How to Evaluate Medical Visualizations on the Example of 3D Aneurysm Surfaces

S. Glaßer¹, P. Saalfeld¹, P. Berg², N. Merten¹, B. Preim¹

¹ Department for Simulation and Graphics, Otto-von-Guericke University Magdeburg, Germany
² Department of Fluid Dynamics and Technical Flows, Otto-von-Guericke University Magdeburg, Germany

Abstract

For the evaluation of medical visualizations, a ground truth is often missing. Therefore, the evaluation of medical visualizations is often restricted to qualitative comparisons w.r.t. user preferences but neglects more objective measures such as accuracies or task completion times. In this work, we provide a pipeline with statistical tests for the evaluation of the user performance within an experimental setup. We demonstrate the adaption of the pipeline for the specific example of cerebral aneurysm surface visualization. Therefore, we developed three visualization techniques to compare the aneurysm volumes. Then, we present a single-factor, within-subject user study, which allows for the evaluation of these visualization techniques as well as the identification of the most suitable one. The evaluation includes a qualitative as well as a comprehensive quantitative analysis to determine statistically significant differences. As a result, a color-coded map surface view is identified as best suited to depict the aneurysm volume changes. The presentation of the different stages of the evaluation pipeline allows for an easy adaption to other application areas of medical visualization. As a result, we provide orientation to enrich qualitative evaluations by the presented quantitative analyses.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation G.3 Probability and Statistics Experimental Design

1. Introduction

Nowadays, many approaches for supporting the clinical expert as well as the clinical researcher regarding diagnosis or therapy involve a number of computer-supported segmentation, visualization and evaluation steps. In this paper, we explain how to evaluate comparative medical visualizations. Although many authors conduct a qualitative evaluation with medical domain experts, those results are hardly reproducible. Often, these domain experts are cooperation partners and co-authors of the presented work, where a subjective bias is hardly avoidable. Nevertheless, their medical knowledge is essential for specific application areas which may justify this procedure. Therefore, we do not neglect qualitative studies, but we want to include quantitative statistical analyses such that they can be easily adapted by other medical visualization researchers for a more comprehensive evaluation.

The segmentation of vessels with pathologic changes such as aneurysms or stenoses is an important research area. To create reproducible results as well as to reduce the work load of clinicians, automatic segmentations of vascular structures are desired. Due to patient-specific anatomies and pathologies, such automatic solutions remain challenging, and aiming for a general automatic segmentation framework is probably illusory [LABEL10]. Our application area is the visualization of cerebral aneurysms. Aneurysms bear the risk of rupture, which may cause severe consequences for the patients. For an improved intervention planning, patient-specific 3D surface models of the aneurysm and the surrounding vascular tree are extracted. They allow for the simulation of the internal blood flow [CCA⁺05, BRB⁺15] or the extraction of morphological parameters [LEBB09]. The results are included into the minimally invasive surgical plan as well as the post-processing applications within the clinical environment.

Our application scenario does not focus on the segmentation technique, but rather on the comparative visualization of different segmentation results. The employed surface meshes were extracted with a threshold-based segmentation, which can be successfully used for cerebral aneurysms [GBNP15]. However, during the segmentation process, the clinical expert requires feedback how parameters influence the segmentation results since small parameter changes may induce enormous changes of the surface mesh. To guide the clinical expert through the segmentation process, we developed three different comparative visualization techniques to show surface mesh variations.

Our quantitative and qualitative evaluation allows for the identification of the most suitable visualization technique. It comprises the visualization of five cerebral aneurysms, each approximated with three slightly different surface meshes. Our conducted user
study determines which visualization technique is best suited to evaluate the perception of small changes in the aneurysm volumes. This is especially necessary when the clinical expert or medical researcher tunes the parameters of the segmentation process and requires feedback, whether the aneurysm extent increases or not. The presented concepts comprising the experimental setup, the study design, the study procedure as well as the statistical evaluation, can be easily generalized and thus, transferred to other medical visualization application areas. Our contributions are:

- We explain which statistical test is suitable for analysis of a user study and order them into a general pipeline for the qualitative and quantitative evaluation of medical visualizations.
- We use the application scenario of cerebral aneurysms to provide three techniques $\text{Vis}_B$, $\text{Vis}_C$, and $\text{Vis}_D$ for the visualization of two similar but not identical aneurysm surface meshes, which mutually penetrate and overlap.
- Based on this example, we demonstrate how to employ the pipeline to determine which visualization technique is best suited for this application.

2. Related Work

In this section, we discuss related work for the qualitative and quantitative evaluation of visualizations with focus on the application area of aneurysms and vessels. We also refer to comparative visualizations of surface meshes extracted from medical image data.

In recent years, findings from psychophysical studies were incorporated to enhance 2D and 3D visualizations [BCFW08] influencing also the evaluation process of visualizations. For the assessment of a visualization’s suitability and performance, user studies offer a scientifically sound method [KHI*03]. Isenberg et al. [IIC*13] provide a systematic review of the evaluation practices in visualization. They employ several evaluation categories and conclude that the Qualitative Result Inspection was most often used by all reviewed papers. Further emphasis on evaluation of algorithmic performance as well as an increasing trend in the evaluation for user experience and user performance were reported.

This finding is also reflected in medical visualizations. Often, a user study is carried out, where the participants provide a subjective rating of the novel algorithm. Gasteiger et al. [GNKP10] carried out a user study for their aneurysm visualization based on the participant’s grade of satisfaction w.r.t. depth perception, spatial relationships, flow perception and surface shape. Subsequently, a more quantitative evaluation was presented by Baer et al. [BGCP11] for this visualization technique amongst two others. They conducted three controlled, task-based experiments and were able to determine statistically significant differences for the visualizations. Borkin et al. [BGP*11] also includes a formal quantitative user study to determine which visualization technique of the endotheial shear stress of coronary arteries is best suited. Hence, the experimental study provided by Diaz et al. [DRN*15] comprises a test setup to evaluate different shading techniques for volume data sets. Their evaluation included a quantitative statistical analysis as well. Also, perceptually motivated medical visualizations often include quantitative evaluations [PBC*16]. However, they focus on abstract information, e.g., depth perception, rather than comparing visualization techniques for a specific medical application area.

Visualizations of vessels are often depicted as 3D surfaces due to their complex and patient-individual shape [BFLC04, SOBP07, PO08]. Furthermore, overview visualizations are possible, e.g., the CoWRadar visualization for cerebral vessels [MMNG15]. Since we intend to employ aneurysm surface meshes for subsequent computational fluid dynamics (CFD) simulations and morphological analyses, we focus on 3D surface visualization methods. The depiction of cerebral aneurysms mostly involves the visual representation of hemodynamic parameters, e.g., scalar parameters are displayed via color-coded surface views [CSP10]. Gasteiger et al. [GNKP10] developed an illustrative visualization of aneurysms using a Fresnel shading to reveal the embedded blood flow. This work strongly motivated our visualization technique $\text{Vis}_B$.

Our comparative visualization is inspired by the image-based rendering of intersecting surfaces [BBF*11]. This technique is based on the approach by Weigle and Taylor [WT05]. Next to the integration of additional local distance cues, they enabled interactive manipulation of the surfaces. Geurts et al. [GSK*15] employed a visual comparison of medical segmentation results to allow for an evaluation of the segmentation quality. They provided additional information with landmark-based clustering to detect similar segmentation results. For the visualization itself, a color-coding of the surface was employed. There also exist illustrative approaches, e.g., the visualization presented by Carnecke et al. [CFM*13]. However, we aim at a fast comparison of cerebral aneurysm volume. Therefore, we want to reduce the visual complexity and choose the concepts provided by Busking et al. [BBF*11] as inspiration for our technique $\text{Vis}_C$.

Our visualization techniques show different segmentation results from the same patient which can be also interpreted as uncertainty visualization. Grigoryan and Rheingans [GR04] presented point-based probabilistic surfaces, which visualize surface models of medical structures such as tumors. Hence, the surface points are displaced to reflect the uncertainty at that point. The method by Pöthkow and Hege [PH11] comprises a feature-based visualization for isosurfaces with uncertainties. Their approach employs color-coding, glyphs and direct volume rendering. A taxonomy of uncertainty visualization approaches is provided by Potter et al. [PRJ12].

3. Medical Background and Image Data

Cerebral aneurysms are pathologic dilatations of the cerebral artery walls which may rupture and cause a subarachnoid hemorrhage with severe consequences for the patient. Treatment is carried out via endovascular intervention or neurosurgical clipping. However, the treatment itself may cause complications such as hemorrhages. The mortality rate associated with treatment is reported to be higher than the rupture rate of small asymptotic aneurysms [Wie03]. Thus, rupture risk assessment is an active clinical research area.

Rupture risk factors in clinical practice mainly comprise the aneurysm’s morphology as well as the type of aneurysm, i.e., asymptomatic or symptomatic [WvdSAR07]. Hence, extraction of surface meshes for aneurysms provide additional information such as the evaluation of the ostium area (i.e., the orifice between the aneurysm sac and the parent artery) [LEBB09]. Further research directions involve the simulation of the internal blood flow since
unstable and complex blood flow was correlated with increased rupture risk [CCA’05, XNT’11]. Again, a patient-specific surface mesh is the prerequisite for volume grid extraction and a subsequently CFD simulation.

For diagnosis of cerebral aneurysms, rotational angiography (RA) is considered as gold standard imaging method [GLR’09] due to the high spatial resolution. Based on RA data, the 3D digital subtraction angiography (DSA) data sets are reconstructed. To obtain the slightly similar surface meshes, we exploit the reconstruction process of the RA data from the DSA suite (Siemens Artis zeego, Siemens Healthcare GmbH, Erlangen, Germany). Five patient-specific cerebral aneurysm data sets ($D_1$-$D_5$) were reconstructed using three different kernels: Hounsfield unit (HU) smooth, HU normal and HU sharp. The five aneurysms stem from five female patients with mean age of 49 years (range 45-59 years). One cerebral aneurysm was located at the anterior communicating artery, one at the posterior communicating artery, two at the segment of the internal carotid artery, and one at the bifurcation of the middle cerebral artery. Their size varied from 2.5 mm to 11.2 mm. All patients were treated with endovascular coiling.

4. Segmentation and Comparative Visualization of Cerebral Aneurysms

In this section, the aneurysm and ostium segmentation is explained. Afterwards, the three visualization techniques $Vis_A$, $Vis_B$ and $Vis_C$ are presented.

4.1. Segmentation of Aneurysm and Ostium

For each patient’s RA data set, the three different reconstruction kernels yield three different DSA data sets. For each patient, a threshold-based segmentation was carried out for the HU normal reconstructed DSA image data. The resulting surface meshes are depicted in Figure 1. Next, the remaining reconstructions of the same patient were carried out such that they exhibit similar contours in a representative slice covering the aneurysm (see Fig. 2). Based on each threshold, the iso-surface is extracted and converted into the triangle surface mesh. Data inspection, threshold segmentation and mesh generation was carried out in MeVisLab 2.7 (MeVis Medical Solutions AG, Bremen, Germany). Hence, the segmentation was not the focus of our work and depending on the medical application, a fully automatic segmentation can be employed as well. For the purpose of our study, we required similar, but not identical aneurysm surface meshes, which could be successfully extracted with the threshold-based segmentation from different reconstructed RA data sets.

Our visualizations focus on the comparison of the volume of each aneurysm without the surrounding vessel tree. Therefore, visual separation between aneurysm and parent vessel has to be provided. The ostium was manually extracted by defining a closed cutting line along the aneurysm surface mesh using Blender 2.74 (Blender Foundation, Amsterdam, The Netherlands). This cutting line was employed twice. First, we create a closed ostium surface by triangulating the cutting line. The aneurysm surface was cut with this surface to extract the aneurysm’s volume for our evaluation. Second, the cutting line’s vertices were extruded to create a ruff-like structure, which supports the participants of our user study. Automatic ostium segmentation was not the focus of this paper, but interested readers are referred to Neugebauer et al. [NLBP13].

Figure 1: Surface meshes of five patient data sets $P_1$-$P_5$ reconstructed with the HU normal kernel are shown.

Figure 2: Segmentations of patient $P_1$. On the left, a direct volume rendering of the DSA data set is depicted. A 2D slice covering the aneurysm is shown on the right, its position is also highlighted in the 3D view. Thresholds for segmentations $S_1$-$S_3$ are selected such that similar segmentations are achieved, see bottom right. The resulting segmentation masks are color-coded.
empirically determined value of influences the visualization of possible inner structures. We use an where f serves as edge-fall-off parameter. This parameter strongly influences the visualization of possible inner structures. We use an empirically determined value of f = 0.7. The visualization technique is realized in MeVisLab using the Open Inventor vertex and fragment shader modules, where the user can directly provide shader codes as input.

4.2. Comparative Visualization Techniques

To evaluate differences of the aneurysm volume, we developed three visualization techniques: the iso-surface view VisA, the boundary-enhancing shading view VisB, and the color-coded map surface view VisC. Each technique shows two aneurysms, where the first one is referred to as ARe f, i.e., the reference aneurysm, and the second one as AComp, i.e., the aneurysm for comparison. Note that the ordering of the aneurysms is important, and employing ARe f first and AComp second yields a different visualization result than the usage of AComp first and ARe f second. In the following, the visualization techniques will be described in more detail.

4.2.1. The Iso-Surface View - VisA

The iso-surface view is a rather straightforward direct visualization of the two surface meshes of the aneurysms ARe f and AComp. It is realized in MeVisLab using the Open Inventor Library. For ARe f an orange, and for AComp a cyan transparent surface mesh is simultaneously visualized with opacity values of 0.5 (see Fig. 4). Beyond mesh extraction, no further preprocessing is required.

4.2.2. The Boundary-Enhanced View - VisB

The second visualization technique VisB (see Fig. 5) is based on the Fresnel shading approach, which was successfully employed for aneurysm visualization comprising an inner blood flow visualization [GNKP10] or the outer vessel wall revealing the color-coded inner vessel wall [GLH*14]. This technique is also referred to as ghosted view or x-ray shading. Although we do not include additional information yet, e.g., the inner blood flow, we do integrate this visualization technique in our user study since we are interested in a possible extension of the visualization with the above-mentioned information in the future.

The opacity o for each surface mesh is assigned in the fragment shader and depends on the normal n and the viewing vector v:

\[ o = 1 - (\vec{n} \cdot \vec{v})^f, \]

where f serves as edge-fall-off parameter. This parameter strongly influences the visualization of possible inner structures. We use an empirically determined value of f = 0.7. The visualization technique is realized in MeVisLab using the Open Inventor vertex and fragment shader modules, where the user can directly provide shader codes as input.

4.2.3. Map-Surface-View - VisC

In contrast to VisA and VisB, the map surface view visually provides quantitative information for the distance between ARe f and AComp. For the gathering of the distance information, the estimation of the nearest vertex pairs from ARe f and AComp is carried out. We calculate the normals of the ARe f surface mesh and approximate the distance based on the intersection with AComp. The normals of ARe f point inwards. If AComp is larger than ARe f, the intersection in negative normal direction is nearer to ARe f’s vertex than the intersection in positive normal direction and the distance value is stored as negative value.

For visual representation, we transfer the extracted distance values to the interval [0, 1] since we want to store them as texture coordinates. Therefore, we clamp the original distance values to the interval \([-0.1, 0.1]\) mm and rescale them to \([0, 1]\). Thus, texture values of 0.5 are assigned to parts where the surface meshes of ARe f and AComp have a distance of almost 0 mm. Finally, we employ the color map depicted in Figure 6 as texture and obtain VisC by using the Open Inventor Vertex Attributes module provided in MeVisLab. The color map is designed such that areas where ARe f is larger than AComp are mapped to orange, whereas the quantitative distance information is provided by the hue’s saturation. Blue areas indicate a larger local extent of AComp.

5. Comparative Study

In this section, we present our pipeline for a qualitative and quantitative evaluation. Afterwards, we describe our experimental setup and the user study in more detail.

5.1. A Pipeline for the Evaluation of Medical Visualizations

Based on the studies presented and discussed in Section 2, as well as discussions with statistical researchers, we created a generaliz-
Figure 5: Depiction of VisB. The mesh extents become best visible at the boundary of the aneurysm (see circular inlay), which requires an interactive exploration of the 3D scene. The visualization shows a larger volume of $A_{ Vis}$ at the aneurysm itself, but not at the aneurysm neck (see rectangular inlay and arrows).

Figure 6: Depiction of visualization VisC. The inlay highlights the aneurysm surface.

5.2. Experimental Setup

The whole study was realized with MeVisLab. Thus, each participant was presented with a graphical user interface (GUI), which guided the participants through the study. The user interface was created with a TabView object using hidden tabs. Each time the participant answered a question via clicking a button, the next tab was shown. At first, the TabView comprises slides for medical background information. Since all visualization techniques were implemented in MeVisLab, they could be easily integrated in the TabView GUI as well. Selection of visualization techniques and data sets for the participants was automatically carried out via Python scripts. Also, the logging of user inputs and time required for each task, i.e., the task completion time, were stored as text files.

5.3. Study Design

For the comparison of the 3D visualizations, we design a single-factor and within-subject study. The independent variable (i.e., the
Figure 7: The proposed pipeline represented as decision tree for the qualitative and quantitative evaluation.

The pseudo-randomization is listed in detail in Tab. 1, Tab. 2 and Tab. 3. For example, for the first test $T_1$ and the first question $q_1$, the user is provided with $V_{3A}$, $V_{3B}$ and $V_{3C}$ of the data sets from patient $P_1$, whereas $S_1$ is employed for the reference aneurysm $A_{Ref}$ and $S_2$ for the comparison aneurysm $A_{Comp}$. In general, for the $i$-th test $T_i$ with questions $q_1$-$q_{18}$, each visualization $V_{3A}$, $V_{3B}$ and $V_{3C}$ was shown six times in the pseudo-randomized order. The patient data $P_1$ – $P_3$ was alternated (see Tab. 2) as well as the segmentations (see Tab. 3). Since the order of the shown data sets was important, each test is repeated for switched segmentation combinations, i.e., $T_1$ is identical to $T_2$ w.r.t. visualization technique and patient but not segmentation.

The pseudo-randomization ensures that each user evaluates different data sets with varying segmentations, i.e., the user does not see the same visualization technique with the same data sets for $A_{Ref}$ and $A_{Comp}$ twice. This also holds for the demonstration of visualizations during the introduction (see Sec. 5.4), where the combinations of patient data and visualization techniques were not identical to the ones used in the test.

Next to the users’ choices regarding the aneurysm volumes, we logged the task completion time as well as the answers to the following questionnaire:

- the age,
- the sex,
- whether the user is familiar with 3D visualizations,
- whether the user is familiar with 3D medical image data,
- a rating for $V_{3A}$, $V_{3B}$ and $V_{3C}$ how much the user liked it.

The ratings were assessed with a 5-point Likert scale ranging from – (i.e. not suitable at all or not preferable at all) to +++ (i.e. very suitable or very preferable).

Table 1: Pseudo-randomization for the visualizations. For the test $T_i$ with questions $q_1$-$q_{18}$, $V_{3A}$, $V_{3B}$, and $V_{3C}$ were shown six times in the depicted order. Each test is repeated for switched segmentation combinations. After $T_{12}$, the sequence is repeated.

<table>
<thead>
<tr>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_4$</th>
<th>$T_5$</th>
<th>$T_6$</th>
<th>$T_7$</th>
<th>$T_8$</th>
<th>$T_9$</th>
<th>$T_{10}$</th>
<th>$T_{11}$</th>
<th>$T_{12}$</th>
<th>$T_{13}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{3A}$</td>
<td>$V_{3B}$</td>
<td>$V_{3C}$</td>
<td>$V_{3A}$</td>
<td>$V_{3B}$</td>
<td>$V_{3C}$</td>
<td>$V_{3A}$</td>
<td>$V_{3B}$</td>
<td>$V_{3C}$</td>
<td>$V_{3A}$</td>
<td>$V_{3B}$</td>
<td>$V_{3C}$</td>
<td>$V_{3A}$</td>
</tr>
</tbody>
</table>
5.4. Procedure

The GUI was presented to each participant, starting with a slide for the medical background information. Afterwards, the three different visualizations VisA, VisB, and VisC were shown. Each of the visualizations as well as the interaction, e.g., zooming and rotating, was explained in detail by the supervisor. The user was also encouraged to explore the scene and get familiar with the user interface for 3D exploration provided by MeVisLab. The test number 1 was assigned to the i-th user. The user had to answer 18 questions q1-q18 and decide, which aneurysm possess the larger volume. Finally, the users answered the questionnaire.

6. Results

This section describes the participants and lists the results of the user study including quantitative and descriptive analyses, based on our evaluation pipeline (recall Fig. 7). Afterwards, the qualitative subjective ratings w.r.t. suitability and preferability are discussed.

6.1. Participants

The participants were recruited from visitors of the Long Night of Sciences. During this event, scientific institutes show experiments and tests to the general public. The majority of our participants were from the university’s computer science and medical engineering departments. As a result, we were able to conduct a user study with 34 participants comprising 5 female and 29 male users, aging from 16 - 66 years. When asked if the users have experiences with medical visualizations, 10 users declined and 24 affirmed. Regarding the experience with 3D visualizations, eight users stated they have no experience. We did not include domain experts or prospective users since we were only interested in a perceptual evaluation of volume change. Hence, no medical knowledge was required.

6.2. Evaluation

The data collection provided by the conducted user study is listed in Table 4. The participants’ answers form the set of observations for VisA, VisB and VisC. We also collect the set of averaged task completion times tA, tB, and tC, each participant needed for VisA, VisB and VisC. All statistical tests were carried out with SPSS 22.0 (IBM, New York, USA). Our statistical analysis comprises three stages (recall Fig. 7):

1. We determine whether there is a significant difference between the visualizations w.r.t. accuracy.
2. In case the visualizations are significantly different, we further analyze which visualization technique is best suited w.r.t. accuracy and task completion time by pairwise comparison.
3. Finally, we provide a descriptive analysis.

6.2.1. Statistical Analysis Regarding the Accuracy

First Stage. The first analysis stage determines whether there is a significant difference between the three visualization techniques w.r.t. the amount of right answers, recall Tab. 4. Box plots for the accuracy for VisA, VisB and VisC are provided in Figure 8. Initially, we employ the Shapiro-Wilk test separately for VisA, VisB and VisC to determine whether the amount of right answers is normally distributed. Hence, the null-hypothesis H0 of the test states a normal distribution of the random variable:

\[ H_0 : The \ random \ variable \ is \ normally \ distributed. \]

The Shapiro-Wilk test yields the following significance levels:

- 0.003 for VisA,
- 0.037 for VisB,
- 0.000 for VisC.

Table 2: Pseudo-randomization for the patient data. For the test \( T_i \) with questions q1-q18, the patient data \( P_1 - P_3 \) was alternated, starting with \( P_1 \) for q1 - q3 and the segmentations \( S_1 - S_2 \) were employed. For example, \( S_1 - S_2 \) indicates segmentation \( S_1 \) for \( A_{Ref} \) and segmentation \( S_2 \) for \( A_{Comp} \). Since the order of the data set was important, the order of segmentations is reversed for odd tests.

<table>
<thead>
<tr>
<th>T1</th>
<th>S1-S2</th>
<th>S2-S1</th>
<th>S1-S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2</td>
<td>S2-S1</td>
<td>S1-S2</td>
<td>S1-S1</td>
</tr>
<tr>
<td>T3</td>
<td>S1-S2</td>
<td>S2-S1</td>
<td>S1-S1</td>
</tr>
</tbody>
</table>

Table 4: Data from the user study. For each user \( U_{1-34} \), the number of correct answers for VisA, VisB and VisC is extracted. This value ranges from 0 to 6, since each participant was confronted with each technique six times. Also, for each user the average time \( t_A, t_B \) and \( t_C \) (provided in seconds) to answer a question is collected.

<table>
<thead>
<tr>
<th>Correct answers</th>
<th>Average required time</th>
</tr>
</thead>
<tbody>
<tr>
<td>VisA</td>
<td>VisB</td>
</tr>
<tr>
<td>tA</td>
<td>tB</td>
</tr>
<tr>
<td>U1</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>U34</td>
<td>5</td>
</tr>
</tbody>
</table>

© 2016 The Author(s)
Since $H_0$ is rejected, if the significance level is smaller than 0.05, the accuracy significantly deviates from a normal distribution for each visualization technique. The next step comprises the analysis whether the visualization techniques are significantly different. We chose the *Friedmann* test, since this test provides an ANOVA for random variables that are not normally distributed. We define the hypothesis:

$$H_0: \text{All visualization techniques achieve similar results.}$$

Advantageously, the Friedmann test is based on ranks and not the actual scores. The Friedman test reveals that the accuracies significantly differ for the three visualizations, with $X^2(2) = 25.382$, $p < .05$, and the hypothesis $H_0$ must be rejected.

**Second Stage.** In the second analysis stage, we compare the visualization techniques to identify the best one w.r.t. accuracy. Based on the previous results, i.e., the amounts of right answers are not normally distributed and are significantly different, we carry out a pair-wise comparison of the visualizations. Since we deal with non-parametric data, we apply the Wilcoxon signed-rank test for $vis_A$, $vis_B$ and $vis_C$. A correction with the Bonferroni procedure [Sha95] was applied, since we carry out multiple tests on the participants’ responses. Thus, all effects are reported at a .0167 level of significance, i.e., a third of 0.05. The amount of correct answers were significantly higher for $vis_A$ ($m = 4.5$) than for $vis_B$ ($m = 3.0$), $U = -3.76$, $p < .0167$, where $m$ denotes the median. Also, the amount of correct answers was significantly higher for $vis_C$ ($m = 5.0$) than for $vis_B$ ($m = 3.0$), $U = -4.07$, $p < .0167$. However, there was no significant difference between $vis_A$ ($m = 4.5$) and $vis_C$ ($m = 5.0$), $U = 0.95$, $p = .354$. The resulting box plots for $vis_A$, $vis_B$ and $vis_C$ are provided in Figure 8.

Since $vis_B$ significantly differs from $vis_A$ and $vis_C$, we analyzed how it competes with guessing, where guessing would result in three correct answers. Hence, a Wilcoxon signed rank test yields a significant difference ($U = -2.094$, $p < .05$ with $\mu_{vis_B} < \mu_{guessing}$). Thus, $vis_B$ may systematically influence the users to provide wrong answers.

**Third Stage.** When using $vis_C$ ($\mu = 4.47$, $\sigma = 1.16$) and $vis_A$ ($\mu = 4.06$, $\sigma = 1.67$), the participants achieved a higher accuracy than

![Figure 8: Box plots of the accuracy for $vis_A$, $vis_B$ and $vis_C$ including the median m, the mean $\mu$ and the standard deviation $\sigma$.](image)

with $vis_B$ ($\mu = 2.41$, $\sigma = 1.52$). Comparison of the mean values of $vis_A$ and $vis_C$ indicates the superiority of $vis_C$.

**6.2.2. Statistical Analysis Regarding the Required Time**

For each visualization technique, the task completion time was logged. We averaged the task completion time for each question, i.e., we extract the average time $t_A$, $t_B$ and $t_C$ required by the users for a single question using $vis_A$, $vis_B$, or $vis_C$, respectively (recall Tab. 4). The boxplots are depicted in Figure 9. Similar to the previous analysis, we first determine whether there is a statistically significant difference between $t_A$, $t_B$ and $t_C$. We employ the Shapiro-Wilk test to determine whether the required times are normally distributed yielding the following significance levels:

- $0.029$ for $t_A$ and $t_B$.
- $0.007$ for $t_B$ and $t_C$.
- $0.006$ for $t_C$.

Hence, all three variables significantly deviate from a normal distribution ($p < 0.05$).

As proposed for statistical analysis of $vis_A$, $vis_B$ and $vis_C$ w.r.t. the accuracy, the second stage determines whether $t_A$, $t_B$ and $t_C$ are significantly different. Therefore, we carry out the Friedmann test, since this test provides an ANOVA for random variables that are not normally distributed. The corresponding null-hypothesis is:

$$H_0: \text{The task completion time differs for $vis_A$, $vis_B$ and $vis_C$.}$$

As a result, the Friedman test reveals no significant difference, i.e., $X^2(2) = 2.8$, and $p > 0.05$. Thus, $H_0$ cannot be rejected.

**Second Stage.** Since no statistically significant difference could be shown by the Friedmann test, we do not carry out a pairwise comparison of the task completion times.

**Third Stage.** Comparing the box plots and test statistics of $t_A$, $t_B$ and $t_C$, the participants performed the tasks in average faster with $vis_C$ ($\mu = 23.80$, $\sigma = 8.38$) compared to $vis_A$ ($\mu = 23.80$, $\sigma = 11.06$) and $vis_B$ ($\mu = 24.04$, $\sigma = 10.17$). Comparing the mean values of $t_A$ and $t_B$, the users required more time to fulfill the tasks with $vis_B$. 

![Figure 9: Box plots of the averaged task completion times $t_A$, $t_B$, and $t_C$ including the median $m$, the mean $\mu$ and the standard deviation $\sigma$.](image)

© 2016 The Author(s) Eurographics Proceedings © 2016 The Eurographics Association
6.2.3. Qualitative Evaluation of Suitability and Preferability

When analyzing the suitability and preference ratings, the same trends are reflected, see Figure 10. Furthermore, the mode value, i.e., the answer that was given most often for each question, as well as the amount of users that provide answer ++ and + is provided. Hence, users mostly rated VisC with ++ for suitability as well as preferability, VisA with + for suitability as well as preferability and VisB with − for suitability as well as preferability. The amount of users rating VisC as suitable and very suitable (i.e., answers + or ++) was highest with 27, followed by 21 for VisA and 9 for VisB. Similarly, the amount of users rating VisC as preferable and very preferable (i.e., answers + or ++) was highest with 29, followed by 16 for VisA and 11 for VisB.

Figure 10: Evaluation results of the participants regarding suitability and preferability of VisA, VisB, and VisC. The mode value, i.e., the answer that was given most often for each question, is marked. Furthermore, the sum of answers ++ and + is provided.

7. Discussion

The statistical analysis revealed a significant difference of VisA, VisB and VisC w.r.t. accuracy. The pair-wise comparison identifies VisB as poorest choice. It does not only achieve lower mean values compared to VisA and VisC, but significantly differs from both as well. VisC is not statistically significant different from VisA, however, due to the higher mean values compared to VisA, it is declared as the best visualization to compare the volume of two aneurysms. A possible conclusion might be that a derived quantity, i.e., the distance, improves the identification of the larger aneurysm. Furthermore, mean and median values of $\text{IC}$ were smaller than the values of $\text{IA}$ and $\text{IB}$. Although no significant difference occurred, these test results rate VisC as best visualization w.r.t. task completion time.

Remarkably, VisB even achieved a lower success rate than guessing. This is interpreted as indication that the users did not fully understand the design of VisB and that VisB is very inappropriate for comparison of surfaces. We assume that users wrongly interpret the ghosting view and thus, do not focus on the border areas but instead on areas with surface normals parallel to the current viewing direction. These areas are pre-dominantly color-coded in cyan, since the $A_{\text{Comp}}$ aneurysm is always drawn after the orange $A_{\text{Ref}}$ aneurysm.

When analyzing the suitability and preference ratings, we found overwhelming preference for VisA and VisC over VisB which further indicates the inappropriateness of the latter. There was also a small trend towards preferring VisC over VisA, identifying VisC as favorite visualization.

8. Conclusion

Researchers involved in medical applications are often confronted with visualization techniques, which are rather difficult to evaluate. Many times, medical visualization papers lack a quantitative evaluation at all. With our proposed user study, a pipeline was presented, which allows the comparative evaluation for three different visualization techniques for the specific application of cerebral aneurysm volume assessment. With focus on accuracy and task completion time, this concept can be easily applied to various scenarios to support qualitative findings with quantitative results.

For the evaluation of the aneurysm volume, the visualization should be reduced to basic information, i.e., no ghosted view techniques should be employed. Providing a color-coded surface visualization with quantitative distance information such as our new visualization technique VisC, helps the users to decide which aneurysm exhibits the largest volume. This was reflected by a statistically significant higher accuracy, a smaller task completion time as well as a better user rating.

For future work, different approaches can be pursued. The visualizations can be improved, for example by including depth cues such as ambient occlusion. From the statistical point of view, a systematic analysis of the influence of the volume change could be carried out. Hence, a visualization technique may be well-suited for the depiction of large volume changes, but rather improperly suited for small volume changes with a second visualization technique exhibiting the opposite behavior. Finally, we chose the employed colors to prevent false interpretations due to red-green color blindness. In future, different color blindness types should be considered and assessed with the questionnaire.

Acknowledgements We warmly thank Alexandra Baer and Maria Luz for fruitful discussions and guidance for the evaluation. We also thank the Institute for Neuroradiology, University of Magdeburg, for providing the medical image data. This work
was partially funded by the Federal Ministry of Education and Research within the Forschungscampus STIMULATE (grant number ‘13GW0095A’).

References


[BCFWF08] Bartz D., Cunningham D., Fischer J., Wallraven C.: The role of perception for computer graphics. Eurographics state-of-the-art-reports (2008), 65–86. 2


© 2016 The Author(s)