is true even when a semi-immersive display with small field-of-view (FOV) is used for tasks requiring mental rotation. Psychophysics studies also suggested that bigger is better when the amount of information is the same, because visual acuity increases with distance through the so-called Aubert-Forster law [6]. The law states that "objectively small objects can be distinguished as two at greater distances from the fovea than objectively larger objects subtending the same visual angle" (page 471 in [35]). These interesting results suggested that we should always choose large displays.

The overarching objective of our work is to systematically understand the relationships between display characteristics and dense data visualization. The present study is the first to study dense streamtube visualizations on size and viewing distance tradeoffs and the uses of stereopsis to understand whether, if we equalize the retinal images between conditions, we need to have a large display. Our work is motivated by the visual complexity of the streamtubes brain researchers use in answering their scientific questions. We sampled every voxel of a DMRI capture from a normal person, yielding tractography with about ten thousand dense lines rendered using tubes and displayed under two monitors: a small $24''$ and a large $72''$ displays with or without stereo (Fig. 2). The retinal images were kept similar in size between conditions, at least between the large and small displays under the same mono/stereo conditions by carefully arranging FOV and viewing distance. Participants could rotate the data during the experiment.

The evaluation covered major and representative fiber bundles in the brain regions suggested by neurologists; the metrics for user task performance were accuracy, task completion time, and subjective comments in post-study interviews. Our data were carefully selected by computer scientists and brain researchers working together.

This paper contributes to the growing literature on design and evaluation of scientific visualization using large displays. It describes new results in which experimental evidence on display characteristics was systematically collected, and discusses how to explore this evidence to examine human task performance and guide display choices based on stereo and size when the retinal images are the same. This paper also contributes to experimental task collection for diffusion MRI studies, as we have made the datasets publicly accessible for benchmarking future experimental studies [8]. In addition, applying these results appropriately to design can also increase the usefulness of high-impact large-display applications in practical scientific visualization.

2 BACKGROUND AND RELATED WORK

Many people have addressed, with quantitative or anecdotal advice, how best to use displays for a variety of cognitive and user performance benefits in 3D [7]. In general, it is believed that when the complexity of 3D data increases, stereoscopic display often provides better insight into the datasets. A major benefit of stereoscopy is binocular disparity that provides a better depth awareness. For example, Ware and Franck studied stereo, interaction, and motion in a network graph data visualization [46]. A stereo display was found to be 1.6 times more accurate than a two-dimensional (2D) display in detecting paths of length two through the complex structures, and stereo combined with head-coupled motion produced the best results. In their visualization comparison, the graph sizes ranged from 24 to 132 nodes with 32 to 176 arcs where participants could still see the structure clearly.

In the present study, where the number of tubes used could reach thousands, we expected some improvement of stereo over 2D but perhaps not as much as in Ware and Franck's study [46].

Bowman summarized and comprehensively discussed the factors related to immersion that can have an impact on visualization experiences [7]. These factors include FOV, field of regard (FOR), display size, stereoscopy, head-based rendering, realism of lighting, and frame and refresh rates. Laha et al. studied three of these (head tracking, FOR, and stereoscopic rendering) in the volume rendering of mouse limb and fossil datasets [24]. Their experimental tasks were mostly open-ended; participants had 1 to 1.5 minutes to describe what they could see and the experimental results were evaluated by experts for quality. Though their study suggested that high levels of FOR, stereo, and head tracking improved task performance in general, they also found stereo worsen task performance on internal feature search and general descriptive tasks. Many of our DMRI structures are related to internal structures in dense tubes, and it is thus worth considering whether or not stereo is helpful when data are highly dense.

Many studies have addressed the benefit of displaying images in large size. For one thing, large size can increase FOV when users are free to move. And the benefit of large FOV seems to be substantial. For example, a large field of view (FOV) has been found to improve task performance and generate better situation awareness and presence in navigating 3D environments [42]. Pausch et al. attributed the benefits to a sense of presence that inspires more efficient egocentric and cognitive strategies for 3D navigation in letter search tasks [29].

The large-size display has also been found to make possible different forms of visual presentation and interaction modalities. For example, a large display with gigapixels can accommodate more data to support more scalable visualizations [47]. Ball, North, and Bowman reported that when display reached the giga-pixel scale, interaction with walking improved task performance [2]. Similarly, in large 3D virtual environments, Ruddle, Payne, and Jones found that participants moved more and that such movement improved task performance [37]. Zanbaka et al. also found the benefits of FOV to improve memory [48]. Our study attempts to separate the interactivity and visualization to keep the interaction techniques constant and study the visualization factor alone. Users can only rotate the data. We also kept a constant FOV and a constant number of pixels so that the retinal projections were the same size among viewing conditions. This setting allowed us to ask a different set of questions and compare the effect of viewing small-close and large-far images.

While the aforementioned studies used an approach called display characteristic-specific study method [7], Swan et al. were among the first to compare the practical uses of displays and justified the value of empirical studies in practical uses [41]. Their study provided a comprehensive evaluation of four display types (desktop, CAVE, workbench, and Wall), stereopsis, movement type, and frame of reference in a map-based battlefield visualization environment. They found that desktop outperformed CAVE. Similarly, Demiralp et al. compared the effect of context on shape perception of a set of potato-shaped objects and found that CAVE were not beneficial in accomplishing tasks; fishtank had sharper images leading to better user experiences [11]. Prabhat et al.'s experiment compared a small fishtank and a CAVE for some volume rendering of biomedical datasets for counting tasks and reported that the CAVE was significantly more efficient than the other two displays [32]. They suggested that the benefits came from the embodied interaction, i.e., from the fact that participants could move their bodies freely and look at large "body-scale" images. In fact, Mizell, Jones, and Slater suggested that the "super-scaling" of visual features made possible by interactivity and zooming in large environments also improved task performance for some engineering tasks [26].

Existing results seem to suggest that high-quality images or large feature size are more important than display size when interaction is relatively the same (e.g., in [11]) and that when the users work in a relatively low-resolution environment (e.g., CAVE), enabling interaction can compensate for the low-resolution by displaying larger features to retain or improve task performance as in [32]. Following the suggestion of Swan et al. to design tasks within the application context

to establish usage guidelines [41], we have designed tasks that brain researchers would perform but have place these tasks in the context of display characteristics studies.

It is not surprising that performance depends on data and task characteristics. Qi et al. compared four volume-visualization problems related to identification and judgment of size, shape, density, and connectivity of objects in a volume using three displays: a head-mounted display (HMD) and a fishtank with and without haptic input [33]. This work suggested that, even though the HMD had larger fields of regard (FOR), fishtanks were better in providing overview and context, both useful in the volume visualization tasks and haptic-input-aided comprehension. Laha et al. were among the first to report that stereo worsens task performance in occluded structures [24]. Inspired by these previous work, here we study very dense datasets with low legibility. We also expand the area of study by running a dataset under a broad range of task conditions, including bundle tracing and some other tasks from the DMRI domain.

3 EXPERIMENTS

Our experiments used a $2 \times 2 \times 5$ within-subject design with the three independent variables of size (small and large), stereo (mono and stereo), and tasks. In general, we wanted to know if size and stereo have an impact on task performance (time and accuracy) on DMRI tasks using streamtube visualizations, especially when the retinal images are about the same. We had two general hypotheses: (1) we followed the bigger was better [6, 42] and hypothesized that adding stereo and increasing display size would improve task accuracy and reduce completion time for DMRI visualizations; (2) furthermore, the greater the complexity of spatial tasks, the greater the benefit of an improved display.

3.1 Display Settings

3.1.1 Equipment

We used two displays for the experiment with stereopsis on or off: a 24″ small display (Alienware's OptX AW2310) and a 72″ large display (Mitsubishi WD-73738 3D Ready TV) with their native 1920×1080 pixel resolution (Fig. 2). Displays were calibrated to be of roughly equivalent brightness and contrast. We set up the displays so that when either display was viewed from a specific distance, the visual angle and hence the size of the retinal image would be identical (Fig. 3). We used a 60° viewing angle, an optimal value suggested by Ware [45] that was achieved by taking into account the horizontal display size and the viewing distance. Coupled with the horizontal screen size of 531.3 mm, we set the participants a comfortable viewing distance of 1.5 ft (460.1mm) for the small display. In order to get identical retinal images, the large display was 4.5 ft (1380.3mm) away from the user. We asked participants to keep their heads as static as possible, although they were free to move closer or further away. In our software implementation, we also kept the software FOV at 60° to equalize the two FOVs, as in our previous study [28] following the suggested optimal solution by Czerwinski, Tan, and Robertson [10]. To minimize the environmental effects, all lights were turned off during the study; and there was minimal reflective light in the room.

During the experiment, participants could rotate the data by left-mouse dragging and the speed of the mouse movement was scaled to the screen size. No zoom was enabled so as to make it impossible to change the retinal images substantially. This setting gave a kinetic depth effect to permit the viewer to integrate spatial information over time and gain a continuous depiction of the spatial structure, similar to [46].

3.1.2 Stereo Implementation

Fig. 4 illustrates the stereoscopic setting. Both displays supported stereoscopic rendering which was turned on or off to control stereopsis. The participants saw the same visual angle of each pixel across displays. The stereo used frame sequential (or quad buffer) stereo in that each eye saw the full resolution of the entire screen, similar to [41]. Participants sat at the same viewing distance as in the mono condition with the software FOV set to 60° for each eye [8]. The
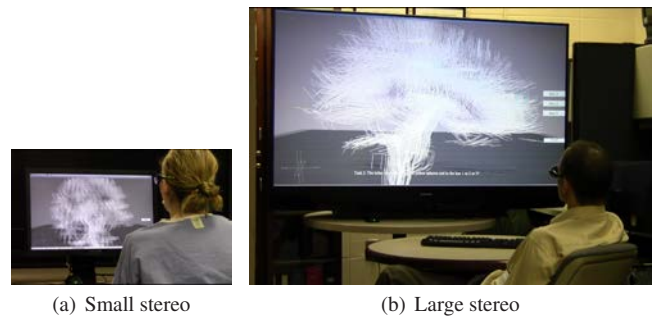

(a) Small stereo   (b) Large stereo

Fig. 2. Brain researchers view 3D diffusion MRI tractography on two displays (here showing the stereoscopic mode).

stereo was implemented using the viewing distance (or focal length), camera (eye) position and separation, software FOV, and viewing frustum. The eyes (camera) looked straight towards the screen along parallel vectors, separated by the viewing distance divided by 30 (30 was empirically chosen to avoid double images; any number between 20 and 30 worked well on our displays). The code was implemented in OpenGL running on a nVidia Quadro 4000 (by PNY) graphics card [8]. Fig. 5 shows two camera shots of images presented on the two displays in stereoscopic mode captured from the eye positions when all lights in the room were off, as in the experiments. The datasets were scaled to fill the screen and be displayed at zero disparity to reduce the discomfort with stereo viewing.

3.1.3 Keep Contrast Constant

Contrast determines the quality of the display in terms of brightness or luminance between white and black, and thus affects our ability to see details in visualizations. We displayed the ANSI 4×4 checkerboard pattern on each display, measured the brightness readings on a AMPROBE LM-100 light meter at the center of all white squares and all black squares, and calculated the contrast ratio using two methods. The first ratio was computed as (average (all white readings) - average (all black readings)) / (average (all white readings) + average (all black readings)), which is a good measure when gray-scale visualizations are presented. The ratios were 0.979 and 0.975 on the small and large displays accordingly. the ANSI ratio was also computed by dividing white to black readings, yielding 92.2 for the small and 78.1 for the large displays. We also took the approach that "human observer is always needed to carry out a color matching experiment" [42] and asked participants if they observed differences of displays that would affect their task performance, Our study participants reported no contrast problems while performing the tasks.

Participants wore the stereo glasses only in the stereo condition. To keep the retina image brightness the same, we could have had the participants to wear the stereoscopic glasses in 2D, similar to [41] and our previous study [15]. We did not do this because the present study was also designed to be faithful to real-world usage so as to ensure external validity. Wearing glasses did make the stereoscopic images dimmer than the 2D mono settings by about 50% measured using the same light meter. No complaints were received in several pilot studies about the low luminance caused by the glasses.

3.2 Tasks

Each participant completed 20 tasks on each display. They provided the answer by clicking on the result button on the screen. The first decision we have to make was to place those visual cues for tasks. Neurologists suggested to cover questions related to five representative fiber bundles in brain anatomy: corpus callosum (CC, interhemispheric fibers), cingulum bundles CG, (ventral-dorsal oriented), corticospinal tract (CST, cranial-caudal oriented), ILF (interior longitudinal occipitotemporal fasciculus, anterior to posterior), and IFO (inferior frontaloccipital fasciculus, anterior to posterior to lateral). The five bundles were chosen also because they represented two distinct
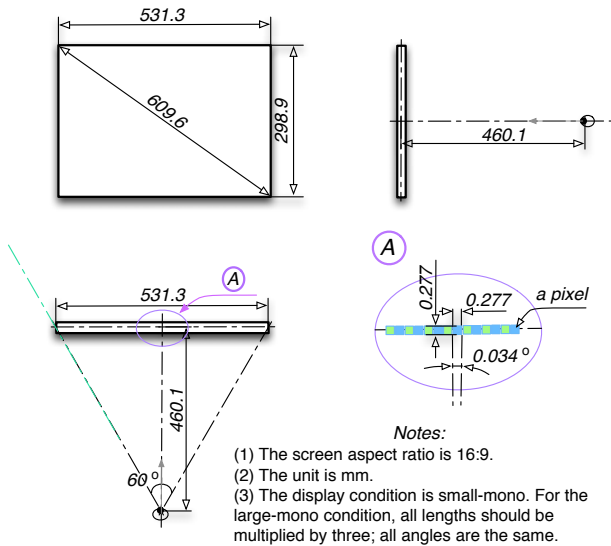
Notes:
(1) The screen aspect ratio is 16:9.
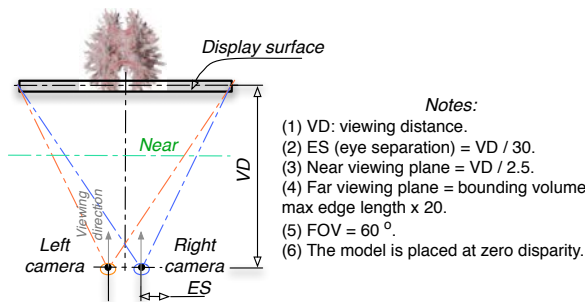(2) The unit is mm.
(3) The display condition is small-mono. For the large-mono condition, all lengths should be multiplied by three; all angles are the same.

Fig. 3. Display setting (mono).

Notes:
(1) VD: viewing distance.
(2) ES (eye separation) = VD / 30.
(3) Near viewing plane = VD / 2.5.
(4) Far viewing plane = bounding volume max edge length x 20.
(5) FOV = 60°.
(6) The model is placed at zero disparity.

Fig. 4. Display setting (stereo).

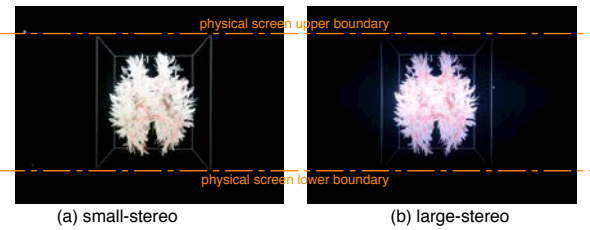(a) small-stereo   (b) large-stereo

Fig. 5. Photos of the two stereoscopic displays: the dataset was scaled to fill the screen during the experiment.

Table 1. Tasks (SIM: simple; CPX: complex).

Complexity	Acronym	Type
SIM	FA	Which of boxes 1 and 2 has higher average FA value?
CPX	TRACING	Which of the boxes covers the ending points of the tubes originating from the yellow spheres?
SIM	NAMING	What is the name of the fiber bundle?
CPX	LESION	Is there a lesion in the given region?
SIM	SAME	Do the yellow fibers belong to the
	BUNDLE	same bundle?

categories of fibers: association (i.e., intra-hemispheric) and commissural (i.e., inter-hemispheric) fibers. Also, the size and numerosity (number count) varied among these fibers, providing a range of suggestions on how visual design could serve knowledge discovery in these distinct structures.

3.2.1 Task Selection Approach

Our approach to selecting tasks was to review the literature and discuss with four brain researchers their activities and expectations for an interactive visualization design. The results indicated that 3D visualizations of fiber tractography could be used to answer questions in the following categories with some subtasks in each category: (1) certain numerical measurement metrics [9], (2) spatial relationships of fiber bundles (e.g., for resolving brain connectivity), (3) pathological condition search and manipulation (e.g., is there a lesion? How can fibers be cut to remove a tumor?), (4) comparisons (e.g., how different are two bundle volumes in two hemispheres? Are two brain tractographies the same? Are they normal?), (5) categorical (e.g., to which anatomical structure does a bundle belong?), (6) tube tracing (e.g., where does a tube bundle go?) and (7) understanding of functional or structural images.

Our experiments included five tasks chosen from these seven categories (Fig. 1) and our selection criteria balanced task difficulty, usefulness, and experiment length as tested in several pilot studies. The first criterion was related to data. We excluded those conditions for which we either did not have a ground truth or could not simulate graphically due to unknown pathological characteristics. One such task was tumor detection. Understanding tumors, in particular knowing if a tumor infiltrated, displaced edematous tracts, or destroyed

tubes, was crucial to surgical planning [20], yet generating tumor effects was unlikely to be grounded because we did not know how the tubes would change. In a previous study, we placed tumors in an area and asked people to judge if tumor and tracts were in contact, and we found that the task became arbitrarily more or less complex depending on tumor location [31]. Thus we excluded such tasks from this study. We also excluded tasks in category two for which we did not have enough data. Finally, we excluded CG fibers in the summative evaluation because most CG fibers from our tractography were too short to be clinically meaningful.

The second data selection criterion was related to tasks. We excluded those open-ended qualitative tasks, such as in category 7, which would deserve a full-blown study, perhaps with the insight-based evaluation method or well-controlled experiment [24]. We did use this task in a pilot study as a complex task, but decided to remove it because pilot participants (neurologists) commented that they would need more than 10 minutes to describe the data features meaningfully. This may suggest that our datasets were far more complex than in Laha et al.'s study [24], where their participants were given about a minute to work on descriptive tasks. Task difficulty was also a consideration. The choice of the tasks related to these seven categories could differ widely and affect task difficulty profoundly. For example, gauging if two bundles are the same or not (a binary-choice task) is much easier than considering the percentage difference between two bundles (a numerosity task). We originally used the numerosity task (task 5) because it would be more interesting, but changed it to the binary-choice one, because pilot study results suggested that participants guessed the answers.

3.2.2 Example Tasks

Our final task set includes five tasks based on the above selection criteria. All tasks only have either right or wrong answers and there is only one correct answer for each question. Participants must make a choice before moving on to the next. Our pilot test results show that task difficulty is within a reasonable range and participants can provide an answer in our settings. The accuracy of each task is defined by the percentage of correct answers. Since we were not interested in differences among tasks, tasks were executed sequentially by alternating the complexity in the order of the task index below. Task completion time, accuracy, and subjective workload were recorded.

Fig. 1a (FA) shows a sample stimulus for comparing the average fractional anisotropy (FA) values in two boxes. The participants were instructed to choose the box covering voxels of higher average FA values. The participant was also told that FA was a quantity used in DMRI to measure the anisotropy in each voxel of the brain volumes. The FA color map was shown on the right: the redder the color, the higher the FA value, following the color mapping in Zhang et al. [49]. Participants indicated the higher FA value by pressing the 'Similar' or '1 is higher' or '2 is higher' button from the three choices. The FA value similarity threshold was set at 0.05, i.e., the two box values were considered similar if the difference was less than 0.05. We chose this value because the small threshold was not clinically interesting given the measurement uncertainty. The chances of selecting each of the three answers were about the same.

Fig. 1b (TRACING) shows a sample interface for the fiber-tracing task in the DMRI streamtube visualizations; the yellow spheres mark the starting points and the three boxes show possible ending positions. Participants were asked to find the box in which the ending points lay. They were told that the marked fibers belonging to the same bundle followed the same orientation. Participants were also told that the three boxes were placed at the end of bundles, each belonging to one of the three anatomical orientations (anterior-posterior, dorsal-ventral, and left-right), and that no two enclosed the bundles on the same orientation. For example, the box 1 in Fig. 1b covers cranial-caudal fibers, the box 2 (the correct answer) encloses anterior to posterior fibers, and the box 3 is at the end of the only inter-hemispheric fibers.

Fig. 1c (NAMING) shows the dataset used in the bundle-naming task: participants were asked to name the fiber bundle marked in yellow. The participants, regardless of their background, were trained to recognize the fiber bundles in order to ensure sufficient knowledge about the task and datasets. During the experiment, a cheat sheet was provided displaying the CC, CST, ILF, and IFO bundles, so they did not have to remember the names.

Fig. 1d (LESION) shows a task condition for the lesion task: participants were asked to locate the lesion and indicate it by right-clicking as close to the center of the lesion area as possible. Participants were also told that the lesion was located in one of the five bundles and they need to right-mouse click the center of the lesion (where the red cross is in Fig. 1d). All points within the lesion area in the screen coordinates were considered to be the correct answers.

Fig. 1e (SAME BUNDLE) shows an example of task 5, in which participants were asked whether or not the fibers in yellow all belonged to the same bundle. 50% of the data were in the same bundle and 50% were not. The choice of distracting fibers was based on fiber orientations; often fibers from the closest perpendicular bundles were selected. The example in Fig. 1e shows that the highlighted fibers do belong to the same bundle, with a set from the CST (cranial-caudal oriented). In the not-the-same-bundle condition, we would add noisy fibers from the CC (inter-hemispheric) bundles and ask the participants to make a judgment.

3.3 Diffusion MRI Datasets

3.3.1 Dimensions of legibility in DMRI visualizations

The ultimate goal of the experimental study is to guide the design of effective visualizations when choosing displays, and thus the data characteristics of the application must be considered [27]. Clutter is a notorious problem in DMRI tractography when a uniform sampling at each voxel of DMRI is used, as witnessed by the continuous efforts to improve spatial structure understanding [14]. Clutter remains a problem in all further stages in the data-analysis pipeline, since it can limit structural recognition and visual segmentation [36]. In graphics, clutter can be addressed by managing occlusion [13] and enhancing legibility so that individual graphics items are unambiguous and can be read by the users (p. 175 [5]) even when displayed in 3D or small pixels [21].

Bertin's three dimensions of legibility for map drawings (Table 2), density, angular legibility, and retinal legibility, can be usefully applied to 3D visualizations. We added the dimension of context [3], i.e., the direct relevance of the surrounding data to the tasks at hand.

We believe that context is orthogonal to the other dimensions applicable to DMRI visualizations: it can provide spatial references and thus aid the tasks at hand, but it can also obscure the internal structures, making diagnosis more difficult due to occlusion. The most obvious examples of this occur when small U-shaped fibers are shown in such a way that we cannot see inside regardless of visualization method, and when the seeding resolution becomes so high that removing irrelevant fibers would require significant user interaction.

Table 2. Dimensions of legibility.

Dimensions	definition
density	number of marks per area
angular legibility	- a global picture
	occupying the right scale on the primary axes
	- shape of a readable variable
retinal legibility	separate foreground and background by
	using the right amount of ink
context	visual objects in surroundings often
	embedded with other related objects

3.3.2 Density

The first factor that could affect our results is the visual clutter caused by dense streamtube visualizations. Given different displays, we would expect adding stereo and increasing size to help legibility in general. In this study, we did not vary density in the datasets.

3.3.3 Context

We collaboratively identified a three-step workflow to choose context presentation, expanded from the information visualization mantra: overview first, zoom and filter, then detail-on-demand [40], and from the report of neurologists' data analysis workflow in [12]. In the first step, initial global examination of the data, the full brain is examined, since this can facilitate the study of related measurement metrics that can provide more robust markers of white-matter structural integrity [9]. The second step is a deeper analysis in which some fiber tracts are removed from the whole brain in such a way that some contexts are preserved. This partial brain study is an intermediate data-exploration stage in which users remove some irrelevant blocking fibers and focus on the study at hand before approaching the region of interest (ROI). The final step is to investigate a certain ROI that is associated with task-relevant bundles, so as to make a more precise pathological assessment without visual occlusion after the surrounding fibers are sufficiently understood. Using this workflow, here we used fibers from the half-hemispherical partial volume to reduce the complexity of seeing the full volume while preserving some context information.

3.3.4 Tube Rendering

Tractography data were computed from the source MRI images captured from a normal human brain at resolution ($0.9375 \times 0.9375 \times 4.52$ mm^3). Diffusion tensors of each seeding resolution are calculated with tricubic B-spline interpolation and then fiber tracts are approximated using the second-order Runge-Kutta solver [4, 49]. During the tractography, the full volume seeding algorithm [43] is adopted for seed selection to produce a tractography sequence.

3.4 Experimental Design

3.4.1 Design

Table 3 shows our experimental design, which followed random ordering with a Latin square design. Each participant (column 1) performed 20 tasks (4 datasets × 5 task types) on each display (row); the order of the displays followed a Latin-square design. We prepared four data groups with each of the CC, CST, ILF, and IFO (column 3) to counterbalance the task difficulties on four displays (column 2). The task difficulties depend on these four datasets and the only changes were

Table 3. Experimental Design.

participant	display	data			
p1-3	SM	CC1	CST1	ILF1	IFO1
	LM	CST2	ILF2	IFO2	CC2
	SS	IFO3	CC3	CST3	ILF3
	LS	ILF4	IFO4	CC4	CST4
p4-6	LM	CC1	CST1	ILF1	IFO1
	SM	CST2	ILF2	IFO2	CC2
	LS	IFO3	CC3	CST3	ILF3
	SS	ILF4	IFO4	CC4	CST4
p7-9	SS	CC1	CST1	ILF1	IFO1
	LS	CST2	ILF2	IFO2	CC2
	SM	IFO3	CC3	CST3	ILF3
	LM	ILF4	IFO4	CC4	CST4
p10-12	LS	CC1	CST1	ILF1	IFO1
	SS	CST2	ILF2	IFO2	CC2
	LM	IFO3	CC3	CST3	ILF3
	SM	ILF4	IFO4	CC4	CST4

the target locations on these bundles. The clinical validity of these data conditions were also confirmed by neurologists.

3.4.2 Participants

Twelve medical residents on rotation to the neurology department volunteered for the study. Half of them were female and their average age was 32.5 years. One reason for choosing this expert group was that we observed in a pilot study the significant main impact of expertise on timing. Though the accuracies were about the same between expert and novice groups and were consistent in all task conditions, we recruited only neuroscientists in the formal study to avoid the confounding factor of participant expertise.

In that pilot study, we compared performance with 10 participants: five novices (avid computer users who had no knowledge in medicine) and five experts (neuroscientists). One reason to run this pilot study was that recruiting busy medical experts had been difficult and time-consuming. If they got similar performances, we might in the future use members of the general public to run this type of study. The other reason was due to participants' expertise. Though past work used novices to replace experts in 3D flow studies [15], we thought doctors might outperform the general population, since they were used to examining images. In addition, we hoped to learn the population usage patterns in the study so that results obtained would be suitable for the end users. The novice group included participants who were undergraduate computer science majors. The medical expert group had faculty and residents in the neurology department of the University of Mississippi Medical Center. That pilot study aimed to discover performance differences between novices and experts. Because we found a significant main effect in that pilot study, we chose to use medical experts only in the formal experiment presented here.

3.4.3 Procedure

Participants were asked to confirm that they had the normal and normal color vision. Participant stereoscopy was tested using our own pictorial presentation; all confirmed during training that they could see stereo on our displays. Participants were told about brain structure and were given a brief description of DMRI techniques and their clinical use. They also had a short (about 15 minutes) warm-up session that presented two trials under each of the tasks, using different data from the formal study for each condition: these training sessions ensured that the participants understood the conditions and the brain anatomical structures. An example training document, with video and audio, is available online at [8]. Participants completed 80 tasks total or 20 tasks on each display. The task completion time was recorded from the time when the model was loaded to when the answer button was clicked. Participants were allowed to change their answers

and that time was included as well. The time between clicking an answer button and clicking the "next" button to advance the task was not recorded. Participants were suggested to take breaks between these two clicks when needed. They filled out a post-questionnaire to rate their experiences in the four display conditions and their overall experience. Participant generally spent 1 to 1.5 hours to finish the experiment and were compensated for their participation.

4 RESULTS AND ANALYSES

We collected 960 data points with 12 participants while performing five tasks using two displays with on and off stereopsis. Before a statistical analysis was conducted, we used a quantile-quantile plot (QQ plot), a graphical method, to test the normality. We removed outliers if they lay more than three standard deviation of the mean in each experimental condition. Overall, we removed about 10 outliers from the 960 samples. All error bars in all graphs in this result section represent one standard error from the mean.

We first looked the overall main effects by performing two types of factorial analyses: three-way (size, stereo, and task) and two-way (display and task). We called the combined size and stereo conditions "display", because it allowed us to examine individual devices. We then performed the Tukey post-hoc analysis on the display factor when there was significant main effect. We summarize overall performance measurement results and test statistics in Fig. 6 and show performance by tasks in Fig. 7. We omit the F and p values in the text part if the values are in the figures.

4.1 Performance by Task

4.2 Performance Summary

We were surprised to see that our first hypothesis that size and stereo would improve task performance was not supported. A three-way ANOVA, with stereotype, size, and task being within-subject factors, was used to analyze task completion time and accuracy. There was a significant main effect by **tasks** (($F(4, 19) = 48.1$, $p < 0.0001$), **size** ($F(1, 19) = 5.28$, $p = 0.02$), and **stereopsis** ($F(1, 19) = 11.1$, $p = 0.001$) settings for task completion time. Only tasks and stereopsis settings were significant for the accuracy (task: $F(4, 19) = 8.36$, $p < 0.0001$; stereopsis: $F(1, 19) = 21.1$, $p < 0.0001$). Size had no effect on accuracy. There was a significant two-way interaction between stereopsis and task ($F(4, 19) = 3.4$, $p = 0.01$).

A two-way ANOVA with display (stereotype and size combined) and task showed that a significant main effect of display on task performance (time: $F(3, 19) = 5.5$, $p < 0.001$; accuracy: $F(3, 19) = 7.17$, $p < 0.0001$). SM (15.4 s) had the best task completion time, followed by LM (18.0 s), SS (19.2 s), and LS (21.2 s) (Fig. 6(c)). LM (0.87) and SM (0.88) also led to the most accurate answers, followed by LS (0.75) and SS (0.77) (Fig. 6(f)). A post-hoc Tukey test revealed two groups: $< SM, LM >$ and $< SS, LS >$ and no significant main effect was found for the displays within the same group. Because task completion time is very influential, the following sections report the results by task only rather than the overall results.

4.2.1 Task Completion Time and Accuracy

We first examined for which tasks we could observe the significant main effect. Fig. 7 plots the performance by tasks. The small display outperformed the large display for the FA and SAME BUNDLE tasks and the main effect of size on task completion time was significant (Fig. 7a 1-2). The mono also outperformed the stereo condition in task completion time for the FA, LESION, and SAME BUNDLE tasks and the main effect of stereopsis on task completion time was significant (Fig. 7b 1-3).

Because both main effects of size and stereo were significant, we collapsed the data by stereopsis for the FA and SAME BUNDLE tasks to learn under which condition, the main effect occurred. We found that the main effect of size on task completion time was only significant in the mono condition but not in stereo (FA: mono: F(1, 1)=4.7, p=0.03, <mean: S: 15.4s vs. L: 19.3s>; stereo: F(1,1)=0.7, p=0.40, <mean: S: 19.7s vs. L: 21.8s>; SAME BUNDLE: mono: F(1,1)=6.3, p=0.01, <mean: S:8.6s vs. L: 11.8s>; stereo: F(1, 1)=2.4, p=0.13,
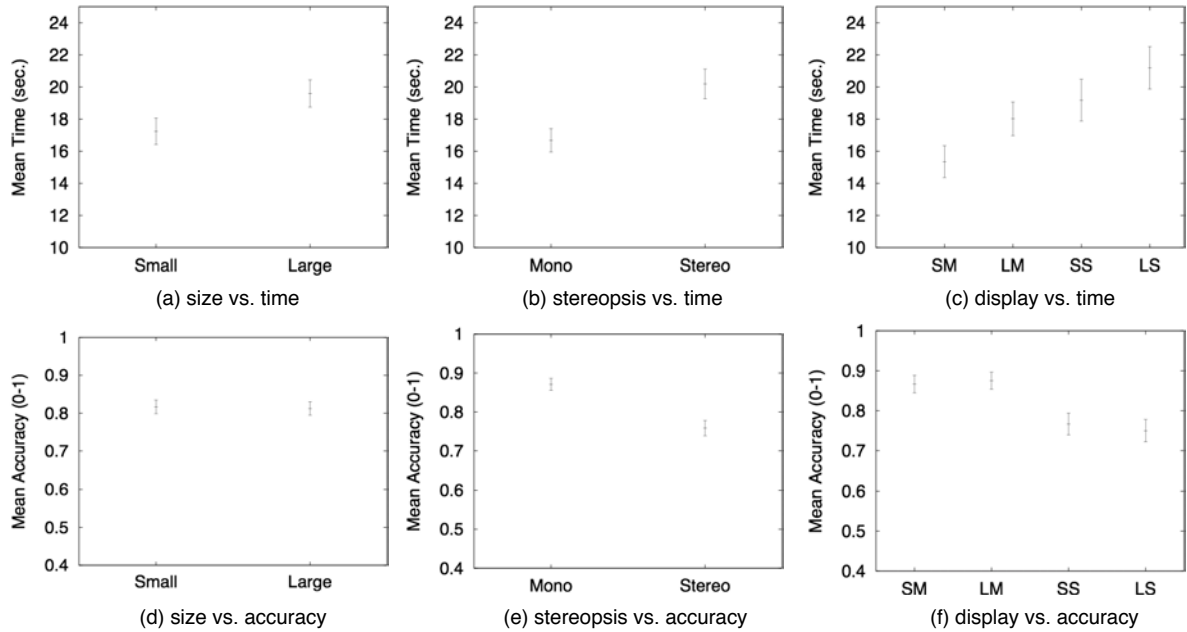
Fig. 6. Task completion time and accuracy by display characteristics across all tasks. All error bars in the graphs in this result section represent one standard error from the mean.

$<$S: 10.8s vs. L: 13.1$>$). For both tasks, the small display had better task completion time than the large display when the display was mono. But under the stereoscopic condition, size did not have a significant impact on task performance.

The increase in accuracy would be more important than efficacy in clinical setting and it was calculated from the percentage of correct answers. The main effect of size on accuracy was not significant (Fig. 7d). The mono conditions led to more accurate answers for LESION and SAME BUNDLE (Fig. 7e 1-2). There was a trend that the mono conditions led to more accurate answers for TRACING ($F(1,2) = 2.6$, $p = 0.1$), though the effect was not significant.

4.2.2 Combined Display Effects

When evaluating the displays by task, we observed that the displays had significant impact on efficiency for the FA and SAME BUNDLE tasks; and the only differences of mean was between SM and LS. The impact of display on efficiency was also significant for the LESION task and two groups were found: $< SM, LM >$ and $< SS, LS >$ (Fig. 7f 1-2).

For FA, though we did not observe significant differences of display on accuracy, LM led to the most accurate answers (about at least 6% higher than all other conditions) (Fig. 7f first column). Considering these tasks were used for medical diagnosis, this accuracy differences were high enough to be considered important. Similar observations could be made for the TRACING tasks where LM also led to the highest accuracy (Fig. 7f second column). We did not find any significant main effect of size and stereo on TRACING and NAMING (Fig. 7). Participants generally felt NAMING was an easy task.

4.3 Subjective Ratings and Comments

We collected subjective self-evaluation data using a post-questionnaire asking about the **effectiveness** (EFFE, this system's capabilities to meet requirements), **overall satisfaction** (SAT, using this system is not a frustrating experience), **ease of use** (EoU), and **efficiency** (EFFI, don't have to spend too much time correcting things with this system) for the four display conditions (Fig. 8) on a scale of 7 with 1 being the worst and 7 the best. In general, participants were positive about all display conditions and were comfortable using them all, as all mean scores are larger than 4.5.
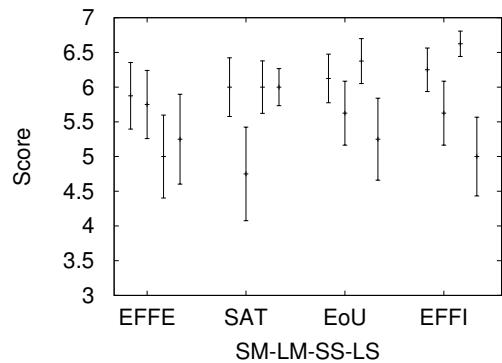

Fig. 8. Subjective rating scores comparing these four displays ordered in Large-Stereo (LS), Small-Stereo (LS), Large-Mono (LM), and Small-Mono (SM) with regard to effectiveness (EFFE), satisfaction (SAT), ease of use (EoU), and efficiency (EFFI).

Participants thought the small displays were easier to use, more efficient, and more effective than the large display. They suggested that they did not need the large display to understand the brain datasets. About half the participants remarked that they would use the stereoscopic setting with the small-stereo display because they felt the stereo allowed better depth so they could see better. They especially commented that they liked small-stereo for the TRACING tasks. On the one hand, they felt they could touch the tubes on the small-stereo with clearer depiction. Tubes on the display were too far to reach especially in the stereoscopic case. This might explain why participants rated SS highest in almost all categories except for effectiveness (Fig. 8 first column). Almost all participants preferred the small display and five of the 12 participants commented that they were confident to infer 3D shapes from mono images because the mono images were used in their textbooks.

Almost all participants felt that the large display seemed overwhelming, especially when stereo was activated. They thought the stereo was "cool" but offered them no benefits for the DMRI tasks, at least in its current form. Participants generally reported that they
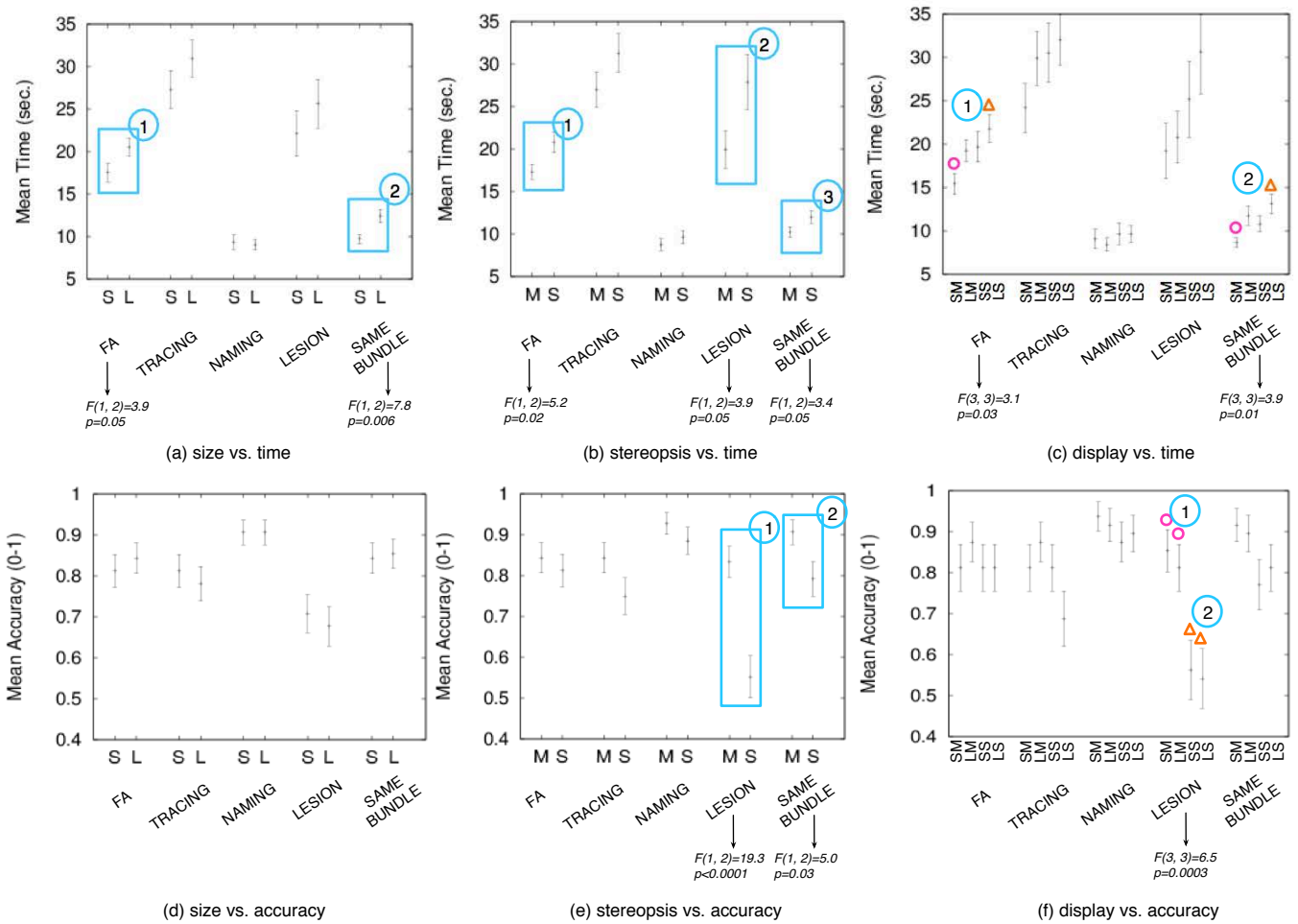
Fig. 7. Task completion time and accuracy by tasks. Different color symbols in c and f represent different Tukey groups from post-hoc analyses.

looked much harder at the stereoscopic conditions simply because the structures became clearer for some tasks and thus more interesting. However, doing this did not improve accuracy. This might explain why participants spent much longer looking in the LS condition (Fig. 7c) but had the worst accuracy (Fig. 7f). They felt the stereoscopic display for LESION was the worst possible combination.

Three participants saw possible value for tasks different from those in our experiment. They would like to have multiple-views on the large displays, so they could place multiple data side-by-side for comparative analyses, a task frequently performed by neurologists. Four participants who preferred small displays reported that the screen size of the small monitor matched their expected size of the human brain and they did not need the large display to understand the data.

Perhaps most remarkably to visualization design, participants also suggested using more non-photorealistic rendering or illustration-based techniques similar to those used in their textbooks to show the bundles. The current tube rendering did not seem to help very much to 'assist' them to 'see' more. Some of them remarked that they liked using hand-held devices for them to 'carry' the data, hold them close-by their eyes, and experiment with how best to see the complex structures.

5 DISCUSSION

5.1 Stereo Experience

Doubt is raised by the observation that stereo viewing leads to worse task performance, against our hypothesis and general findings that two eyes are better than one. Most participants commented that the stereoscopic viewing of dense tubes triggered pictorial depth, but to our surprise few perceived any benefit in using stereoscopic display. In

LESION and SAME BUNDLE, participants were more accurate with monocular than with binocular presentations. Note that both tasks required the participants to derive patterns from tubes.

We think this negative effect is perhaps more related to some intrinsic drawback with stereoscopic displays for pattern analysis tasks that require seeing internal structures. This lesion structures are somewhat similar to the internal structures in the Laha et al.'s study where that study also presents detrimental stereoscopic performance [24]. Their paper reports that such negative stereo experience could have been caused by participants' eye strains from the stereoscopic viewing. The other explanation for the worse task performance on the large stereo display is related to cue conflicts. One important distance cue is that our eyes converge more for near than for far objects [17]. This convergence effect also explains why we always look smaller in mirrors. or images on the Wheatstone stereoscopic display look smaller even though the retinal image is not changed.

The visual system "thinks" it is looking at a closer object and scales the visual perception in the direction of the object's physical size. This effect might explain why participants felt the lesion areas 'shrank' with the distance in stereo. To some extent, the lesion area "looked smaller" in the stereo mode as participants commented, thus making the visual search difficult. On the other hand, we are likely to get more correct stereo-readings on the near-small display than the far-large display. In our setting, the disparity increases with the viewing distance and does not exactly match physical eye separations. We did this purposefully to equalize the retinal images. Then the stereo disparity for the near view would be more close to the physical eye separation compared to the far view, which would have slightly exaggerated disparity, though

the effect might be small [44].

Additionally, the stereoscopic images were darker than the mono displays or that the stereoscopic display had low contrast. An interesting future direction would be to study how performance varies with differences in screen contrast and brightness. One way to do this is to compare 3D TV with some high-end display with compatible brightness and contrast to examine if performance differences will be found.

Our results also indicate that our knowledge about stereoscopic viewing for dense tube-based visualizations is limited. We need to understand shape as a collection of mutually dependent lines for shape representations. One way is to alter the data rendering method to increase internal structure legibility. One solution might be to allow effective surface structure detection by adding a "cap" at the end of a tube to allow the perception of surfaces and provide a clear view of those broken fibers, for example, depicting the tensor field with glyphs or volume rendering methods [22]. We may alter the data-rendering method to increase visual legibility, especially since the rendering method will alter the perceived structure, for example by applying flow maps [16] or showing topological structures [30, 39]. One possibility is to study the tradeoffs between display characteristics and more advanced depth-enhancement techniques, e.g., adding halos and providing better color design [19].

5.2 Size Experience

In general, contrary to our hypothesis, the large screen did not improve task performance. This result could have at least two explanations. First, it might be argued that the brain structures are familiar to medical school clinicians and that they are used to examining 2D structures in textbooks as well as 3D structures. This may explain why participants dislike the large displays or in another words, they would suggest a different use in multiple-views. The brain tractography visualization on the small monitor was presented at its expected actual size and that there might not be a benefit from presenting it larger-than-life. Also, participants were experts who had good mental model about the brain structure, that they did not need the large display to provide situation awareness, as in other settings [28]. Second, we might posit that the size experience is enhanced in 'super-scale' as large as our body size [26], but our large display did not support zooming to bring large model to participants' eyes.

5.3 Retinal Images Experience

We observed that the small display improved task completion time only when the display was in mono mode. Under the stereo display condition, the two sizes performed equally well. Since the far-large and near-close should project the same retinal images when the stereotype was the same, we might predict that the retinal image was a stable estimate of task performance only under the stereoscopic condition. With the mono display, retinal image was generally a good estimate of accuracy but not of task completion time, at least for the tasks used here.

There did appear to be a penalty associated with large-stereo viewing. A number of participants stated that they found viewing the same DMRI tubes in the LS mode somewhat stressful. Part of this stress may be due to the difficulty of the tasks: trying to find the structures from several tens of tubes in a tangle of thousands of tubes is not easy. They also felt some loss of control of the data when they sat further away. For example, while performing the TRACING tasks, participants used their fingers to point to the tubes to trace them; but they could not do this on the far-large settings. We also suspected that coupling a body-centric view so that participants could move their bodies around or to zoom would improve viewing accuracy in the stereoscopic mode.

5.4 Tasks

Not surprisingly, we found a main effect of task type on completion time, since the tasks differed in difficulty and cognitive activity involved. The stereo and size had no significant effects on accuracy for NAMING tasks, suggesting that NAMING might not require stereo viewing or large displays. Indeed, all participants described that this was the simplest task. Originally, we thought FA and SAME BUNDLE were very different tasks because FA was similar to visually mapping the colors to numerical values while SAME BUNDLE involved counting and searching for differences, and because in the FA tasks the target fibers constituted only a small chunk of the overall structure, while participants needed to find only one difference in order to answer the SAME BUNDLE questions. The two tasks apparently shared some similarities with regard to the display choices. Participants also reported that they mostly focused their attention on the fibers within the two boxes.

5.5 Implications for Design

In this study, we used DMRI data carefully selected with neurologists in a controlled experiment to learn the impact of stereo and size on a set of DMRI tasks. The results can be generalized to cases where users examine large dense tube visualizations. We have the following recommendations for designers in choosing size and stereo: (1) Stereo seems to have a greater impact on performance than size at least for the tasks and visualizations in this study. We will need to design better stereoscopic display experience that makes it useful for visual feature extraction from dense dataset; (2) When a stereo display is chosen, the task execution time did not vary with display size as long as the retinal image is unchanged. Thus the retinal image is a fairly good estimate of task execution time for the task and datasets we have studied; and (3) For tasks that require seeing a fiber track (e.g, FA and TRACING tasks), large-mono display may be a better choice. For complex tasks that require shape understanding from thin tubes along the depth dimension (e.g., LESION), further considerations are needed to balance the stereo and mono displays.

6 CONCLUSIONS

Overall, we think we still know very little about complex dense dataset visualizations in 3D environments. One future direction is perhaps to revisit those complex 3D shape understandings for better visualization experience [1, 18, 34]. Previous work has shown that large displays and stereoscopic displays can increase task performance in many types of environments. This study focuses on the impact of displays on streamtube visualizations in real-world DMRI tasks measured with medical doctors. The major contribution of this article is to provide the first quantitative estimate of the benefits of stereo and size for perceiving dense structured tube data based on the taxonomy of legibility. We provide a taxonomy of legibility and a rare counterexample in which small mono displays were sufficient. Our results surprised us: the easy-to-understand large stereo did not yield the best performance as we expected. All the possibilities are to be further investigated, but they do not detract from the practical utility of our findings.

REFERENCES

[1] A. Bair, D. House, and C. Ware. Texturing of layered surfaces for optimal viewing. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1125–1132, 2006.

[2] R. Ball, C. North, and D. Bowman. Move to improve: promoting physical navigation to increase user performance with large displays. In *Proc. of the SIGCHI conference on Human factors in computing systems*, pages 191–200, 2007.

[3] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[4] P. J. Basser, S. Pajevic, C. Pierpaoli, J. Duda, and A. Aldroubi. In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, 44:625–632, 2000.

[5] J. Bertin. *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin Press, 1983.

[6] E. Boring. *Sensation and perception in the history of experimental psychology*. Appleton-Century-Crofts Inc., 1942.

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

[8] J. Chen, A. Auchus, and D. Laidlaw. DMRI display study, https://sites.google.com/site/simplevisualizationlanguage/.

[9] S. Correia, S. Lee, T. Voorn, D. Tate, R. Paul, S. Zhang, S. Salloway, P. Malloy, and D. Laidlaw. Quantitative tractography metrics of white matter integrity in diffusion-tensor MRI. *Neuroimage*, 42(2):568–581, 2008.

[10] M. Czerwinski, D. Tan, and G. Robertson. Women take a wider view. In *Proc. of the SIGCHI conference on Human factors in computing systems*, pages 195–202, 2002.

[11] C. Demiralp, C. Jackson, D. Karelitz, S. Zhang, and D. Laidlaw. Cave and fishtank virtual-reality displays: A qualitative and quantitative comparison. *IEEE Transactions on Visualization and Computer Graphics*, 12(3):323–330, 2006.

[12] S. Diepenbrock, J. Prassni, F. Lindemann, H. Bothe, and T. Ropinski. 2010 IEEE visualization contest winner: Interactive planning for brain tumor resections. *IEEE Computer Graphics and Applications*, 31:6–13, 2011.

[13] N. Elmqvist and P. Tsigas. A taxonomy of 3D occlusion management for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(5):1095–1109, 2008.

[14] M. Everts, H. Bekker, J. Roerdink, and T. Isenberg. Depth-dependent halos: Illustrative rendering of dense line data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1299–1306, 2009.

[15] A. Forsberg, J. Chen, and D. Laidlaw. Comparing 3D vector field visualization methods: A user study. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1219–1226, 2009.

[16] M. Hlawatsch, J. Vollrath, F. Sadlo, and D. Weiskopf. Coherent structures of characteristic curves in symmetric second order tensor fields. *IEEE Transactions on Visualization and Computer Graphics*, 17(6):781–794, 2011.

[17] J. Hollerbach, W. Thompson, and P. Shirley. The convergence of robotics, vision, and computer graphics for user interaction. *The International Journal of Robotics Research*, 18(11):1088–1100, 1999.

[18] V. Interrante. Perceiving and representing shape and depth. *PhD thesis, UNC-Chapel Hill, Department of Computer Science*, 1996.

[19] V. Interrante and C. Grosch. Visualizing 3D flow. *IEEE Computer Graphics and Applications*, 18(4):49–53, 1998.

[20] B. Jellison, A. Field, J. Medow, M. Lazar, M. Salamat, and A. Alexander. Diffusion tensor imaging of cerebral white matter: a pictorial review of physics, fiber tract anatomy, and tumor imaging patterns. *American Journal of Neuroradiology*, 25(3):356, 2004.

[21] D. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

[22] G. Kindlmann, R. Whitaker, T. Tasdizen, and T. Moller. Curvature-based transfer functions for direct volume rendering: Methods and applications. In *IEEE Visualization*, pages 513–520, 2003.

[23] R. Kosara, C. Healey, V. Interrante, D. Laidlaw, and C. Ware. User studies: why, how, and when? *IEEE Computer Graphics and Applications*, 23(4):20–25, 2003.

[24] B. Laha, K. Sensharma, J. Schiffbauer, and D. Bowman. Effects of immersion on visual analysis of volume data. *IEEE Transactions on Visualization and Computer Graphics*, 18:597–606, 2012.

[25] J. LaViola, A. Forsberg, D. Laidlaw, and A. van Dam. Virtual reality-based interactive scientific visualization environments. *Trends in Interactive Visualization*, pages 225–250, 2009.

[26] D. Mizell, S. Jones, M. Slater, and B. Spanlang. Comparing immersive virtual reality with other display modes for visualizing complex 3D geometry. *University College London, technical report*, 2002.

[27] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[28] T. Ni, D. Bowman, and J. Chen. Increased display size and resolution improve task performance in information-rich virtual environments. In *Proc. of Graphics Interface*, pages 139–146, 2006.

[29] R. Pausch, D. Proffitt, and G. Williams. Quantifying immersion in virtual reality. In *Proc. of the 24th annual conference on computer graphics and interactive techniques*, pages 13–18, 1997.

[30] R. Peikert, H. Hauser, H. Carr, and R. Fuchs. *Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications*. Springer Verlag, 2012.

[31] D. Penney, J. Chen, and D. Laidlaw. Effects of illumination, texture, and motion on task performance in streamtube visualization of diffusion tensor MRI. pages 97–104, 2012.

[32] Prabhat, 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*, 14:551–563, 2008.

[33] W. Qi, R. Taylor, C. Healey, and J. Martens. A comparison of immersive HMD, fish tank VR and fish tank with haptics displays for volume visualization. In *Proc. of the 3rd Symposium on Applied Perception in Graphics and Visualization*, pages 51–58, 2006.

[34] P. Rheingans and D. Ebert. Volume illustration: Nonphotorealistic rendering of volume models. *IEEE Transactions on Visualization and Computer Graphics*, 7(3):253–264, 2001.

[35] J. Roeckelein. *Dictionary of theories, laws, and concepts in psychology*. Greenwood Pub Group, 1998.

[36] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *Journal of Vision*, 7(2):1–22, 2007.

[37] R. Ruddle, S. Payne, and D. Jones. Navigating large-scale virtual environments: what differences occur between helmet-mounted and desk-top displays? *Presence: Teleoperators & Virtual Environments*, 8(2):157–168, 1999.

[38] T. Schultz. Feature extraction for DW-MRI visualization: The state of the art and beyond. In *Proc. of Schloss Dagstuhl Scientific Visualization Workshop*, 2010.

[39] T. Schultz. Topological features in 2D symmetric higher-order tensor fields. 30(3):841–850, 2011.

[40] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. of IEEE Symposium on Visual Language*, pages 336–343, 1996.

[41] J. Swan, J. Gabbard, D. Hix, R. Schulman, K. Kim, et al. A comparative study of user performance in a map-based virtual environment. In *Proc. of IEEE Virtual Reality*, pages 259–266, 2003.

[42] D. Tan, D. Gergle, P. Scupelli, and R. Pausch. Physically large displays improve performance on spatial tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(1):71–99, 2006.

[43] A. Vilanova, G. Berenschot, and C. van Pul. DTI visualization with streamsurfaces and evenly-spaced volume seeding. In *Proc. of the Eurographics Symposium on Visualization*, pages 173–182, 2004.

[44] C. Ware. Dynamic stereo displays. In *Proc. of the SIGCHI conference on Human factors in computing systems*, pages 310–316, 1995.

[45] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, second edition, 2004.

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

[47] B. Yost, Y. Haciahmetoglu, and C. North. Beyond visual acuity: the perceptual scalability of information visualizations for large displays. In *Proc. of the SIGCHI conference on Human factors in computing systems*, pages 101–110, 2007.

[48] C. Zanbaka, B. Lok, S. Babu, A. Ulinski, and L. Hodges. Comparison of path visualizations and cognitive measures relative to travel technique in a virtual environment. *IEEE Transactions on Visualization and Computer Graphics*, 11(6):694–705, 2005.

[49] S. Zhang, C. Demiralp, and D. Laidlaw. Visualizing diffusion tensor MR images using streamtubes and streamsurfaces. *IEEE Transactions on Visualization and Computer Graphics*, 9(4):454–462, 2003.