It is also interesting to determine the difference between the performance of interactive visualizations and static visualizations, since it is generally assumed that interactive visualizations tend to outperform static visualizations [37]. Some of the prior work has also looked into this for 3D visualizations [19], but in general the influence of interaction in visualization is not very well understood empirically [37]. Unlike the interest in evaluating the performance of 3D renderings, cross section based approaches remain a subject of lesser interest. However, there are some earlier works examining the performance difference for specialist medical tasks [16, 38].

Even though some empirical studies about the relative performance of techniques have been done, there is a need to make an empirical examination of whether there is a general performance difference between 3D renderings and cross sections. Motivated by this, the present study aims to determine whether there is a performance difference between the two visualization techniques for the task of identifying a known surface. To ensure that the study would measure general performance, a simple and common task was chosen. Being able to correctly identify features in a data set is considered to be a basic task for 3D spatial data [26], but there is a lack of studies examining this task for visualizations of 3D spatial data. We also examined both interactive and static versions of the two techniques, to determine if the interactive versions outperform the static versions, which is generally assumed in the visualization community. When studying the usefulness of a visualization technique, the experience required to use it in practice must also be taken into account. It is often assumed that highly experienced participants outperform those with little experience, particularly with cross section visualizations [10]. Therefore, we also investigated whether highly experienced participants with regards to using image data perform better than those with comparatively little experience. An empirical study of the performance difference between 3D renderings and cross sections will result in a better understanding of when to use the two techniques and serve as a basis for further investigation of how to approach the question of 2D vs. 3D.

2 Related Works

2.1 3D Visualization and Graphics

There have been a number of studies in 3D visualization comparing similar techniques for common tasks. Reaction time and number of correct responses are commonly used measures for performance in studies comparing different visualization techniques and approaches [13, 4, 19, 36]. However, there are also some studies that measures the participant's subjective perception of the techniques [18, 5].

Pike et al. argued that when examining the performance of visualization techniques, the relationship between method of interaction and task must be investigated. This will enable the correct interaction mode to be made available for each task [37]. For 3D spatial data there is a lack of empirical research into the utility of interaction, with some exceptions such as Grosset et al.'s study on depth of field [19] and Baer et al.'s investigation of Ghosted Views for vascular structures and embedded flow [5]. Further, Pike et al. identified quantitative evaluation with interaction isolated as an experimental variable as an important topic of research [37]. This should be measured according to the following variables: efficiency, effectiveness and satisfaction [21, 37]. Single experiments or studies usually measure either efficiency and effectiveness with the reaction time (RT) and number of correct responses (CORR) or the participant's overall satisfaction with each technique [29]. We expand the RT and CORR pair to include a confidence (CONF) variable to measure the participant's satisfaction with each response. A similar approach was used by Baer et al. where in addition to measuring CORR and RT, the participants were asked on a questionnaire to state which techniques they preferred after the experiment [5]. Kersten-Oertel et al. also presented a questionnaire at the end of the experiment where the participants were asked to rate the ease with which they could conduct the task [24]. Our approach was an improvement over the questionnaire approach because it tests how satisfied the participants are that they have successfully completed an individual task. This is a more specific question than asking the participants to judge an aggregate of a number of individual tests.

Another interesting work by Penney et al. examined the effect of global vs. local illumination together with texture and motion on the effectiveness of streamtube visualization [36]. Their experiment tested 26 participants on their ability to compare depth, continuity and intersections with foreign objects. The results showed that the effect of global vs. local illumination depended on the task. On the other hand motion tended to increase the reaction time but lead to higher accuracy. The study expanded upon previous work by Weigle and Banks [45]. However, they only examined local vs. global illumination and had a relatively small sample size of 5 participants.

2.2 Medical Case Studies

Comparing different visualization techniques for medical Computational Tomography (CT) data has been a topic of interest for medical researchers. Fox et al. presented a study comparing the performance between CT slices, 3D reconstructions and Multi-Planar Reformation (MPR) for identifying maxifillofacial (head and neck) fractures. They found that the 3D and axial CT slice visualizations outperformed MPR when it came to correctly identifying whether fractures were present, with no difference between 3D and CT slice visualizations. The results showed that the number of fractures detected resulted in a ranking with the highest numbers correctly identified for CT slices, then 3D renderings and lastly MPR. However, the study could be considered a case study since it only had three participants, although they had a large number of stimuli per participant (108). Furthermore, the study investigated tasks that were highly dependent on medical knowledge and by doing so confounded variation in professional expertise with the visualization technique [16]. The same could be argued for a study by Dos Santos et al. [15]. The study consisted of 2 participants, and 56 stimuli were used to identify the performance differences between axial slices, MPR and 3D renderings for the diagnosis of maxillofacial fractures. This was examined by investigating the sensitivity and specificity of the visualization techniques in identifying fractures. The study found that the results depended on which anatomical region the test was performed in. Axial CT slices and MPR outperformed 3D rendering for sensitivity in the maxillary buttress and MPR and 3D renderings outperformed CT slices for sensitivity in the orbit region. No significant results were found for specificity. Remmler et al. conducted a study to determine the utility of 3D CT and MPR for nasoorbitoethmoidal fractures. Their results were more varied: 2D CT outperformed 3D CT in inspections of the medial orbital wall and 3D CT outperformed 2D CT for diagnosis in the medial maxillary buttress [38]. A questionnaire based study examining the utility of 2D reformatted and 3D rendered CT images was conducted by Alder et al. [1]. The study was performed with 29 expert participants who reported their perceived utility of the two techniques. They found that 97% of responders stated that the 3D renderings were useful and provided additional information, but only 34% stated that the 3D renderings were essential. It should be noted that no comparison was made between the two techniques in that study. A weakness with respect to the general applicability of studies of 2D vs. 3D CT visualizations from the medical domain is that there is a tendency to focus either on case studies [27] or the perceived utility of techniques without direct comparison of the visualization techniques [1]. Additionally, the use of expert tasks enforces a reliance on prior experience which makes it difficult to generalize from the results.

Based on the present studies, it could be argued that the performance of 3D renderings vs. cross sections for general use is an open question, particularly when considering non-expert users. The fact that some studies, such as the one by dos Santos et al., showed varied results for the same task for different regions of the data set indicates that to have a more generalizable result less complex data and tasks should be examined [15]. It could also mean that no meaningful general statement can be made about the actual performance, because it is too dependent on the task and data. Furthermore, because the prior studies use few participants, it is difficult to determine this without conducting further studies using larger samples. These problems are addressed in the present study by using a larger sample of participants, comparing high experienced vs. low experienced participants and by

using a non-expert task.

2.3 Greebles

Some previous work from perceptual psychology has examined 3D shape perception using physical ground truths. In particular abstract physical shapes called Greebles have been used to research object and facial recognition. However, their relation to general object and facial recognition remains controversial. As an example, Gauthier et al. demonstrated that it was possible for patients with severe visual object agnosia and dyslexia, but intact face recognition, to fail to recognize Greebles [17]. Vuong et al. on the other hand, showed that incorrect pigmentation of Greebles and faces resulted in longer reaction time and a lower number of correct responses for both [44]. By using fMRI scans of healthy participants, Brant et al. found evidence of an inversion effect when examining the Greebles [8], which is most commonly associated with face recognition. Based on the literature it is difficult to determine the exact relation between facial and object recognition for Greebles. However, there is a general consensus in the perception community that faces and general objects are processed differently by the brain [8]. To ensure that our objects would be processed through object recognition rather than facial recognition, we chose common objects that most people would be familiar with as our physical ground truths.

2.4 Virtual Reality and 3D User Interfaces

It is generally held by the 3D interaction community that 3D input based interaction outperforms 2D input for 3D renderings [7]. This has been backed up by user studies. As an example, Hinckley et al. [20] demonstrated that 3D trackers outperformed mouse and arc-ball for 3D rotation tasks. Results from the Virtual Reality (VR) domain, demonstrated that there is a difference between user performance in 3D VR and 3D projected onto a 2D surface (e.g. a screen). Especially interesting are 3D interaction instruments that have been developed in that field [34, 23, 22]. For instance, Jackson et al. presented a 3D interaction instrument that was specialized towards the visualization of fiber structures [23]. This could lead to different results from the present study's 2D input. Possible performance differences between 3D and 2D interaction were not taken into account by the present study, but examining the performance of this and other 3D interaction techniques could make for an interesting follow-up study.

3 Method

In order to test the assumption that 3D renderings are better at gaining an overview of the data while cross sections are more suited for detailed analysis of the data [28, 33], the present study examined the performance of 3D renderings and cross sections for the task of identifying a known surface. An argument for using the identification of a surface rather than depth tasks is that prior case studies comparing 3D renderings with cross sectional views within the medical domain had been carried out for other feature identification tasks [16, 15]. Regarding prior work on depth perception, it is believed that high resolution form and depth perception may utilize different neurological pathways [31]. This means that results taken from the studies examining depth perception may not be directly generalizable to tasks relating to the identification of or location of shapes. Since it should be possible to compare our findings to those of prior case studies [16, 15] this further strengthens the argument for using a form identification type task for the present study.

Because of the possibility of prior knowledge affecting the outcomes, it was important to take into account the participants' relevant experience. In particular participants with extensive experience working with image data may outperform participants with less experience, especially in the case of cross sectional or MPR views. To determine the effect of experience with image data, the participants were divided into a group with high and a group with low image data experience. The objects for the identification task were chosen from relatively common shapes such as hands, cups, etc. This was done to ensure that differing user familiarity with shapes would not become a confounding factor.

There is a general preference for interactive visualizations in the visualization community and an assumption of better performance for interactive techniques. It could be argued that this is caused by multimodal integration, which is the concept of how the integration of several sensory modalities are used by perception to gain an understanding about an observed phenomenon [2]. The integration of multiple senses has been shown to be useful in other identification tasks, such as the combination of facial and vocal information for person identification [11]. By utilizing interaction, more sensory modalities are made available. For the visualization community, interactivity is often considered a key feature in the performance of many visualization approaches, especially in fields like Visual Analytics [35, 14, 37]. In high throughput environments, such as the medical profession, it has been more common to use static visualizations [10]. Since the performance of interaction in visualization is not fully understood, the present study was designed to test both interactive and static versions of the visualizations.

An advantage of taking all of these factors into account in the same experiment is that the results are based on data that have been collected from the same participants under the same conditions. This reduces the risk of incorrect results that may occur when comparing results from multiple different studies with different conditions and assumptions. It also reduces the statistical error rate from analyzing many different experiments separately.

Based on the problem statement and reasoning above, the following primary hypotheses were derived:

1.a Our first hypothesis was that 3D techniques should have improved results over 2D techniques. This will be shown as a main effect of visualization techniques.

1.b Furthermore, the interactive versions should be superior to their static versions. This will be shown as a main effect of the degree of interactivity.

1.c An interaction effect of technique by degree of interactivity was expected to occur, resulting in a ranking of performance of the different techniques, with interactive 3D better than static 3D, then interactive 2D followed by static 2D.

2.a The second hypothesis was that highly experienced participants should outperform the low experience participants. This will be shown as a main effect of experience.

2.b Furthermore, we expect this to be most obvious when comparing the slicing technique since it is the most commonly used technique by our expert users. This will be shown as an interaction effect between degree of experience and visualization technique.

The experiment was set up by using physical objects as the known surfaces (ground truth). A virtual version of the physical object and similar virtual objects were used to provide a set of surfaces from which the participant would do the identification. The experiment was conducted as a series of four tests. For each test the user had to identify a surface based on a physical object. The tests showed four surfaces to choose from and used the same visualization technique for all, which allowed the results of the each test to be examined on the basis of the technique. To ensure that the experiment gives a full picture of the participants' performance, the measures of efficiency, effectiveness and satisfaction [21, 37] were used. This was accomplished by using the following three variables: 1. Whether the participant identified the correct (CORR) surface. 2. What the participant's reaction time (RT) for making the choice was. 3. How confident (CONF) the participant was that the identification was correct.

3.1 Procedure

Before the participant started the experiment, (s)he began by filling in a questionnaire registering background information. The questionnaire contained information about gender, the participant's prior experience working with image data, educational background, age, profession, whether the participant was a trained radiographer and ratings

of their experience using computer graphics. The participants self-reported their image data and computer graphics experience according to a Likert scale (1 to 5), where 1 meant no experience whatsoever and 5 meant very experienced.

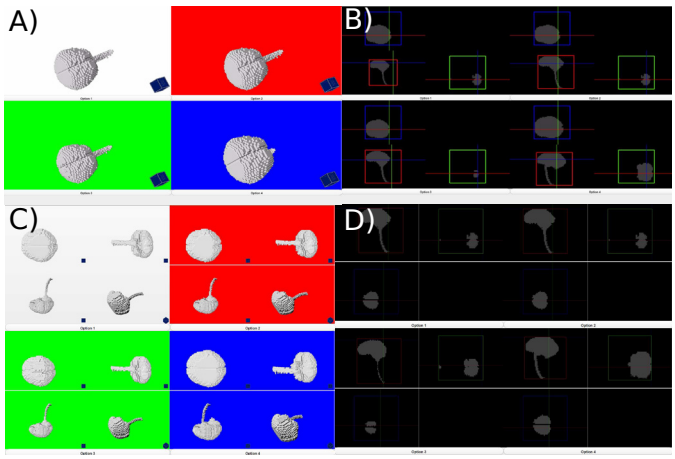

Fig. 2. Examples of the four techniques used in the experiment; A) Interactive 3D rendering (I3D), B) Interactive cross sections using Multi-Planar Reformation (I2D), C) Static 3D rendering (S3D) and D) Static cross sections (S2D). For each technique, four variations of the surface were shown. Only one of the four techniques was used for each participant per test. The participant selected the surface (s)he believed was the correct one by pressing the option button below the visualization of that surface. To make it easier for participants to distinguish between the four alternatives, especially for the static version, the 3D renderings were shown with different colored backgrounds.

Fig. 3. The figure shows the experiment setup during a static cross sections (S2D) test. The printed cup object was covered with a white sheet prior to commencement of the test to avoid participants spending an unequal time familiarizing themselves with the object beforehand. Once the visualizations were finished loading the participant was allowed to remove the white sheet and start the task of identifying the virtual version of the object.

In the next step, the participant was given a short tutorial on how the experiment was to be carried out. The tutorial took 3-5 minutes and introduced the basics of how cross sections/MPR and 3D rendering work, as well as how to interact with the visualizations. The tutorial used an ear object as a physical ground truth. This object was not used later in the experiment. All the alternatives for identification of the surfaces were identical and correct, meaning that there was no actual identification of surfaces during the tutorial. This was done because the focus for the tutorial was not to teach the participants to identify surfaces, but rather to familiarize them with the setup and the basics of how the visualization techniques and interaction worked. The tutorial showed one example for each of the four versions of the techniques. For each example the window was divided into four quadrants with each quadrant visualizing one of the four versions of the surface. The techniques used were interactive and static versions of cross sections/MPR and a 3D solid surface rendering. We use the following abbreviations in the text: I3D for Interactive 3D, I2D for interactive slices, S3D for static 3D and S2D for static slices. Examples can be seen in Figure 2. The I3D tests had a single interactive 3D visualization in each quadrant as can be seen in Figure 2.A. It could be rotated, translated and scaled by mouse interaction. The original viewpoint was chosen as a random side of the object. For the I2D, slice visualizations were shown with one Multi-Planar Reformation (MPR) visualization per quadrant (see Figure 2.B). The slices were shown from the three cardinal directions and were located inside a square with a colored outline. Each slice also had a colored line representing the intersection with the slice inside the square with that color. The participant was able to move the slices with the mouse either by clicking and dragging the colored lines to move the intersecting slice or by clicking and dragging inside of a slice to move that particular slice. The original position of the slices was set as a random pre-chosen position intersecting the object. For each of the four surfaces the S3D tests showed static snapshots from the 3 cardinal directions (top, front and side), as well as a snapshot of the object angled 45 degrees around both the X and Y axes. This last direction was chosen because it was found in prior studies that users prefer views from oblique angles [18]. Combined with the three cardinal directions this has the advantage of very concisely presenting most of the surface, as can be seen in Figure 2.C. Lastly, for S2D the same visualization scheme as for the I2D was used with the cross sections chosen manually beforehand by the experimenters. The cross sections were chosen to best highlight the differences between the surfaces (see Figure 2.D). The advantage of choosing the same number of images as in I2D and S3D was that it preserved a similar type of layout across the different visualizations. This helped to ensure that the approaches were evaluated based on the chosen visualization technique and not for instance the number of views.

After the tutorial the participants performed the experiment. Each test in the experiment was conducted using different surfaces in each of the four quadrants and the participant had to identify which one of the four corresponded to the physical object. For these tests the reaction time (RT) was the time it took the participant (in milliseconds), from when the visualizations were finished loading until (s)he selected a surface. Then the participant had to rate his/her confidence (CONF) from 1 (not confident) to 5 (completely certain). It was also registered whether the participant selected the correct (CORR) surface. A total of four tests were performed (see Figure 4 for the objects used in each experiment).

To ensure that a complete picture was given, both for each test participant as well as for each of the four surfaces, the visualization techniques being presented were varied in a semi-random way for each test. All participants were presented all techniques and all objects were visualized.

The exact procedure of each individual test was that the participant begins the test by being shown a screen stating which type of test will come next (I2D, I3D, etc.) and a ”Show Test” button. When the participant pressed the ”Show Test” button the visualization started loading. The physical object that was to be identified was located to the participant’s right or left side in a queue, with a white sheet covering each physical object (see Figure 3). Once the visualization was finished loading the timer started on the RT measure. The participant could now remove the white sheet and start to identify which of the four visualizations corresponded to the physical object. For the objects made by 3D printing the participant could pick them up and move them around as (s)he wished. The Lego object was more fragile

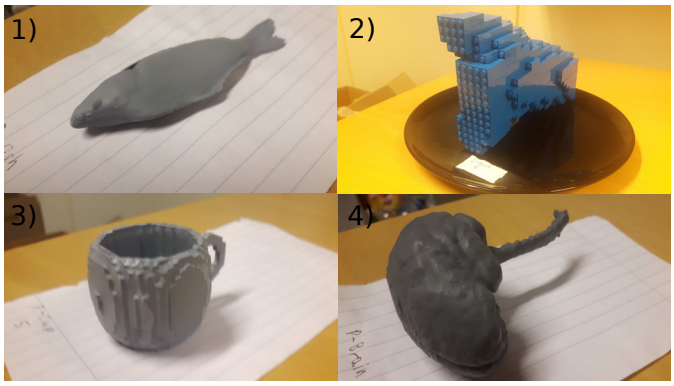

Fig. 4. The physical ground truths shown in the order they were presented: 1. Printed fish, 2. Lego hand, 3. Printed cup and 4. Printed brain.

and was therefore presented on a plate (see Section 3.4 for more information on the 3D printed vs. Lego objects). The plate could be lifted up and moved around, as well as rotated around the axis orthogonal to its flat surface. Limited pitch and roll was also possible. In order to record the CORR variable, the participant selected a surface by pressing one of the four option buttons under the four visualizations. The RT variable was registered and the participants had to rate their confidence (CONF) by pressing a button with a label between 1 and 5 to represent their score. The test was then completed, the current object moved out of the queue and the procedure was repeated for the next test.

The participants were instructed to take as much time deciding as they felt they needed to be comfortable with their selection. This was done to reflect real life scenarios where people spend as much time as needed to make an adequate response. It should be noted that due to this instruction, the results are probably erring more on the side of accuracy than speed.

3.2 Pilot Studies

Before the present study was carried out, two pilot studies were performed with 8 participants in each of the two pilots. Although the pilot samples were too small to perform statistical tests, they were helpful in refining the process. Among the changes made was adding a tutorial to familiarize the participants with the basics of how the visualization techniques worked and how they were supposed to react to the practical portions of the experiment. This ensured that all the participants knew the basics of what the techniques were and how they worked before they started. Also a question was added to the pre-experiment questionnaire adding meta-data about prior graphics experience for each participant. These improvements were mainly made after the first pilot study, with the second pilot primarily serving to check that the experiment was running as expected.

3.3 The Participants

The experiment was carried out on a total of 45 participants. Of these a total of 16 where women and 29 were men. Thirty-two participants were recruited primarily from the students and staff at the Department of Informatics of the University of Bergen, as well as 13 domain experts from the Radiography Department at the University Hospital of Bergen. The age of the participants examined ranged from 18 to 61 of which 1 was aged below 20, 22 between 20 and 29, 9 between 30 and 39, 4 between 40 and 49 and 9 above 50. Thus, the participants to some degree reflected the general working population. The data were gathered isolated from random noise at the offices of the Department of Informatics University of Bergen and the Radiography Department at the University Hospital of Bergen.

3.4 Setup

The visualization part was implemented using the standard ray casting and slicing plug-ins for VolumeShop [9]. The participants interacted with the virtual part of the experiment through a setup using the VolumeShop webserver for access to the VolumeShop visualization capabilities and HTML and JavaScript for the questionnaire and experiment setup. This made for a setup that could be rapidly altered to suit our needs during the pilot studies.

The physical objects were made by a process of first extracting a suitable surface from volumetric CT scans using Paraview [42]. The surface was then either printed with a 3D printer or modeled in Lego. It was converted to the Lego-format .ldr and edited into four versions using Lego Digital Designer. The Lego Digital Designer's Building Guide Mode provided an interactive instruction set for building the physical Lego object [32]. Lego models were chosen because they are practical to use and because editing the surfaces in Lego Digital Designer was relatively straight forward. Approximately 500 individual bricks ended up being needed for the model, even with heavy down-sampling of the data-set. In the last step the four versions of the surfaces, three of which had been edited, were converted back to a voxelized representation and used for the visualizations. The original Lego object was made because we originally had easier access to Lego than to 3D printing. Because the identification of the Lego object did not require pitch or roll, we do not believe that the limited ability to manipulate the model influenced the results of the study.

Collection of data on the participants' performance was done through several means. The questionnaire and performance data were stored in a MySQL database. The website interfaced with the MySQL database through a Laravel-server based solution, which was accessed by cross-site scripting from the visualization implementation on the VolumeShop server. Screen capture videos of the participants entering the answers to the questionnaires and working on the tests were used as a backup solution.

The physical setup consisted of a Dell UltraSharp 2408WFP with the dimensions 22x19.6 inches running at a 1920x1200 resolution and 59Hz. The programs were run on an Alienware Desktop PC with a 3.40 GHz CPU, 32 GB of RAM and GeForce GTX 660 graphics card running Windows 10. The lighting setup and the location of the physical objects varied between the different experiment locations.

Assignment of participants to the high and low experienced group was done based on the median split of the self evaluation of experience with imaged data. There were 27 participants in the low experience and 18 in the high experience group. All 13 participants from the radiography department ended up in the high experience group. The RT variable was square root transformed in order to achieve better homogeneity of error variance, thus normalizing the scores [25].

4 Results

The data were analyzed using a 2 (high vs. low degree of experience) $\times$ 2 (3D vs. 2D) $\times$ 2 (interactive vs. static) ANOVA design. A separate ANOVA was run for each of the three variables that were investigated (CORR, RT and CONF). Fisher Least Squares Differences (Fisher LSD) were used for post-hoc analysis. No post-hoc corrections were performed, since only a few contrasts were of interest [3]. Since the hypotheses were defined with a direction of means, one-tailed versions of the Fisher LSD tests were used where appropriate. The one-tailed tests are marked *one-tailed* in the text. For the same reason and because only a limited number of specific contrasts were of interest, non-significant interaction effects were followed up with Fisher LSD [46]. The results of the statistical tests are given with 5 decimal points of precision, if the value was lower than what can be shown with 5 decimal points of precision the value set is shown as 0.

A visual summary of the findings can be seen in Figure 5. Tables 1 and 2 show the specific values of the tests.

A main effect of visualization techniques was found on the CORR data with 3D outperforming 2D, $F(1, 43) = 16.856$, $p = 0.00018$. Also, a main effect of the degree of interactivity on the CORR data was found with interactive outperforming static, $F(1, 43) = 4.3289$, $p = 0.04346$. No interaction effect between visualization technique and

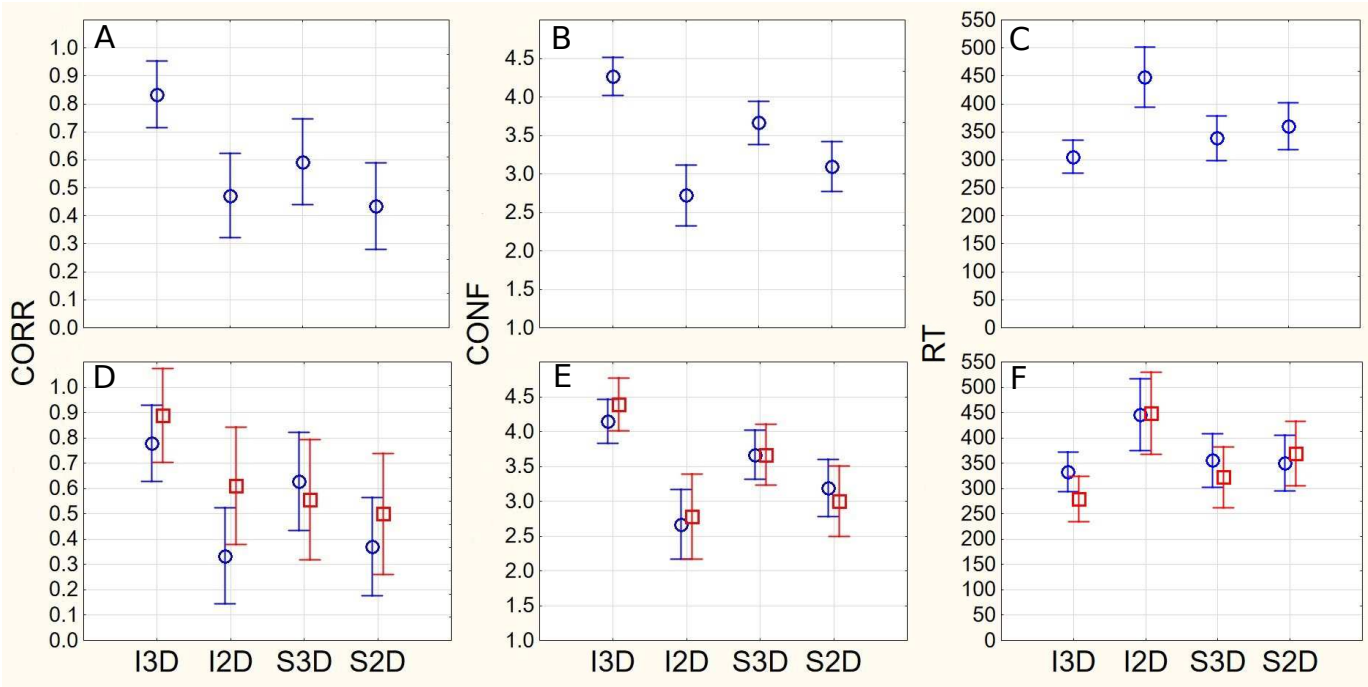

Fig. 5. The figure shows the means for Interactive 3D (I3D), Interactive 2D (I2D), static 3D (S3D) and Static 2D (S2D). The data were based on the two-way interaction of visualization techniques by degree of interactivity (panels A-C). The figure also shows means separated for the high (Red) and low (Blue) experienced group, based on the three-way interaction of Groups by visualization techniques by degree of interactivity (panels D-F). The error bar denotes 95% confidence interval and the panels show the data for number of correct responses (CORR), degree of confidence (CONF) and reaction time in milliseconds (RT).

degree of interaction was found, $F(1, 43) = 1.634$, $p = 0.20806$. A follow-up of the interaction effect showed that I3D had a higher number of correct responses compared to I2D ($p = 0.001378$), S2D ($p = 0.000765$) and S3D ($p = 0.025228$, *one-tailed*).

Another main effect of the visualization techniques was found on the CONF data with 3D outperforming 2D, $F(1, 43) = 47.841$, $p = 0.00000$. No main effect of the degree of interaction was found on the CONF data, $F(1, 43) = 0.69382$, $p = 0.40947$. An interaction effect between visualization technique and degree of interaction was also found, $F(1, 43) = 15.546$, $p = 0.00029$. The follow up LSD test revealed that participants reported higher confidence in their choices with I3D as compared to I2D ($p = 0.000000$), S3D ($p = 0.001542$) and S2D ($p = 0.000000$). Participants were shown to have higher confidence in S3D than in I2D ($p = 0.000001$) and S2D ($p = 0.002232$). Lastly, the participants demonstrated a higher confidence in S2D than in I2D ($p = 0.023913$).

A main effect of the visualization techniques was also found on the RT data with 3D outperforming 2D, $F(1, 35) = 18.516$, $p = 0.00013$. No main effect of the degree of interaction was found on the RT data, $F(1, 35) = 3.8951$, $p = 0.05636$. However, this constituted a borderline significant effect, where static outperformed interactive. An interaction effect between visualization technique and degree of interaction was also found, $F(1, 35) = 17.403$, $p = 0.00019$. A follow up LSD test showed that participants were slower using I2D than I3D ($p = 0.000000$), S2D ($p = 0.000108$) and S3D ($p = 0.000008$). It was also demonstrated that participants were faster with I3D than with S2D ($p = 0.019957$).

This means that hypothesis 1.a is well supported by our findings, since a significant main effect of the visualization technique was found for all three measures. We found some support for hypothesis 1.b, since a main effect of the degree of interaction was identified for the CORR measure and there was a borderline main effect for RT. There is also some support for hypothesis 1.c, and this will be examined further in the discussion section.

No main effect of degree of experience was found on the CORR data, $F(1, 43) = 2.128$, $p = 0.15191$, CONF data, $F(1, 43) = 0.043$, $p = 0.83684$ or RT data, $F(1, 35) = 0.274$, $p = 0.60391$. Because of this, hypothesis 2.a was rejected. Also, no interaction effect between degree of experience and visualization techniques was found on the CORR data, $F(1, 43) = 2.150$, $p = 0.14985$, CONF data, $F(1, 43) = 0.264$, $p = 0.61025$ or RT data, $F(1, 35) = 2.115$, $p = 0.15474$. However, the non-significant interaction effects for hypothesis 2.b were followed up with LSD post-hoc analysis due to a specific hypotheses of the direction of means [46]. The results showed no significant difference in performance between the high and low experience group on 2D CONF ($p = 0.883914$) and RT ($p = 0.762299$). However, the high experience group performed better on 2D CORR ($p = 0.042643$). No significant differences between the groups were found for 3D CORR ($p = 0.851988$), CONF ($p = 0.635336$), or RT ($p = 0.235011$).

No interaction effects between degree of experience and degree of interactivity were found on the CORR data, $F(1, 43) = 1.5584$, $p = 0.21866$, CONF data, $F(1, 43) = 0.93360$, $p = 0.33933$ or RT data, $F(1, 35)$ 0.4183, $p = 0.522028$. Also, no 3-way interaction effects were found on the CORR data, $F(1, 43) = 1.634$, $p = 0.90804$, CONF data, $F(1, 43) = 0.0135$, $p = 0.051391$ or RT data, $F(1, 35) = 0.007$, $p = 0.93642$.

4.1 Summary of Results

The results showed a main effect of visualization techniques with improved performance of 3D techniques over 2D techniques. This was the case for all dependent measures; CORR, CONF and RT. Furthermore, a main effect of interactivity with superior performance of the interactive techniques was found for CORR. A follow-up of visualization technique by degree of interactivity showed that there was indeed a ranking of the techniques. No interaction effects of degree of experience by visualization techniques were found. However, a follow-up of the non-significant interaction revealed that the high experience group outperformed the low experience group on number of correct responses for I2D.

From this it can be concluded that there was significant support for

	I3D			I2D			S3D			S2D		
	CORR	CONF	RT	CORR	CONF	RT	CORR	CONF	RT	CORR	CONF	RT
I3D				.001378	.000000	.000000	.025228 **O**	.001542	.123922	.000765	.000000	.019957
I2D	.001378	.000000	.000000				.166095	.002232	.000008	.841451	.023913	.000108
S3D	.025228 **O**	.001542	.123922	.166095	.000001	.000008				.114712	.002232	.394443
S2D	.000765	.000000	.019957	.841451	.023913	.000108	.114712	.002232	.394443			

Table 1. The matrix shows the significant results of the post-hoc Fisher Least Squares Distances based on the interaction effect between visualization techniques and degree of interaction. In the first column and the first row are listed the four compared techniques. In the second row the three variables that are being tested are listed. Each cell contains the p value of the individual comparison, and significant results are colored green for the better results on the y-axis and yellow for better results on the x-axis. One-tailed tests are marked as **O**.

	CORR			CONF			RT		
	DF = (1,43)			DF = (1,43)			DF = (1,35)		
Analysis	F	p	Obs. Power	F	p	Obs. Power	F	p	Obs. Power
Degree of Experience	2.128	.15191	.297106	.043	.83684	.054716	.274	.60391	.080206
Visualization Techniques	16.856	.00018	.979932	47.841	.00000	.666666	18.516	.00013	.986867
Degree of Interactivity	4.329	.04346	.529545	.694	.40947	.128767	3.895	.05636	.48353
Experience × Technique	2.150	.14985	.299665	.264	.61025	.079345	2.115	.15474	.293122
Experience × Interactivity	1.558	.21866	.230570	.934	.33933	.156844	.418	.52203	.096435
Technique × Interactivity	1.634	.20806	.239426	15.546	.00029	.970847	17.403	.00019	.981904
Experience × Technique × Interactivity	.014	.90804	.051480	.013	.91083	.051391	.007	.93642	.050700

Table 2. The table presents Degrees of freedom (DF), F and p values as well as observed Power for all main and interaction effects in the present study. Instances with $p \leq 0.05$ denote a significant effect.

3D outperforming 2D. There was also some support for interactive outperforming static. The results revealed a ranking of the techniques with I3D showing the best performance followed by S3D then S2D and lastly I2D. Due to no significant main effect of degree of experience, no support was found for experience generally affecting performance. However, some support was found for high experience users outperforming low experience users with I2D in the follow up LSD tests on the CORR variable.

5 Discussion

Regarding hypothesis 1.a and 1.b we found that 3D renderings do indeed outperform cross sections. This was expected because 3D renderings provide more information about surfaces per view than cross sections. Hypothesis 1.b, that interactive outperforms static, was partially supported. Specifically, support was found in number of correct responses and a borderline significance in the reaction time. A possible reason for the more limited support of hypothesis 1.b was that the task in the present experiment did not require complex investigation since the physical objects in this experiment were of known objects. Therefore an overview could be sufficient. This view is in line with Munzner's statement that the main advantage of interactivity is that it allows investigation and analysis of complex data [33].

Further insights into this can be gained by examining the follow-up of the significant interaction effect between visualization technique and degree of interactivity. Figures 5.A-C shows that I3D consistently outperformed or broke even with S3D, whereas I2D consistently either underperformed or broke even with S2D. This could be because manually picked slices present sufficient information and the extra information presented by the interactive version did not add to the users' understanding of the object. I2D underperforming compared to S2D was puzzling since it contradicts conventional wisdom that more investigation increases the participant's understanding of the data [33]. A possible explanation for the underperformance of S2D was that the average participant had low competence with I2D and that the interaction did not provide further information. Some support for this explanation was provided by the fact that the high experience group slightly outperformed the low experience group in the number of correct identifications using I2D.

Hypothesis 1.c stated a ranking of the technique/degree of interactivity. The ranking found based on the CONF data (Figure 5.B) was: 1. I3D, 2. S3D, 3. S2D and 4. I2D. In general parallel results were found for the CORR (Figure 5.A) and RT data (Figure 5.C). For CORR there was no difference between I2D and S2D. This indicates that the participants were able to correctly assess the performance of each technique. It should also be noted that the relative ranking of S2D and I2D was the opposite of what was expected, which could partially help to explain why hypothesis 1.b was only partially supported.

This also backs up the common practice of using the participant's subjective perception of the techniques. The rankings, however, were very clear compared to rankings found by Baer et al. [5]. Although they found that Ghosted and Depth Enhanced Ghosted view outperformed non-Ghosted views in self-reported preference and the accuracy, they were unable to demonstrate similar results for Ghosted vs. Depth Enhanced Ghosted views. A comparison between the present study and Baer et al.'s study is shown in Table 3.

We did not find support for hypothesis 2.a that experience leads to better performance. A likely reason for this result was that the tasks and objects were chosen in such a way that the focus was on the visu-

Comparisons	Present Study	Fox et al. [16]	dos Santos et al. [15]	Baer et al. [5]
Number Of Participants	45	3	2	86
Number of Stimuli	4	108	56	Surface Orientation Task: 36 Flow Task: 32 Depth Task: 36
3D vs. 2D Techniques	3D Superiority	Mixed Results Based on Layered Study	Contradictory Results Based on Layered Study	Not Examined
Interactive vs. Static	Interactive somewhat better	Only examined static	Not compared	Mixed Results
Experienced users have better results	Only for Interactive 2D	Not Examined	Not Examined	Not Examined
Confidence Predicts Performance	Yes	Partially Discussed	Not Examined	Inconclusive
Variables	CORR, RT & CONF	CORR, RT & CONF	CORR & CONF	CORR, RT & Preference

Table 3. The table compares the results of the of the present study with the studies by Fox et al. [16], dos Santos et al. [15] and Baer et al. [5]. The studies by Fox et al. and dos Santos et al. investigated the participants' performance identifying maxifillofacial (head and neck) fractures for multiple regions. Baer et al. investigated the influence of ghosted views for determining depths, surface orientation and detection of flow. When referring to techniques the table refers to Multi-Planar Reformation marked as 2D and 3D rendering marked as 3D. The preference that Baer et al. measured was part of a survey after the experiment where the participants were asked to rate their preference of the examined techniques against each other. This differs from CONF in that whereas Preference measures the compared impressions after a number of different stimuli and techniques, CONF measures the participants' certainty that the individual decision was correct. An important consequence of this is that since the CONF measure is part of the experimental data itself it can be included in an ANOVA with CORR and RT and that it reflects each individual data point rather than the overall impression.

alization techniques and that prior knowledge was controlled for. This means that when looking at a task and data that novices and experts within the field are equally familiar with, equal performance should on average be expected. This could arguably represent a positive aspect of the techniques, since it indicates that they can be mastered quickly. An exception from this was found for number of correct answers using I2D. As seen in Figure 5.D, the high experience group slightly outperformed the low experience group. This finding was in line with hypothesis 2.b, that any difference would be particularly visible for the slice technique since this was the most commonly used method by our domain experts.

A general take-away from this study was the importance of choosing the correct visualization technique. Of the two techniques we tested, the technique that showed highest performance was 3D rendering. For interactive vs. static, which approach worked best depended on which of the two visualization techniques were used and previous experience was not important.

5.1 Comparison With Case Studies

There have been some prior studies within the medical domain examining the differences between 2D and 3D renderings of CT data. Table 3 compares results from two of these studies with results from the present study. Fox et al. used three expert participants in radiology, with the task of examining CT data of the skulls from nine cadavers, to attempt to identify fractures [16]. The reaction time for making a diagnosis in this study was shown to be shorter with 3D compared to axial CT slices. MPR performed similar to axial CT slices when fractures were absent and similar to 3D when the fractures were present. The findings were supported by our study, since no difference was found between the reaction time of S2D and S3D. Because their axial CT slices are not directly comparable with our MPR slices, only 3D vs. MPR will be considered from here on.

For the orbit region, 3D renderings were shown to detect the presence of injuries with higher accuracy than MPR, although no difference was found in the maxilla. This is in line with the present study finding that S3D outperformed S2D for the CORR variable (see Figure 5.A). The same relationship was observed when participants were asked to count the number of fractions. S3D outperformed S2D both in the present study and in Fox et al.'s study, which strengthens the conclusion that there is little support for a difference in how the performance of the visualization techniques ranks between participants with high and low experience. The fact that the same ranking between S3D and S2D was shown in a case study using an expert task and requiring expert domain knowledge also indicates that our results have a degree of generalizability.

Fox et al.'s results for the axial CT images were more complex. In general they found that slices tend to outperform both the 3D rendering and the MPR. However, slices are similar to the MPR and it seems reasonable to assume that they will perform similarly. A possible reason for the axial CT slices' performance could be that radiologists have expert knowledge on how to perform this particular task using slices sampled along an axis. Further studies would be needed to clarify why the axial CT slices outperform MPR in diagnosing maxillofacial trauma [16].

Fox et al. also ranked the techniques according to confidence. For the orbit region they found a ranking of the reported confidence going from axial CT slices, 3D renderings, to the poorest ranking of MPR. The rankings of the 3D renderings and the MPR corresponded to the correct detection rate for the orbital region. These findings are in agreement with our own results. However, it should be noted that the study by Fox et al. did not present a ranking of confidence between 3D rendering and MPR for the remaining three regions. This means that it is hard to draw conclusions about possible differences in confidence and performance from Fox et al.'s study.

Another interesting case study by dos Santos et al. also used expert participants to identify fractures in skulls [15]. Even though the two studies compared the same three methods (axial CT slice, 3D and MPR), the results were quite different. The study examined both sensitivity and specificity, which required the authors to measure both number of correct responses and confidence ratings. Since the present study has been looking at a positive identification task, we will only be discussing the sensitivity.

In the maxillary buttress, dos Santos et al. found that both axial slices and MPR outperformed 3D renderings. This stands in contrast to Fox et al.'s work which found that CT slices and 3D renderings tended to outperform MPR [16]. In the orbit region dos Santos et

al. found 3D renderings and MPR to outperform CT slices and in the zygomatic-maxillary complex very little or no difference was found. The results of dos Santos et al.'s study seemed to vary more between the regions than was the case with Fox et al. and the results in general run contrary to the findings of the present study. It is not readily apparent why the two studies differ in their results. A possible reason could be that since the case studies are conducted with few participants, an outlier may have been affecting the results. Other possibly confounding variables are differences in the experimental setups, differing data sets or variation in professional competence of participants. Because of the similarities in the results we conclude that Fox et al.'s study [16] was more representative of the findings of the present study. The results of dos Santos et al.'s study differed from our own in a more substantial way, in particular because the relative performance of MPR and 3D rendering was dependent on in which region the tests were conducted [15].

6 Conclusion

In the current experiment we examined the utility of slices vs. 3D renderings for identifying surfaces, as well as the effect of interactivity. A novel approach of using a physical object as a ground truth was employed to ensure a good basis for evaluating the task according to the variables reaction time, confidence and whether the correct surface was selected. Based on their prior experience with image data, participants were divided into high and low experience groups and we examined whether the experience affected the results. The physical objects were chosen to be items that all the participants should have a similar degree of familiarity with. This controlled for the confounding factor of the variation in the participants' prior domain specific expertise. This ensures that the tests are measuring the performance of the actual technique. We found support for 3D renderings outperforming slice renderings. Some support was also found for interactive visualizations outperforming static visualizations. We also found a ranking of the techniques: 1. interactive 3D, 2. static 3D, 3. static 2D and 4. interactive 2D. The only significant result for the low vs. high experience groups was that the high experience group showed more correct results for interactive 2D slices. This indicates that the best technique for identifying surfaces was 3D rendering. The use of interactive versions of the techniques was found to be beneficial for 3D renderings, but the opposite effect was found for static cross sections vs. MPR. The experience of the participants was found to be largely unimportant for this task.

Acknowledgments

The authors wish to thank the radiography department at the University Hospital of Bergen for letting its employees take the time to take part in this study. This work has been supported by the MedViz lighthouse project IllustraSound.

References

[1] M. E. Alder, S. T. Deahl, and S. R. Matteson. Clinical usefulness of two-dimensional reformatted and three-dimensionally rendered computerized tomographic images: literature review and a survey of surgeons' opinions. *Journal of oral and maxillofacial surgery*, 53(4):375–386, 1995.

[2] R. A. Andersen. Multimodal integration for the representation of space in the posterior parietal cortex. *Philosophical Transactions of The Royal Society of London B: Biological Sciences*, 352(1360):1421–1428, 1997.

[3] R. A. Armstrong. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508, 2014.

[4] A. Baer, F. Adler, D. Lenz, and B. Preim. Perception-based evaluation of emphasis techniques used in 3D medical visualization. In *Proceedings of the Workshop on Vision, Modeling, and Visualization*, pp. 295–304, 2009.

[5] A. Baer, R. Gasteiger, D. Cunningham, and B. Preim. Perceptual evaluation of ghosted view techniques for the exploration of vascular structures and embedded flow. *Computer Graphics Forum*, 30(3):811–820, 2011.

[6] C. Boucheny, G.-P. Bonneau, J. Droulez, G. Thibault, and S. Ploix. A perceptive evaluation of volume rendering techniques. *ACM Transactions on Applied Perception*, 5(4):23, 2009.

[7] D. Bowman, E. Kruijff, J. J. LaViola Jr, and I. Poupyrev. *3D User Interfaces: Theory and Practice, CourseSmart eTextbook*. Addison-Wesley, 2004.

[8] M. Brants, J. Wagemans, and H. P. O. de Beeck. Activation of fusiform face area by Greebles is related to face similarity but not expertise. *Journal of Cognitive Neuroscience*, 23(12):3949–3958, 2011.

[9] S. Bruckner and M. E. Gröller. VolumeShop: An interactive system for direct volume illustration. In *Proceedings of IEEE Visualization*, pp. 671–678, 2005.

[10] S. Bruckner, P. Kohlmann, A. Kanitsar, and M. E. Gröller. Integrating volume visualization techniques into medical applications. In *Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 820–823, 2008.

[11] S. Campanella and P. Belin. Integrating face and voice in person perception. *Trends in cognitive sciences*, 11(12):535–543, 2007.

[12] A. Cockburn and B. McKenzie. Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In *Proceedings of ACM CHI*, pp. 203–210, 2002.

[13] J. Díaz, T. Ropinski, I. Navazo, E. Gobbetti, and P.-P. Vázquez. An experimental study on the effects of shading in 3d perception of volumetric models. *The Visual Computer*, 2015.

[14] H. Doleisch. SimVis: Interactive visual analysis of large and time-dependent 3D simulation data. In *Proceedings of the Winter Simulation Conference*, pp. 712–720, 2007.

[15] D. T. dos Santos, A. P. A. C. e Silva, M. W. Vannier, and M. G. P. Cavalcanti. Validity of multislice computerized tomography for diagnosis of maxillofacial fractures using an independent workstation. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 98(6):715–720, 2004.

[16] L. A. Fox, M. W. Vannier, O. C. West, A. J. Wilson, G. A. Baran, and T. K. Pilgram. Diagnostic performance of CT, MPR and 3DCT imaging in maxillofacial trauma. *Computerized Medical Imaging and Graphics*, 19(5):385–395, 1995.

[17] I. Gauthier, M. Behrmann, and M. J. Tarr. Are Greebles like faces? using the neuropsychological exception to test the rule. *Neuropsychologia*, 42(14):1961–1970, 2004.

[18] J. Giesen, K. Mueller, E. Schuberth, L. Wang, and P. Zolliker. Conjoint analysis to measure the perceived quality in volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1664–1671, 2007.

[19] A. Grosset, M. Schott, G.-P. Bonneau, and C. D. Hansen. Evaluation of depth of field for depth perception in DVR. In *Proceedings of IEEE PacificVis*, pp. 71–76, 2013.

[20] K. Hinckley, J. Tullio, R. Pausch, D. Proffitt, and N. Kassell. Usability analysis of 3D rotation techniques. In *Proceedings of the ACM UIST*, pp. 1–10, 1997.

[21] Ergonomic requirements for office work with visual display terminals (VDTs). Standard, International Organization for Standardization (ISO), Geneva, Switzerland, 1998.

[22] B. Jackson and D. F. Keefe. Sketching over props: Understanding and interpreting 3D sketch input relative to rapid prototype props. In *Proceedings of the IUI Sketch Recognition Workshop*, 2011.

[23] B. Jackson, T. Y. Lau, D. Schroeder, K. C. T. Jr., and D. F. Keefe. A lightweight tangible 3D interface for interactive visualization of thin fiber structures. *IEEE transactions on visualization and computer graphics*, 19(12):2802–2809, 2013.

[24] M. Kersten-Oertel, S. J.-S. Chen, and D. L. Collins. An evaluation of depth enhancing perceptual cues for vascular volume visualization in neurosurgery. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):391–403, 2014.

[25] R. E. Kirk. *Experimental Design: Procedures for the Behavioural Sciences*. Brooks/Cole, Belmont, 1969.

[26] B. Laha, D. A. Bowman, D. H. Laidlaw, and J. J. Socha. A classification of user tasks in visual analysis of volume data. In *Proceedings of IEEE SciVis*, pp. 1–8, 2015.

[27] P. M. Lambert. Three-dimensional computed tomography and anatomic replicas in surgical treatment planning. *Oral surgery, oral medicine, oral pathology*, 68(6):782–786, 1989.

[28] R. S. Laramee, C. Hansen, S. Miksch, K. Mueller, B. Preim, and C. Ware. 2D VS 3D. IEEE SciVis Conference Panel Discussion, 2014.

[29] K. Lawonn, A. Baer, P. Saalfeld, and B. Preim. Comparative evaluation of feature line techniques for shape depiction. In *Proceedings of the Workshop on Vision, Modeling, and Visualization*, pp. 31–38, 2014.

[30] F. Lindemann and T. Ropinski. About the influence of illumination models on image comprehension in direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1922–1931, 2011.
[31] M. S. Livingstone and D. H. Hubel. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *The Journal of Neuroscience*, 7(11):3416–3468, 1987.
[32] S.-J. Luo, Y. Yue, C.-K. Huang, Y.-H. Chung, S. Imai, T. Nishita, and B.-Y. Chen. Legolization: optimizing lego designs. *ACM Transactions on Graphics*, 34(6):222, 2015.
[33] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.
[34] J. W. Nam and D. F. Keefe. Spatial correlation: An interactive display of virtual gesture sculpture. In *Proceedings of the IEEE VIS Arts Program*, pp. 91–94, 2014.
[35] S. Oeltze, H. Doleisch, H. Hauser, P. Muigg, and B. Preim. Interactive visual analysis of perfusion data. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1392–1399, 2007.
[36] D. Penney, J. Chen, and D. H. Laidlaw. Effects of illumination, texture, and motion on task performance in 3d tensor-field streamtube visualizations. In *Proceedings of IEEE PacificVis*, pp. 97–104, 2012.
[37] W. A. Pike, J. Stasko, R. Chang, and T. A. O'Connell. The science of interaction. *Information Visualization*, 8(4):263–274, 2009.
[38] D. Remmler, A. Denny, A. Gosain, and S. Subichin. Role of three-dimensional computed tomography in the assessment of nasoorbitoethmoidal fractures. *Annals of Plastic Surgery*, 44(5):553–563, 2000.
[39] K. Risden, M. P. Czerwinski, T. Munzner, and D. B. Cook. An initial examination of ease of use for 2d and 3d information visualizations of web content. *International Journal of Human-Computer Studies*, 53(5):695–714, 2000.
[40] M. M. Sebrechts, J. V. Cugini, S. J. Laskowski, J. Vasilakis, and M. S. Miller. Visualization of search results: a comparative evaluation of text, 2d, and 3d interfaces. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–10, 1999.
[41] V. Šoltészová, D. Patel, and I. Viola. Chromatic shadows for improved perception. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-photorealistic Animation and Rendering*, pp. 105–116.
[42] A. H. Squillacote and J. Ahrens. *The Paraview guide*. Kitware, 2007.
[43] Y. S. Tanagho, G. L. Andriole, A. G. Paradis, K. M. Madison, G. S. Sandhu, J. E. Varela, and B. M. Benway. 2D versus 3D visualization: impact on laparoscopic proficiency using the fundamentals of laparoscopic surgery skill set. *Journal of Laparoendoscopic & Advanced Surgical Techniques*, 22(9):865–870, 2012.
[44] Q. C. Vuong, J. J. Peissig, M. C. Harrison, and M. J. Tarr. The role of surface pigmentation for recognition revealed by contrast reversal in faces and Greebles. *Vision Research*, 45(10):1213–1223, 2005.
[45] C. Weigle and D. Banks. A comparison of the perceptual benefits of linear perspective and physically-based illumination for display of dense 3d streamtubes. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1723–1730, 2008.
[46] R. R. Wilcox. New designs in analysis of variance. *Annual Review of Psychology*, 38(1):29–60, 1987.