PERCEPTION-GUIDED EVALUATION OF 3D MEDICAL VISUALIZATIONS

DISSERTATION

zur Erlangung des akademischen Grades

Doktoringenieurin (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von DIPL.-ING. ALEXANDRA BAER

geb. am 17.08.1980, in Leipzig

Gutachter:

Prof. Dr. Bernhard Preim
Prof. Dr. Douglas W. Cunningham
Prof. Dr. Timo Ropinski

Magdeburg, den 16.06.2015

Dipl.-Ing. Alexandra Baer:
Perception-Guided Evaluation of 3D Medical Visualizations,
Dissertation, Otto-von-Guericke Universität Magdeburg © 16.06.2015

EHRENERKLÄRUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; verwendete fremde und eigene Quellen sind als solche kenntlich gemacht. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Ich habe insbesondere nicht wissentlich:

- Ergebnisse erfunden oder widersprüchliche Ergebnisse verschwiegen,
- statistische Verfahren absichtlich missbraucht, um Daten in ungerechtfertigter Weise zu interpretieren,
- fremde Ergebnisse oder Veröffentlichungen plagiiert,
- fremde Forschungsergebnisse verzerrt wiedergegeben.

Mir ist bekannt, dass Verstöße gegen das Urheberrecht Unterlassungs- und Schadensersatzansprüche des Urhebers sowie eine strafrechtliche Ahndung durch die Strafverfolgungsbehörden begründen kann. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, den 16.06.2015

Dipl.-Ing. Alexandra Baer

ZUSAMMENFASSUNG

Illustrative 3D Visualisierungen werden heutzutage im klinischen Alltag immer häufiger zur Präsentation patientenindividueller Bilddaten eingesetzt. Die vorliegende Doktorarbeit präsentiert Studien, die medizinische 3D Visualisierungen und dafür verwendete illustrative Techniken betrachtet. Ihr Potenzial zur effizienten Darstellung von patientenspezifischen Bilddaten und daraus abgeleitete Informationen wird im Rahmen von vier experimentellen Evaluierungen analysiert. Diese Arbeit konzentriert sich auf segmentierte 3D Strukturen aus patientenspezifischen Bilddaten, die zur Unterstützung für die Diagnostik, für die Behandlungsplanung und zur chirurgischen Ausbildung visualisiert werden. Die vorgestellten Studien basieren auf Richtlinien und Erkenntnissen psychophysischer Methoden, Theorien und Grundlagen, die Aspekte der visuellen Wahrnehmung betrachten. Wahrnehmungsbasierte Evaluierungen ermöglichen eine Analyse der Eignung und des Potenzials illustrativer Techniken für eine effiziente visuelle Exploration. Es werden etablierten Evaluierungsstrategien der visuellen Wahrnehmung analysiert und erforderliche Änderungen für die Evaluierung von 3D medizinischen Visualisierungen präsentiert und diskutiert.

Die entwickelten und durchgeführten Studien untersuchen illustrative Visualisierungen einzelner Strukturen wie Aneurysmen, Schädel oder Oberschenkelknochen und komplexe Szenarien wie Mittelohr-, Hals- und Thorax-Anatomien. Alle verwendeten Stimuli (Reize) wurden aus und basierend auf patientenspezifischen Bilddaten generiert und die Aufgaben von realen klinischen Aufgaben abgeleitet, die für die Diagnostik, die Behandlungsplanung und die intraoperative Navigation benötigt werden. Neben dem autostereoskopischen Monitor und der neuen zSpace-Technologie als zwei stereoskopische Ansichten wurden drei Hervorhebungstechniken, zwei ghosted view-Techniken und sechs Merkmalslinien Techniken betrachtet. Ihre Fähigkeit, effizient wichtige Informationen zu kommunizieren und medizinische Strukturen realistisch und ästhetisch zu veranschaulichen, wurde im Rahmen der vorliegenden Arbeit analysiert. Effizienz und Aussagekraft einer Visualisierung oder Technik wurden in Bezug auf ihr Potenzial zur Aufmerksamkeitslenkung und der Unterstützung von Form, Tiefe und räumlicher Wahrnehmung untersucht. Qualitative und quantitative Studien mit bis zu 129 Teilnehmern wurden konzipiert und durchgeführt. So wurde wahrnehmungsbasiert die Wirkung der Techniken, Visualisierungen und Ansichten durch objektive Messungen und subjektive Meinungen analysiert. Die gesammelten Ergebnisse der Probanden wurden mit Methoden der deskriptiven Statistik und Inferenzstatistik ausgewertet.

ABSTRACT

Nowadays, illustrative 3D visualizations are used more frequently in the clinical routine to present patient-specific image data. This thesis presents four experimental evaluations analyzing 3D medical visualizations and applied illustration techniques in terms of their potential to efficiently communicate patient-specific image data and derived information. This work focuses on segmented 3D structures from patient-specific image data visualized to support the diagnostic, treatment planning and surgical training process. The presented evaluations are motivated by guidelines and findings from psychophysical methods, theories and fundamentals that examine aspects of the visual perception. Perception-guided evaluations analyze the technique's suitability and potential to provide and support an efficient visual exploration. In this thesis, well-established evaluation strategies for visual perception are analyzed and required modifications to evaluate 3D medical visualizations are presented and discussed.

The designed and conducted evaluations investigate the visualization of single structures such as aneurysms, skull or femur bones as well as complex scenarios such as the middle ear, neck and thorax anatomy. Stimuli are generated from patient-specific image data and the tasks are derived from clinical tasks required for the diagnostic process, the treatment planning and the intra-operative navigation. Besides the autostereoscopic display and the new zSpace technology as two stereoscopic views, three emphasis, two ghosting and six feature line drawing techniques were investigated. Their ability to efficiently communicate essential information and to illustrate medical structures realistically and aesthetically were analyzed. Efficiency and expressiveness is evaluated in terms of the technique's attention guidance potential and the shape, depth and spatial perception support. Qualitative and quantitative evaluations with up to 129 participants were designed and conducted. Thus, the perceptual effect of techniques, visualizations and devices was analyzed with respect to objective measurements and subjective preferences. The gathered results were analyzed using descriptive and inferential statistics, respectively.

DANKSAGUNG

Als erstes möchte ich meinem Doktorvater Prof. Bernhard Preim für seine sehr gute Betreuung danken. Danke dass du mir mit deinen Anregungen und Tipps stets weitergeholfen hast, immer da warst und für alles ein offenes Ohr hattest.

Ein großer Dank geht an Douglas Cunningham, Daniel Lenz, Friederike Adler, Kerstin Kellermann, Antje Hübler sowie Mandy Scherbinsky und Maria Luz mit deren Hilfe meine wissenschaftlichen Arbeiten entstanden sind und publiziert werden konnten. Weiterhin möchte ich mich bei Prof. Dr. Karl Oldhafer, Christoph Logge, Dr. Gero Strauß, Dr. Jörg Franke, Dr. Volker Dicken, Dr. Gabor Janiga und Dr. Ulrich Vorwerk für die Bereitstellung der Datensätze und für die notwendigen medizinische Grundlagen bedanken.

Sehr dankbar bin ich für alle Kollegen der AG-Visualisierung mit denen ich zusammen gearbeitet und auch viele schöne und lustige Momente auf Konferenzen, beim Grillen, Eis essen, Bowlen, Tischtennis und täglich in der Kaffeeküche erlebt habe. Zunächst natürlich die Kolleginnen, die diese Arbeitgruppe ganz besonders bereichert haben: Jeanette und Sylvia! Ich bedanke mich sehr für die schöne gemeinsame Zeit in der Universität. Für die gute Zusammenarbeit und das tolle Büroklima danke ich Paul Klemm, Benjamin Köhler, Christian Tietjen, Konrad Mühler, Mathias Neugebauer, Tobias Mönch, Kai Lawonn und Monique Meuschke. Besonderer Dank geht an Rocco Gasteiger, der mit seiner positiven Art, Unterstützung und vielen Komplimenten meine Promotionszeit bereichert und verschönert hat. Daneben danke ich auch Patrick für die vielen interessanten Diskussionen über experimentelle Studien und deren Design sowie über Janet Siegmund und "Anatomy". Es war sehr schön und ich danke dir für eine lustige, interessante Bürozeit und für die vielen Möglichkeiten meine Nerf Trefferquote zu verbessern. Weiterhin danke ich Dr. Steffen Oeltze-Jafra für die gemeinsamen Stunden im Kreise vieler interessanter Prüfungen und Studentenbegebenheiten. Fehlen dürfen natürlich nicht die Petras aus dem Sekretariat, Thoro und Heiko sowie Steffi, die bis zum Schluss alle Rechtschreibfehler in jedem einzelnen Kapitel gesucht hat.

Zuletzt möchte ich mich bei meinem Papa, meiner Mama, Ute, meinen Schwestern sowie Katharina und Jürgen bedanken, die mich immer unterstützt und stets liebevoll um meine Kinder gekümmert haben. Ein ganz großer Dank geht an meine Omi Rotraud Schmidt. Die Unterstützung meiner Kinder Oscar und Alegra war eher praktischer Natur. Sie haben mir im Rahmen einer real-world task Studie gezeigt, das mit $p \leqslant .001$ ein statistisch signifikanter Unterschied gemessen in Zeit, Geduld und Anzahl von geschlafenen Stunden zu einer Promotion ohne Kinder besteht. Ich liebe euch dafür. Allen voran aber danke ich Michael für seine Unterstützung und Liebe, die mich immer aufgebaut und motiviert hat.

CONTENTS

I PRELIMINARIES 1

1 MOTIVATION AND CONTRIBUTIONS 3
- 1.1 Introduction . . . 3
- 1.2 Goals and Leading Questions . . . 4
- 1.3 Organization . . . 5

2 EXPERIMENTAL EVALUATIONS 7
- 2.1 Goal Definition . . . 8
  - 2.1.1 Research Questions and Hypotheses . . . 8
  - 2.1.2 Qualitative versus Quantitative Evaluation . . . 11
  - 2.1.3 Variable Identification . . . 12
- 2.2 Design . . . 13
  - 2.2.1 Tasks and Response Measurements . . . 14
  - 2.2.2 Experimental Design . . . 17
  - 2.2.3 Participants . . . 19
  - 2.2.4 Stimuli and Procedure . . . 20
- 2.3 Analysis and Result Interpretation . . . 22
  - 2.3.1 Descriptive Analysis . . . 22
  - 2.3.2 Statistical Analysis . . . 26
- 2.4 Summary . . . 30

3 VISUAL PERCEPTION AND 3D MEDICAL VISUALIZATION RESEARCH 33
- 3.1 3D Medical Visualizations . . . 34
  - 3.1.1 Medical Tasks and Questions . . . 34
  - 3.1.2 Customized Visualizations . . . 38
- 3.2 Visual Perception and Attention . . . 43
  - 3.2.1 Low- and Higher-Level Perception . . . 45
  - 3.2.2 Depth Cues and Perceptual Factors . . . 46
- 3.3 Experimental Evaluations of 3D Medical Illustrations . . . 47
  - 3.3.1 Point-Based and Line Drawing Illustrations . . . 49
  - 3.3.2 Smart Visibility Illustrations . . . 55
  - 3.3.3 Stereoscopic Views . . . 60
- 3.4 Conclusion . . . 68

II MAIN CONTRIBUTION 71

4 A QUANTITATIVE EVALUATION OF THREE EMPHASIS TECHNIQUES 73
- 4.1 Medical Background . . . 74
- 4.2 Visual Search and Signal Detection Theory . . . 74
- 4.3 Emphasis Techniques . . . 75
- 4.4 Experimental Design . . . 77
  - 4.4.1 Participants . . . 77
  - 4.4.2 Stimuli . . . 78
- 4.5 Apparatus and Procedure . . . 81
- 4.6 Analysis and Results . . . 82

4.6.1 Descriptive Analysis . . . 83
4.6.2 Statistical Analysis . . . 85
4.7 Qualitative Results . . . 86
4.8 Result Discussion . . . 87
4.9 Summary . . . 88
5 A COMPARATIVE EVALUATION OF FEATURE LINE TECHNIQUES 91
5.1 Feature Line Methods . . . 92
5.2 Experimental Design . . . 93
5.2.1 Participants . . . 93
5.2.2 Stimuli . . . 95
5.3 Apparatus and Procedure . . . 95
5.4 Qualitative Analysis and Results . . . 97
5.4.1 Frequency Distribution . . . 97
5.4.2 Schulze Method . . . 99
5.4.3 Ranking Results . . . 100
5.5 Result Discussion . . . 102
5.6 Quantitative Analysis and Results . . . 103
5.6.1 Descriptive Analysis . . . 104
5.6.2 Statistical Analysis . . . 106
5.7 Lessons Learned . . . 110
5.8 Summary . . . 111
6 GHOSTED VIEW TECHNIQUE EVALUATION 113
6.1 Medical Background and Data Flow Pipeline . . . 114
6.2 Visualization Techniques . . . 115
6.2.1 Semitransparency . . . 115
6.2.2 Ghosting . . . 116
6.2.3 Ghosting with Depth Enhancement . . . 116
6.3 Experimental Design . . . 117
6.3.1 Participants . . . 118
6.3.2 Task Methodology . . . 119
6.3.3 Stimuli . . . 120
6.4 Apparatus and Procedure . . . 123
6.5 Quantitative Analysis and Results . . . 125
6.5.1 Shape Perception Results . . . 125
6.5.2 Smart Visibility Results . . . 128
6.5.3 Spatial Perception Results . . . 130
6.6 Qualitative Results . . . 132
6.7 Result Discussion . . . 133
6.8 Lessons Learned . . . 135
6.9 Summary . . . 136
7 2D VERSUS TWO 3D DISPLAYS FOR A MEDICAL IMPLANT PLACEMENT TASK 139
7.1 Medical Background . . . 140
7.1.1 Treatment of Deafness . . . 141
7.1.2 Tympanoplastic Surgery Workflow . . . 142
7.2 Experimental Design . . . 143
7.2.1 Participants . . . 144
7.2.2 Stimuli . . . 145

7.3 Apparatus and Procedure . . . 148
7.4 Analysis and Results . . . 151
7.4.1 Depth Perception . . . 152
7.4.2 Interaction . . . 155
7.4.3 Task Completion Time . . . 157
7.5 Qualitative Results . . . 158
7.6 Result Discussion . . . 159
7.7 Lessons Learned . . . 161
7.8 Summary . . . 163

iii SUMMARY AND CONCLUSION 165
8 SUMMARY 167
8.1 Conclusion . . . 167
8.2 Future Work . . . 172

iv APPENDIX 175
A APPENDIX - CRITICAL VALUES OF DISTRIBUTIONS 177
A.1 Chi-Square Distribution . . . 177
A.2 F-Distribution . . . 178
A.3 T-Distribution . . . 179

BIBLIOGRAPHY 181

Part I

PRELIMINARIES

1

MOTIVATION AND CONTRIBUTIONS

1.1 INTRODUCTION

A visual representation of data is used to understand, explore, navigate or to visually convey complex and relevant features and relationships to the viewer. A growing research domain is the visualization and simulation of 3D patient-specific data. Medical volume data, e.g., from computer tomography (CT) or magnetic resonance imaging (MRI), and derived segmentation and simulation information, e.g., blood flow, are visualized for the diagnosis and treatment planning as well as for documentation, training and education purposes. Visualizations are generated to support the viewer in data understanding, exploration and answering therapeutic questions to make disease-specific diagnosis and select appropriate treatment options. A patient-specific data analysis enables a detailed surgery or intervention planning and therefore minimizes the injury risk and surgery stress for the patient and the physician. Furthermore, to train and explore different disease patterns, characteristics and treatment options, training systems benefit from the integration of individual anatomy compared to artificially generated datasets or illustrations in anatomic atlases. The increasing amount of 3D data, e.g., generated by scanners and simulation processes or acquired biomedical and medical datasets demands adequate techniques and visual representations to efficiently explore, analyze, and communicate essential data information.

However, the selection of *adequate techniques and visualizations*, the determination of an *efficient exploration* and the *communication of essential information* is challenging. Initially, a definition of *essential information* is required. Due to the therapeutic question or medical purpose an appropriate medical imaging modality is used. An acquired dataset should support the identification of potential pathologies and risk structures. Since a patient-specific dataset is very complex including lots of structures tightly located, a structure categorization and importance-driven identification depending on the therapeutic question is suitable. The *information communication*, in this case the structure illustration, demands *adequate illustration techniques and visualizations* to guide the viewer's attention and support the exploration process. Many illustration techniques as well as in- and output devices have been developed and refined to support the viewer and enable a *perception-guided exploration*. 3D visualization techniques have a great potential to convey the anatomy of a particular patient, to show pathologic structures realistically, and to reveal their spatial relations to adjacent risk structures. It is, however, difficult to decide which techniques or devices should be used for particular applications

and therapeutic questions, how they should be combined and how parameters should be adjusted. Perceptual theories and experimental evaluations form the basis to analyze and verify the suitability and the potential for an *efficient exploration*. Physiologists and psychologists investigate the human's perception for over 150 years. Signal and information processing and representation are determined and analyzed using experimental evaluation studies. Recently, computer scientists try to benefit from such knowledge and evaluation experiences to provide a validated perception-guided exploration of medical visualizations [17, 84, 109]. However, it is in most cases not possible to only adapt well-established experimental tasks, methods and guidelines. Complex and real-world visualizations and tasks used in computer science hamper an evaluation design, since they often violate some of the traditional experimental design assumptions [40]. Novel pitfalls arise that have to be addressed to obtain a visual perception analysis.

1.2 GOALS AND LEADING QUESTIONS

The aim of this thesis is *to design and to conduct experimental evaluations* analyzing 3D medical visualizations and illustration techniques in terms of their potential to effectively communicate essential information. This work focuses on segmented 3D patient-specific structures visualized to answer therapeutic questions for diagnostic and treatment planning purposes. Primarily, these visualizations are used for the clinical workflow as well as for training scenarios, education and documentation. This thesis presents experimental evaluations that investigate the effectiveness of emphasis techniques, the individual characteristics of two ghosted view techniques, a stereoscopic view illustration and the personal preferences. In detail, qualitative and quantitative evaluations were designed, conducted and thus, the perceptual impact of techniques, visualizations and devices was analyzed based on objective measurements and subjective preferences. All evaluations are motivated by guidelines and findings from psychophysical methods, theories and fundamentals that examine and analyze aspects of the human perception. Controlled perceptual evaluations can be performed such that the results are more general and more objective than an informative evaluation or observation. It is, however, difficult to adapt common psychophysical experiments that are primarily based on rather simple stimuli and tasks compared to complex 3D medical visualizations. The thesis analyzes well-established evaluation strategies and introduces possible adaptations and required modifications to perceptually evaluate 3D medical visualizations with patient-individual structures.

In summary, the following research questions are investigated in this thesis:

- What is an effective and expressive 3D medical visualization?
- How can psychophysical guidelines be applied to complex 3D isosurface visualizations of medical patient-specific image data to evaluate the effectiveness and expressiveness of the visualization?

- How is an appropriate evaluation for a 3D medical visualization characterized?
  - Are qualitative or quantitative evaluations necessary?
  - Which stimuli, tasks and measured parameters are necessary?

1.3 ORGANIZATION

Based on the previously mentioned motivation and leading questions, the thesis provides several results and novel contributions. The thesis is organized as follows:

Chapter 2 presents the individual steps of the design process of experimental evaluations and required decisions. In this context, guidelines and findings as well as design conventions are introduced and analyzed according to their adaptability for evaluating 3D medical visualizations. Statistical fundamentals and definitions are outlined as well as limitations and pitfalls of statistical analysis methods.

Chapter 3 presents a short overview of medical tasks and areas of application followed by the specification of the major requirements to generate customized 3D medical visualizations of patient-specific image data. Different visualization concepts and the fundamentals of visual perception including preattentive and attentive perception as well as important depth cues are presented. Perception-guided research and evaluations investigating illustration techniques and stereoscopic views generated for the exploration and presentation of 3D medical visualizations are outlined and discussed.

Chapter 4 presents a visual search task experiment to quantitatively analyze and compare the effectiveness of three illustrative emphasis techniques. The ability to guide the users' attention were investigated. The visual search and the signal detection theory were applied to clinical datasets and a therapeutic question of lymph node detection. To validate the first evaluation results regarding generalization for lymph node detection, a second experiment is presented using thorax datasets.

Chapter 5 presents and discusses the design, conduction and analysis of a comparative feature line technique evaluation. Six feature line drawing techniques are evaluated with 129 participants. Six models were used to evaluate the techniques' capability of illustrating surfaces according to realism and aesthetics. The analysis is divided into a pure qualitative analysis based on the Schulze method and an inferential statistics analysis to test the postulated hypothesis.

Chapter 6 presents three controlled task-based experiments investigating individual technique characteristics of a ghosting visualization of cerebral aneurysm anatomy with embedded flow derived from five clinical datasets. Quantitative and qualitative evaluations were performed to evaluate and compare a common semi-transparent visualization technique with a ghosted view and a ghosted view with depth enhancement technique. The design, conduction, analysis and results discussion of three studies analyzing the techniques' capability to support the shape and the spatial representation of the aneurysm models as well as the smart visibility characteristic are presented.

Chapter 7 presents a comparative experimental study investigating the effectiveness of 2D versus 3D displays for an otologic training scenario. Therefore, a training scenario for a tympanoplastic surgery was developed and used as stimuli. The design, conduction, analysis and results discussion for the evaluation and comparison of a 2D display with a glasses-free a 3D autostereoscopic display and the new 3D zSpace technology are presented.

Chapter 8 concludes the thesis results and provides recommendations. Furthermore, remaining challenges for future developments are outlined.

2

EXPERIMENTAL EVALUATIONS

Psychology is a comprehensive discipline that aims at analyzing the human mind. It can be distinguished between *research psychology*, which investigates the humans' mind by conducting experiments and the *applied psychology*, which seeks to help people with their problems. Different disciplines within the field of psychology study why people behave, think, and feel the way they do. There are several subdisciplines of psychology including social psychology, clinical psychology, occupational health, and cognitive psychology with each taking a somewhat different approach to understand and analyze the human mind. Gustav Fechner, a German psychologist born in 1801, is considered to be the founder of the experimental psychology. Fechner established the branch called *psychophysics* that investigates the relationship between physical stimuli and the sensations and perceptions they affect [50].

Psychophysical experiments and experimental evaluations, respectively are designed to investigate this functional relationship. Since one cannot directly measure perception, behavior is measured using experimental evaluations including specific tasks [40]. The properties of a stimulus are systematically varied along one or more physical dimensions to study the caused effect on a participant's behavior or perception, and thus to analyze the perceptual processes with scientific methods. Due to the effort associated with running an evaluation, it is valuable and important to design the experiment carefully. A poorly designed study leaves too much scope for further alternative result explanations, and thus leads to inaccurate or, in the worst case, to biased or useless results [53]. A well-designed evaluation isolates the causal factors and provides *valid*, *reliable* and *generalizable* results.

- **Valid** in the sense that they actually reflect what the evaluation intends to show and reliable that the results can be confirmed by other similar experiments, and thus that they are reproducible.
- **Reliable** in terms of the precision of the measurements and of the control of factors that were not intended to be studied [28].
- **Generalizable** in the sense that the findings of the performed evaluation are adaptable to other similar domains; in our case further anatomical structures (e.g. with similar shape or surface properties), surgical procedures or illustration techniques and anatomical visualizations.

Since there exist several terms in human-computer interaction (HCI) evaluation, a basic set of expressions will be used within this thesis. An *experimental evaluation*

also denoted as "user study" or simply "evaluation", "study", or "experiment" is defined as an orderly procedure carried out as a human participant research with the goal of verifying, refuting or establishing the validity of a *research question* or *hypothesis*. It generally consists of one or more user experiments (except for an entirely observational evaluation). During an user experiment, *participants* are generally asked to perform different *tasks* that are representative for the research questions or the experimental hypotheses. An experimental condition is a complete set of values for the experiment's *independent variables*, which are variables that potentially affect the task performance or subjective preference [47]. Furthermore, when designing an evaluation one of the main classifiers is whether it should be a *qualitative* or *quantitative* evaluation. Based on that, several *response measure* methods are available and an *exploratory*, *descriptive* or *inferential statistics* analysis is possible and required.

The individual steps of the design process and required experimental decisions will be introduced within this chapter. Starting in Section 2.1 with the definition of the main research aim comprising the decision of a qualitative or quantitative evaluation and the identification of required and important variables. Followed by the experimental design process introduced in Section 2.2 including the specification of appropriate tasks and potential response measure methods as well as the recruitment of participants and the stimuli generation and presentation. Finally, analysis strategies and methods will be presented in Section 2.3.

2.1 GOAL DEFINITION

Even though the definition of an evaluation is easy and well-known, it seems to be very difficult to design a proper empirical evaluation for visualization beyond time and error [28]. On the one hand, this is caused by the complexity of visualization tasks, which makes an evaluation design relevant and the analysis difficult [47]. On the other hand, the evaluation is often designed poorly, starting at the beginning with an insufficient and vague goal specification. Primarily, the major part of the design process is the goal definition, starting with the research question. Experiments vary greatly in their goal and scale, but always rely on a repeatable procedure and logical analysis of the results. The initial question heavily influences the choice of research strategies, the types of stimuli and data, the methods of result collection and analysis [100]. A poor or weak goal analysis in advance hampers the experimental design and task specification. Imprecise tasks may confuse the participants or produce non-reliable or, in the worst case, useless results.

2.1.1 *Research Questions and Hypotheses*

Experimental evaluations are performed to answer research questions and/or to verify hypotheses. Thus, finding a research question to answer is the first step when designing an experimental evaluation. Since this thesis focuses on evaluations of illustration techniques, visualizations and devices developed and used for the clinical workflow, informal or vague questions such as *"What is the difference*

between visualizations?" or *"What happens if .... ?"* are rather seldom. These question are commonly used when comparing existing techniques and the difference or the effect on the perception, respectively is not jet specified nor obvious. Generally, medical visualizations are generated to answer diagnostic or therapeutic questions or to promote the patient's education, the training process of medical students or the documentation of surgical interventions. However, for each of these target applications general or specific questions should be answered with patient-specific visualizations. Either these therapeutic or diagnostic questions are already the research questions or they establish the major requirements and thus define the research aim and indicate the required questions. Section 4 to Section 7 present four different medical areas of application and domains including various research questions and hypotheses motivated by specific medical questions of the diagnostic or therapy planning process.

The initial aim, for example, when introducing a new illustration technique or device, is to analyze the potential benefit compared to the common techniques or methods. Thus, a possible research question is usually rather general, e.g., *"Is there a difference in surface perception with the new technique or method?"* or *"Is there a difference in task performance with the new technique, method or device?"* However, if a benefit in different specific properties is assumed, and thus improvements for the structure assessment or perception, the research question is more specific, e.g., *"Does the new input device enable a more accurate implant positioning?"* or *"Does the new developed technique enable a more accurate depth perception of vessel branches?"* Generally, the recommended research question finding strategy starts with a general topic and is refined down to a particular question.

A first question should be *"Which clinical workflow or process is supported?"* followed by *"How is it supported with the new technique / visualization / device?"* Commonly, the answers of the second question are too general, e.g., *"A better depth judgment is provided."*, *"The new method is better than the common method."* or *"The structure's surface perception is more accurate."*. Now, it is necessary to define words like better, improved or more accurate (*"What is better or more accurate and how is it defined?"*). A specific definition of these describing verbs and adjectives indicates the required experimental tasks, requirements and the measured variables that enable a detailed research question verification and analysis. The more research questions are defined, the more experimental design choices including measured variables, tasks and stimuli have to be defined and potential bias variables exists. However, the questions and the tasks should be as close to the real-world as possible, but still be adaptable to a controlled experimental environment [40]. If "better" or "more accurate" are defined as measured variables, e.g., a correct implant position or distance estimation, a quantitative design is required. If "improved" or "better" are defined as subjective preferences, a qualitative design is necessary. These design choices will be introduced and explained in Section 2.1.2.

Overall, it is recommended to define the research question as specific as possible including concrete definitions of the expected difference or advantage, e.g., acceleration in terms of shorter task completion times or better in terms of a more accurate surface normal estimation.

This refinement process of the abstract high-level into a more concrete low-level description and specification is based on questions such as:

- *What is .....?*
- *What does ..... mean?*
- *How is ..... defined?*

For example, *"What is the difference?"*, *"How is interaction or perception effort defined?"*, *"How is accuracy or suitability defined?"* or *"What does facilitated structure identification mean?"*. These questions often lead to concrete expectations and assumptions, and thus to the postulation of hypotheses. A hypothesis is an assumed explanation for a phenomenon. A scientific hypothesis requires scientific methods such as statistical analysis techniques to test the proposed hypothesis. Postulated expectations in terms of hypotheses prevent a fishing for results and combine theory with empirical investigations. Additionally, hypotheses specify required tasks and actions, and therefore define and influence further evaluation design choices and variable determinations. Only when a quantitative analysis is targeted, hypotheses are required.

There are two different kinds: the *null hypothesis* and the *alternative hypothesis*, which is also called only *hypothesis* in this thesis. The null hypothesis is the premise that whatever relationship was formulated as the hypothesis for the evaluation does not occur and is due to chance alone, e.g., "*A stereoscopic display **does not** enable a more accurate depth judgment than a common 2D display.*" and "*There is **no** difference between a stereoscopic display and a common 2D display measured in depth judgment accuracy.*". Thus, this hypothesis predicts no effect or no difference and is often defined as H_0 [49]. A null hypothesis is required when the evaluation goal is to employ statistical tests on the gathered data to verify an alternative hypothesis, compare Section 2.3.2. The alternative hypothesis comprises a statement that is expected to be true instead of the null hypothesis and is denoted within this thesis as H with a defining index, e.g., H_{time} or $H_{accuracy}$ [49].

- **One-tailed hypotheses** predict a directional relationship between groups, e.g., "*A stereoscopic display enables a more accurate depth judgment than a common 2D display.*". These hypotheses specify in which direction the difference will exist. Therefore, they are also called *specific hypotheses*.
- **Two-tailed hypotheses** are *non-specific hypotheses*, since no directional relationship is predicted. Instead, a two-tailed hypothesis predicts a difference with unknown direction between groups, e.g., "*There is a difference between a stereoscopic display and a common 2D display measured in depth judgment accuracy.*".

A one-tailed hypothesis is recommended if the predicted direction is based on knowledge and experience. Since a statistical test for one-tailed hypotheses provides more power to detect an effect, it may be tempting to only postulate one-tailed hypotheses. However, a potential effect of the other untested direction (tail) will be unnoticed and remains undiscovered, and thus the consequence of missing

an effect in the other direction needs to be considered. If the expected direction is undertheorized, a two-tailed hypothesis is favored [53].

Research questions and derived hypotheses need to be defined in advance. The more questions and hypotheses are drafted, the more trials are needed to properly test each of them. Therefore, experiments are often restricted to the most important conditions.

2.1.2 *Qualitative versus Quantitative Evaluation*

In order to answer the research questions or verify the postulated hypotheses, in principle, two perception-based empirical strategies are possible:

- **Qualitative** experiments that investigate the subjective opinion, preferences, acceptance or appropriateness or
- **Quantitative** or controlled perceptual experiments that investigate the perceptual effectiveness and expressiveness by validating objectively measured experiment parameters.

Pure qualitative evaluations are observations, notes, transcripts, etc. that collect in-depth and free-form data that can generally not be measured and is used for exploratory or explanatory purposes [47, 28]. Qualitative evaluations investigate the *why* and *how* of decision making, and thus offer potential for improved understanding of existing practices by analyzing subjective preferences, experiences, perspectives and thinking processes. Since it is hard to avoid that such feedback is more than just subjective preferences, even though using standardized questionnaires, it is difficult to generalize from them. However, qualitative experiments reveal conventions in specific application areas which should be considered in designing computer-based visualization systems.

In contrast, quantitative evaluations focus on objectively measured parameters, e.g., neural activities, speed, error rate, correctness of positioning or orientation, that can be analyzed using statistical methods. The precision is relatively high in quantitative methodologies and they can be performed such that the results are more objective and some generalization to a larger population is possible. These kinds of evaluations are part of traditional scientific research, and therefore have become established. They are based on the postulation of hypotheses, the control and manipulation of independent variables in order to measure the observed effect (dependent variables), and the application of statistical analysis methods to understand the result's importance and to specify the confidence with which the results can be taken [28]. A very important property of such experiments is the reproducibility of the results.

However, in most evaluation studies qualitative methodologies are often used in conjunction with or as a part of quantitative methods, since qualitative evaluation techniques can be incorporated into all types of studies. Qualitative data is collected while performing a mainly quantitative evaluation, e.g., free-form comments or think-aloud protocols. Vice versa, quantitative measures can also be gathered in a qualitative experiment, e.g., recording interactions in a structured inter-

view or quantifying the subjective opinions by using rating or Likert scales (compare Section 2.2.1). If qualitative and quantitative measurements are combined, the results show the quantitative nature of the behavior that occurs and explain why it occurs or why not.

For medical visualizations that are generated to answer diagnostic and surgical questions, a precise and correct visualization and perception is the major goal. An exploration and perception of structures, structure relationships and simulation processes, e.g., blood flow, shall be promoted and facilitated. In principle, quantitative methodologies are recommended to investigate the evaluation conditions (e.g. illustration techniques, devices, etc.) to ensure and to analyze their accuracy. However, the subjectively perceived (qualitative) experience and personal opinions should not be neglected, since a visually pleasing illustration has a higher acceptance potential. As noted by Isenberg et al. [76], a sign in Albert Einstein's office which read *"Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted"* depicts this aspect and reminds of including qualitative research about data that cannot necessarily be measured or counted, respectively. Thus, the primary goal is a most accurate exploration with the most preferred visualizations. Such a combined evaluation might be: participants explore anatomic scenarios with respect to specific diagnostic or treatment planning tasks (distance estimation to risk structures), and besides that they are asked whether they perceive the relevant information and whether the visualization technique is considered as appropriate and easy to interpret.

In summary, for an evaluation goal definition the important questions are "*What is my research question?*" and "*What is the best method for answering/investigating my research question?*" Qualitative and quantitative research both have advantages and limitations, but they pose different kinds of questions. The chosen method should be consistent with the kinds of questions and answers the evaluation is searching for. The research question specification and the decision for a qualitative or quantitative methodology are performed simultaneously. Commonly, when defining the investigated benefit or difference during the research question specification, a qualitative or quantitative methodology is chosen and followed by the specification of the measured parameters and potential evaluation tasks. However, if hypotheses are postulated, a statistical analysis is necessary, and thus quantitative measurements are required.

2.1.3 *Variable Identification*

When the research goal is defined, the major variables have to be identified. Primarily, the independent and dependent variables need to be identified. Moreover, bias and latent or hidden variables have to be considered, since they influence the gathered results, and thus the investigated effect.

As described by Field and Hole [52], the independent and dependent variables will be obvious when appropriate research questions and hypotheses are formulated, since such scientific statements usually contain a cause and an effect. The cause is the independent variable that is studied. Within this thesis, the indepen-

dent variable is denoted as *factor* and each controlled manipulation is denoted as *level*. For example, illustration technique is the factor and each of the investigated techniques (e.g. hatching, silhouettes and ghosting) is a level of this factor. The levels may or may not affect the research statement and are investigated with the desired evaluation. The effect that needs to be measured is denoted as *dependent variable*, the perceptional and behavioral result and response, respectively. The dependent variables can either be quantitative or qualitative, based on the chosen methodology and defined research goal, e.g., accuracy or preference. Moreover, a dependent variable can be measured as interval, ordinal or nominal variable depending on the research question specification. For example, a hypothesis might be: "*A 3D user interface speeds up the exploration of a 3D anatomical visualization*". The independent variable is the 3D user interface and the measured (dependent variable) is time to determine a potential acceleration. If different user interfaces are investigated, each of them is a factor of the independent variable, and thus a controlled manipulation. Experiments that investigate more than one independent variable follow a multi-factorial design. It is recommended to limit the number of independent variables, since the collected results become enormously complicated to interpret.

Ideally, to exactly measure the caused effect, no other factor except the independent variable changes between the evaluation conditions that investigate the different levels. However, this is desirable but unfortunately not common. There are various variables called *bias variables* that represent a systematic error, which influences the effect, and thus biases the results, for example:

- Response bias: Participants try to give answers that they think the experimenter wants. This is also known as *demand characteristics* [40, 43].
- Cultural bias: Ethnicity, culture and gender strongly influence how and what people report on specific classes of opinion or belief [122, 174].
- Sampling bias: The collected data may not be accurate or represent the recruited participants, since there is a potential error that arises due to the sample selection [40, 174].
- Order effects: The effects that the order of presenting the stimuli or tasks have on the dependent variable is a bias factor, since practice effects or motivation changes occur during the evaluation and prejudice the measured results [148].

Moreover, varying and non-uniform evaluation instructions, different evaluation environments as well as a varying number of practice trials between participants or evaluation trials influence the dependent measured variables. In summary, bias factors exist for the participants, the experimental design including the stimuli design and for the experimental procedure, and it is necessary to identify, minimize and control these factors as well.

2.2 DESIGN

Lam et al. [100] stated that common experimental evaluations lack of realism, since they are mostly laboratory-based using basic visual search tasks with non-target

users. One way to ensure validity is to ensure realism in tasks, participants, data and workflow. Experimental evaluations that investigate anatomical visualizations for the clinical workflow, provide a realistic workflow including realistic patient-individual data to answer therapeutic questions. Thus, the realism aspect to ensure validity is basically existing. Now, it is necessary to adapt and to modify specific therapeutic questions and patient-individual anatomy visualizations to design an experimental evaluation that measures the defined dependent variables and enables a generalization to further anatomic and application domains.

This section introduces potential response measure methods for qualitative and quantitative experiments, discusses the two major experimental design variants and presents the basics to design appropriate tasks, stimuli and the evaluation procedure as well as recruiting participants.

2.2.1 *Tasks and Response Measurements*

The task definition step is called operationalization, where procedures are defined to measure the determined dependent variables, see Section 2.1.3. Since it is difficult to decide precisely what the participants should do, a detailed goal definition step is the major requirement. A concrete and clearly specified research goal facilitates the task definition process, including the decision what would serve as an answer. Cunningham and Wallraven [40] presented the interaction between possible research questions and answers as lying along a continuum, shown in Figure 2.1. For general questions, very flexible tasks such as free descriptions are recommended. For very specific questions, constrained tasks are recommended that are easy to interpret. However, choosing a task requires knowledge about the task's potential to answer a specific research question. Besides the task definition,

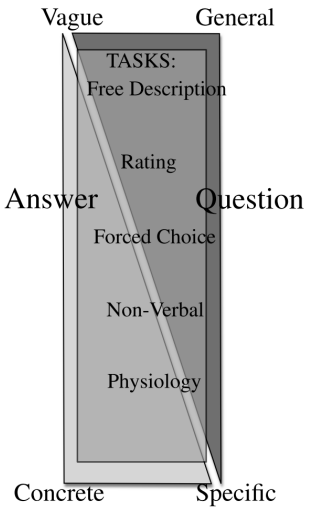
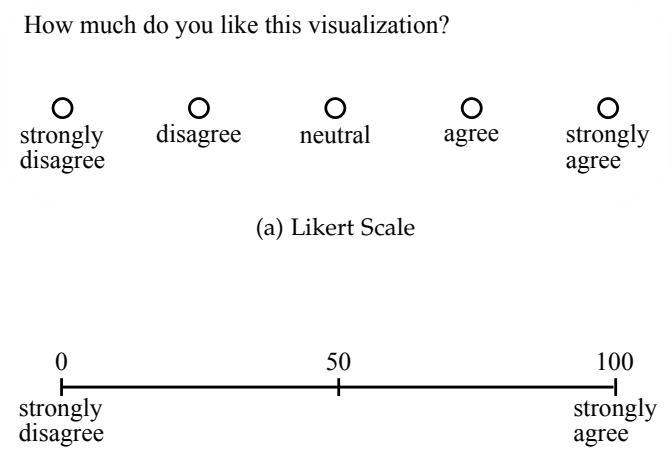


Figure 2.1: Cunningham and Wallraven [40] illustrated the correlation between question and answer from general and vague until concrete and specific. Potential tasks and response measure methods are integrated corresponding to the different levels of research questions and answers. (Republished with permission of TAYLOR & FRANCIS BOOKS, from Cunningham and Wallraven [40]; permission conveyed through Copyright Clearance Center, Inc.)

variations of answers and their analysis have to be considered at the beginning of the experiment design process. Otherwise, this may lead to an experiment and collected data that cannot be analyzed [52]. Data or results can be measured at the ordinal, ratio or interval level. Depending on the chosen design, e.g., qualitative or quantitative, different measurement methods and techniques exist.

Tasks for qualitative experiments are, for example, to describe what the participants think, believe or what they prefer. This can be realized with free descriptions, interviews and questionnaires, denoted as *meta tasks* by Cunningham and Wallraven [40]. It is suitable to use rating scales such as ranking, semantic differential or forced choice scales to obtain the participants' opinions, preferences or degree of agreement with the research statement. Field and Hole [52] denote these scales as *self-report measures*, since they rely on the subjective experience of the participants. Figure 2.2 presents an example of a 5-point bipolar Likert (semantic differential scale) and of a visual analog rating scale. It is important to choose appropriate scales and sensible labels that cover the full range of the question and thus, provide content validity. A Likert scale is a psychometric response scale, where the scale can be determined using different scale items and labeling intervals, e.g., each scale number or only the endpoints are labeled [108]. However, to provide a quantitative measurement of attitudes and an interval scale property, it is recommended to use well-established and analyzed labels, e.g, strongly agree, agree, neutral, disagree and strongly disagree, 1 to 5 and - -, -, 0, +, ++, respectively [40]. Since there are Likert scales from 3 points up to 11 points, it is recommended to determine the required qualitative division level in advance. For example, when designing the task and choosing a rating scale, one should ask: "*For this current research question, is there an important difference between a 5-point and a 7-point rating scale result?*" and "*What does this difference represent or mean for the research question?*".



(a) Likert Scale

(b) Visual-Analog Rating Scale

Figure 2.2: (a) A 5-point Likert scale with adjective labeling and a (b) visual analog rating scale (VAS) to answer the research question "*How much do you like the visualization?*". The VAS is simply a line with numerical labels. Participants can mark their response in both scales with a cross.

In practice, a 7-point bipolar Likert scale with - - -, - -, -, 0, +, ++, +++ is presented and the question is "*How much do you like the visualization?*". Is a difference between ++ and +++ (strongly agree and extremely agree) crucial or is it sufficient to divide between + and ++ (agree and strongly agree, respectively in Figure 2.2a)? Moreover, do the participants distinguish between 7 levels for this kind of question? Sometimes an even-point scale without the "neutral" option is more appropriate. Participants shall be forced to select the scale item that is most preferred without being able to select "neutral", since this is also chosen when the participant is unsure. These scales can be seen as a forced choice method.

In summary, the analysis of well-established rating scales is easier than of word-based free description, that vary widely in length and informativeness. However, it is important to formulate the research question as objective as possible to address the response bias and prevent influencing the participants' answers. Field and Hole [52] noted that content validity is achieved when representative, well distinguishable items (labels) are used. Content or logical validity refers to the extent to which the elements within a measurement procedure, such as items of a rating scale, are representative and relate to the targeted measured aspects [53]. More guidelines and recommendations for addressing the different kinds of validity (e.g. criterion, factorial, internal or external) can be found in Field and Hole [52], Cunningham and Wallraven [40] and Carpendale [28]. Moreover, it is recommended to integrate individual comments to gain more insights into the given answers or problems with choosing the desired answer.

The most common task in quantitative experiments is the *direct task* or *real-world task* [40]. Participants have to actually perform with the stimuli such as object placement or navigation tasks and their task completion time, accuracy or speed is measured meanwhile. Generally, these tasks are used for real-world questions, and therefore the surrounding environment (stimuli) has to be as close to the real-world as possible, which aggravates the evaluation design as well as the task definition and answer analysis. The task has to be as realistic as possible but still general to enable the recruitment of participants with a broad spectrum of knowledge. However, there are specific tasks that require experts as participants such as diagnosis tasks, e.g., tumor staging, and an evaluation with non-expert users as participants is not recommended due to limited relevance and generalization (external validity). The analysis of deviating answers has to be considered and the gold standard or correct answer has to be defined and determined to validate the results. When designing and choosing tasks, the following questions should be addressed:

- What is a correct answer?
- How is a correct answer analyzed?
- Is a deviating answer possible?
- Is a deviating answer totally wrong or is it still correct, for example, up to a specific percentage?

For example, the participants' task is to define a characteristic region of the aortic arc, where lots of blood flow turbulences can be found. There is more than one cor-

rect answer and the accuracy depends on the desired location or size correctness. As the individually defined regions can vary in size and location, the correctness of an answer has to be defined precisely and possibilities of correctness percentage should be considered. This example illustrates the importance of task specification and answer estimation in advance.

The most specific questions require the most concrete answers, and thus *physiological tasks* are used, compare Figure 2.1. Experiments such as electroencephalography, electromyography or even eye-tracking evaluations integrate that kind of task to analyze electrical activity and eye movement or gaze durations. The major advantage is that participants cannot affect or influence their low-level processes. However, they can affect their response, e.g., influence their gaze duration.

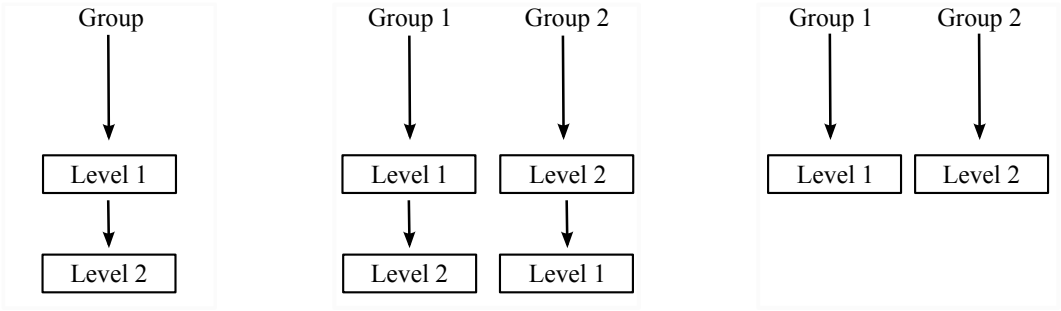
Overall, the measurements' validity has to be considered when choosing a task and response measurement technique. Even though quantitative and physiological measures are usually valid, it must be considered that, for example, participants naturally have different speeds at which they can react or some are more nervous than others, unmotivated or not enough concentrated. In detail, the goal of the tasks and measurements method is to measure what it is designed for. Otherwise, the obtained results are not due to the investigated manipulation of the independent variable but caused by other factors.

2.2.2 *Experimental Design*

Experimental design refers to how participants are allocated to the different manipulations of the independent variable in an experiment. Traditionally, in psychological experiments the participants are divided into two groups: the experimental group and the control group. A change is introduced for the experimental group but not for the control group. There are two major experimental design choices to allocate the participants to the different levels of an independent variable: a *within-participant design* and a *between-participant design*. A combination is called a *hybrid* or *mixed design* and comprises the within-participant and the between-participant design [53].

A WITHIN-PARTICIPANT DESIGN is also known as a *within-group* or *repeated measures design*. It is characterized by one participant group, where each participant is given the same kind and amount of stimuli, and therefore is exposed to each independent variable level, see Figure 2.3a. Thus, just a few participants are needed and statistical tests are more comfortable in terms of conduction and analysis because of the same participant and stimuli dimensions. Still, it is recommended to test more than 30 subjects as a reliable sample size of the population. A sample is the group of participants, explained in Section 2.2.3. Advantages of this design are economy and sensitivity [52]. Each participant performs several experiments and therefore the measured effect is not due to noise caused by the variability of using different participants for each condition compared to a between-participant design. However, order effects (carry-over effects) are a major disadvantage of this design [37]. The order of the presented conditions has an effect on the participants' behavior, and therefore biases the measured results. There

is a learning effect, since participants know what they have to do after a while and get better practiced, respectively. Their performance as well as the chance of getting fatigued or board will increase. Field [53] denoted these effects as *practice* and *boredom effects*. Thus, the measured responses are influenced and biased by the presented order of stimuli and conditions. To avoid this, the order of the presented different conditions (stimuli) has to be randomized between the participants or it has to be balanced using techniques such as matched pairs or counterbalancing. Counterbalancing means that the group of participants is split into two groups: group 1 does level 1, then level 2 and group 2 does level 2, then level 1, as seen in Figure 2.3b. Hence, order effects balance each other out in the results, since they occur equally in both groups.



(a) Within Design (b) Counterbalance Method (c) Between Design

Figure 2.3: (a) In evaluations with a within-participant design each participant is exposed to all levels of the independent variable. (b) To counterbalance order effects, the group of participants is divided and each subgroup is exposed to a different order of the presented levels. (c) A between-participant design uses separate groups of participants for each investigated level.

A BETWEEN-PARTICIPANT DESIGN is also known as a *between-group* or *independent measures design*, where two or more separate groups of participants are used, one for each level, compare Figure 2.3c. The allocation should be done randomly to ensure that each participant has the same chance of being in one group or another, and thus to maximize the internal validity and address the sampling bias introduced in Section 2.1.3. The internal validity reflects the extent to which the observed effects are due to the manipulation of the independent variable and not caused by other factors. The advantages of this design are the simplicity due to no carry-over effects, the less occurrence of practice, fatigue or motivation effects and that the evaluation duration is shorter [37]. If participants cannot participate in both conditions (e.g., an evaluation based on different ages or knowledge), this design is more appropriate, too. Indeed, a between-participant design requires more participants than a within-participant design. As most evaluations comprise on average 15 participants per group and two or more equally sized groups have to be used, the recruitment and the evaluation easily result in a time-consuming and laborious process. Additionally, this experimental design is less sensitive to the investigated conditions, and thus will less likely detect an effect, since more than one group has to be tested [174]. Ideally, nothing else than the experimental manipulation (level) and the participants should vary between the groups. Practically, the evaluations are carried out on different days, locations or other environmental

conditions. In summary, all these non-systematic variations between the groups hamper the detection of the systematic manipulation effect [52]. Additionally, the increasing number of conditions lead to more groups, and thus more parameters that have to be analyzed.

In summary, the more independent and dependent variables are evaluated, the more complex is the experimental design including required balancing methods. Overall, randomization is important for the allocation of participants and the presentation order of the investigated conditions to isolate the manipulation effect of the independent variable. The goal is to eliminate a potential bias due to the participants and the stimuli as well as the participant by stimulus interaction variance terms. To facilitate the recruitment of participants and the statistical analysis a within-participant design is recommended unless it is impossible or absolutely inappropriate, for example, due to a limited number of participants or recognition effects based on a limited number of stimuli.

2.2.3 *Participants*

A *participant group* or a *sample* is the group of people who take part in the evaluation. These people are named as *participants*. Sampling or recruitment is the process of selecting people from the population. The aim is to recruit people that are typical or representative for the target population and, thus, to ensure generalizability. A sample is biased if certain members are underrepresented or overrepresented relative to others in the population. This sampling bias is a systematic error that can influence the evaluation results such that they may not be accurate or represent the group, e.g., restricted number of participants or over-use of special participant groups. A biased sample can sometimes be identified by being very thoughtful and comparing the characteristics of respondents. For example, demographic characteristics might have an important relationship to their answers. Therefore, a carefully selected and balanced e.g., by gender, age or experience sample minimizes a potential sample bias and maximizes the internal and external validity. The external validity defines the extent to which the results of an evaluation can be generalized to other settings and participants, e.g., to other medical applications or medical experts [52].

In conclusion, controlled evaluations using anatomical visualizations and therapeutic questions, respectively such as the experiments that are presented from Section 4 to 7 exclusively require medical domain experts as participants. This requirement is very restrictive. Moreover, it neglects that these visualizations are also generated to inform patients as well as to train medical students. Thus, also participants with less to medium anatomical knowledge are required. Since the recruitment of only medical experts, medical students and patients is difficult to realize unless only a very small number of participants is sufficient, an alternative is to provide an understanding of the anatomical situations where some of these requirements can be released, for example, using non-domain expert participants. This can be done by simplifying the tasks and stimuli as well as by providing an extensive evaluation instruction and training sessions. Controlled perceptual experiments that use participants from the general population can also provide very

useful insights. Apelt et al. [3] demonstrated that controlled perceptual studies with a random sample are expressive, too. It is likely that the measured results can be applied to prospective users (e.g., medical doctors) concerning the perceptual effectiveness, even though medical doctors may achieve better results concerning the accuracy because of their clinical experience with the anatomical structures. Medical doctors, however, have to familiarize with the visualization techniques, the 3D models or used input and output devices as well as non-domain expert participants. Thus, it is not necessary to restrict participants to the narrow group of experienced medical doctors who perform a specific medical task regularly.

The required number of participants can be calculated for a quantitative evaluation using the estimated or desired effect size, which will be explained in Section 2.3.2. Contrary, there is neither an exact way of determining a sample size for a qualitative evaluation, nor a right answer in the same way a power and effect size calculation determines a sample size in quantitative research. Theoretically, a perfect sample size is defined by theoretical saturation, which means that sampling and data analysis have to be continued until no new data appear and all concepts in the theory are well-developed [145]. Morse [117] stated that the size practically depends on a number of factors, e.g., the quality of data, the scope of the study, the nature of the topic and the used qualitative method and study design.

Overall, there are rules of thumb that base on experience and gained knowledge. A controlled experiment such as the ones that will be presented in this thesis, that investigates one factor and two levels requires at least 15 to 20 participants. However, Field and Hole [52] stated that large samples give more confidence, since observations based on large samples will be relatively similar while behavior observed from small samples is likely to be rather variable. Additionally, experiments with larger sample sizes of both participants and stimuli always have greater power than experiments with smaller sample sizes [117].

2.2.4 *Stimuli and Procedure*

As mentioned in the introduction of this chapter, properties of a stimulus are systematically varied along one or more dimensions to study and analyze the caused perceptual effect. A stimulus is a specific item that causes a change or a reaction, since behavior is the result of stimulus and response. Within an evaluation a stimulus is something that is presented to the participants and their behavioral reaction or response is the caused effect, which is measured and later on analyzed. Generally, a stimulus can be everything that induces a reaction, e.g., images, light, noise or music or even electrical impulses. Traditionally, simple stimuli, e.g., single letters, simple geometrical objects, e.g., circle, triangle or square were used to evaluate low-level and higher-level visual perception. As the evaluations have evolved over the last 150 years where physiologists and psychologists have been performing experiments, the stimuli have become more individual and complex, too. For example, a stimulus to investigate visual perception can be a complex 3D object [84], a scene consisting of several 3D objects [151, 155], an immersive virtual reality [78, 98], a complex vector field [99] or even a 2D image of a facial expression that the participants have to identify [41]. In summary, the more complex or individ-

ual the stimuli such as patient-specific data or other biological datasets, the more difficult is the isolation of the caused effect. Thus, it is recommended to consider a few requirements to generate stimuli that enable a response analysis for visual perception experiments:

- Similarity: Generation of stimuli that are as similar as possible, where only the manipulation of the investigated variable differs between two stimuli.
- Simplicity: Generation of stimuli that are easy to understand, and thus facilitate the task understanding.
- Relevance: Generation of stimuli that comprise only the most relevant objects and investigated features for the task and research question, to promote the isolation of the caused effect by minimizing potential bias factors and to promote the stimuli and task understanding.
- Representative: The chosen stimuli have to be as representative as possible for the investigated domain.

The presented information shall be as highly controlled as possible such that it is relevant to the task and research question [40]. However, when using real-world tasks with realistic datasets or objects, a stimulus has to be as realistic as possible. If different stimuli categories such as degrees of difficulty can be determined, a uniform number of stimuli representing each category is recommended. Therefore, a stimulus that represents a trade-off between realism and the mentioned requirements has to be generated. The number of stimuli for each manipulation as well as the represented domain or category has to be equal to ensure an equal number of results for each category, and thus to guarantee a proper analysis and a valid results comparison.

Similarly, the stimuli presentation has to be as controlled as the stimuli generation process. Order effects and response biases are only two main issues that have to be considered and minimized when presenting stimuli, recall Section 2.2.1 and 2.2.2. Ideally, all subjects have to be tested under the same conditions to produce meaningful results and to avoid discriminations. Due to the effort associated with running an experiment, it is valuable to conduct a pilot study with a few participants. The pilot studies allow a test and refinement of the experimental design, stimuli and tasks including the provided response actions or options before starting a full-fledged study with many more participants.

Generally, an experimental procedure starts with an introduction, except, for example, behavior research experiments where no instruction is required. Moreover, demographic data, e.g., age, gender, profession, experience or knowledge and additional data that might correlate with the investigated domain such as color blindness or stereo vision deficits is recorded. To minimize influencing participants and to prevent varying introductions, an experiment and task description in written form is preferred. This initial step is usually followed by practice trials to relax and to familiarize the participants with the situation and the task as well as to practice the provided response possibilities and actions, respectively. Simple tasks and stimuli require only one practice trial to ensure that the participants understand the tasks. For more complex experiments, two or three practice trials

are recommended. The complexity and required number of practice trials can be determined during the pilot study. The stimuli presentation should be realized with an appropriate software to guarantee accurate timing as well as an accurate saving of the acquired data and error handling, e.g., missing values. The gathered data has to be uniquely assigned to the corresponding participant, while each participant and the results are saved anonymously. Available psychological toolboxes can be found in Cunningham and Wallraven [40].

In summary, the stimuli generation as well as the presentation procedure have to be as controlled and equal between the stimuli and the participants as possible to enable reproducible and valid results.

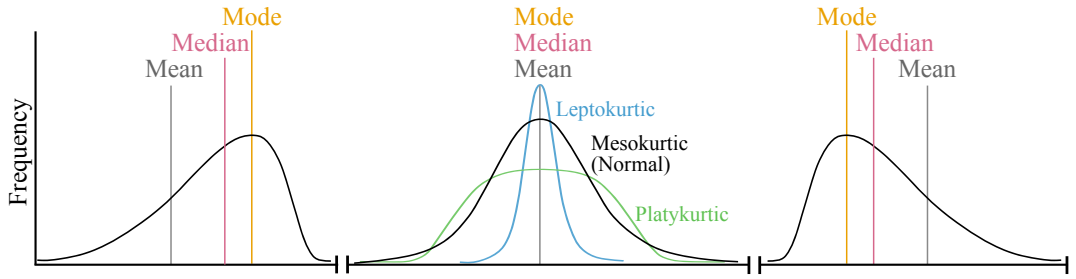
2.3 ANALYSIS AND RESULT INTERPRETATION

The measured dependent variables that represent the responses of the participants have to be summarized, analyzed and interpreted at the end of the experimental evaluation. There are two major methods: the *descriptive* and the *inferential statistics*. The former analyzes only what has been measured, and is therefore restricted by the measured data. Since researchers are generally interested in answering general questions, they are looking for general statements or rules about the population, biological systems or consumers. The measured dependent variables represent the responses of a sample drawn from an underlying statistical distribution of possible values [40]. Thus, the inferential method tries to generalize from the collected responses of a sample about the population by analyzing the data based on statistical assumptions, and thus goes beyond the measured data. Generally, the descriptive statistics is a first explorative analysis step to summarize the data and check for outliers or missing values. Quantitative evaluations and measured values require an inferential statistics to test specific postulated hypotheses and to be able to generalize from the results. Therefore, the inferential statistics is performed after a first descriptive data overview. However, a few methods such as the effect size calculation can be used before an evaluation is run to determine the appropriate number of required participants to obtain an effect with the targeted size, see Section 2.3.2.

This section gives a short overview of descriptive and inferential statistics, outlines type I and type II errors and the determination of the statistical power and effect size calculation.

2.3.1 *Descriptive Analysis*

The goal of the descriptive statistics step is to summarize, to uncover patterns or trends, to detect outliers or errors, and thus to quantitatively characterize the collected experimental results. Generally, a descriptive analysis starts with summarizing the data by calculating a summary statistics or illustrating the frequency distribution of the measured data (responses). Further on, to characterize the data, the type of distribution, the ranges (minimum and maximum values), the mean, the median, the mode and the standard deviation are determined [49].



(a) Left Skewed (b) Normal Distribution (c) Right Skewed

Figure 2.4: A frequency distribution is characterized by its shape that can be described by the skewness and kurtosis of the distribution. (a) A left skewed distribution exists when the most frequent values are clustered at the right part of the scale and the mean is less than the median. (b) The mode, median and mean values are the same when the frequencies are normally (symmetrically) distributed. The more leptokurtic the distribution's shape, the smaller the standard deviation, and thus the more accurate is the representation of the data values by the mean. (c) A right skewed distribution exists when the most frequent values are clustered at the left part of the scale and the mean is greater than the median.

A frequency distribution represents how often each value and value range, respectively occurs in the measured data. The data, therefore, is categorized and can either be displayed in a table format or as a histogram to facilitate a visual interpretation, as shown in Figure 2.5a. Ideally, the measured data values are normally distributed such that the majority of values lies around the center of the distribution, compare Figure 2.4b. Mathematically, a distribution is normal when an axis of symmetry exists [49]. A distribution is called a skewed distribution when it deviates from a normal distribution and the values' frequencies are not symmetrically but the most frequent values are clustered at one end of the scale [52]. The two major types of skewed distributions are the *left* and the *right skewed* distribution, as seen in Figure 2.4a and 2.4c. When the frequency distribution is left skewed, also referred to as *negatively skewed*, then $\bar{x} < md < mode$ and vice versa for a right (*positively*) skewed distribution ($\bar{x} > md > mode$). The type of distribution is essential for choosing an appropriate statistical analysis method, since a parametric test should be used for normally and a non-parametric test for not-normally distributed data, see Section 2.3.2.

The *mode*, the *median* and the *mean* are the three major measures that are commonly used to calculate the center of the frequency distribution [53]. The value with the highest frequency is the mode that is easy to calculate but can change dramatically if only one single value changes. Moreover, there are frequencies that exhibit two or more modes that are called *bimodal* and *multimodal frequency distributions*. Determining the middle score called median, above which 50% of the values lie, is the second method of quantifying the center of the distribution.

$$md = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n = \text{odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{if } n = \text{even} \end{cases} \tag{1}$$

The median md is the middle value when the values are ranked in order of magnitude and the number of values n is odd. If there is an even number of values, the median can be calculated as the average of the two middle values, see Equation 1. Another possibility is to chose one of the two middle values as median. A median calculation is less affected by a skewed distribution and outliers. Outliers are defined as extreme values at either end of the distribution. However, the median does not take all data values into account. In contrast, a mean calculation uses every value and is the most accurate summarizing method [52].

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

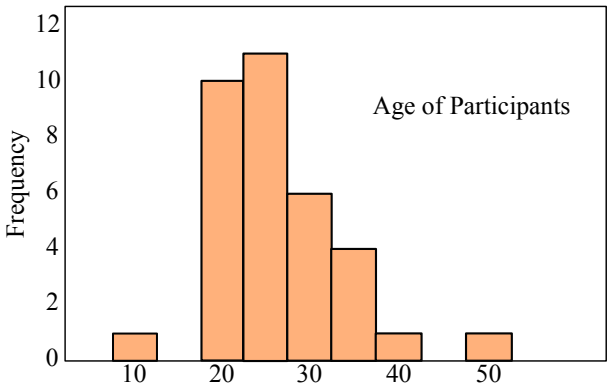
The mean $\bar{x}$ is defined as the sum of all elements x_i divided by the number of elements n (see Equation 2). However, compared to the median, the mean is sensitive to outliers and skewed distributions. If the mean is a good representation of all data values, the difference between the calculated mean value and each data value is small. To analyze how good the mean represents the measured values, the accuracy of the mean has to be calculated. The accuracy is defined by the deviation to each data value $x_i - \bar{x}$. The squared deviations between the mean and each data value are summed up (sum of squared errors) and a mean squared error called *variance* is calculated.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{3}$$

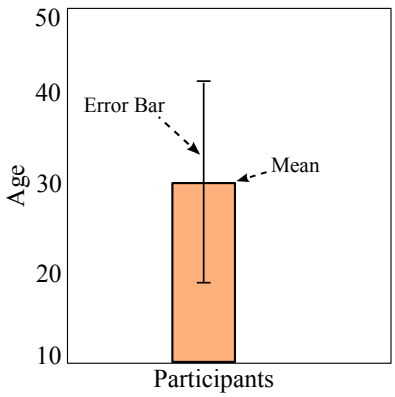
The square root of the variance is the *standard deviation* σ and represents the accuracy of the calculated mean (Equation 3). The smaller the deviation the more accurate is the representation of the data values by the calculated mean value, which is also characterized by a leptokurtic shape of the frequency distribution (blue curve in Figure 2.4b). As represented by the green curve in Figure 2.4b, a higher standard deviation exists for a platykurtic shaped distribution. Additionally, a standard error of means $\frac{\sigma}{\sqrt{n}}$ can be calculated that indicates how well the population is represented. The standard error is also known as the standard deviation of sample means over all possible samples, since this is a measure of how much variability exists between means of different samples drawn from the population [53]. Thus, this almost refers to the inferential statistics and can be seen as the transition, since the standard error is used to generalize from the collected data values of one sample about the population.

A histogram, a bar chart or any kind of boxplot are possible frequency graphs that visually communicate the summarized data and promote the identification of the above-mentioned characteristic measures. Figure 2.5 depicts each of these possibilities by taking the example of the age distribution of 33 participants. While the height of columns in a histogram represents the frequency for a specific range of values (see Figure 2.5a), the column's height in a bar chart shows the frequency for a category or investigated characteristics (see Figure 2.5b). Calculated error bars that represent, for example, the standard deviation or the standard error can be integrated as black lines on top of each bar, see Figure 2.5b. As mentioned above, even though the standard error or confidence interval refers to the inferential statistics by definition, error bars in bar charts commonly represent those measures to depict the variability of means and to provide a generalization.

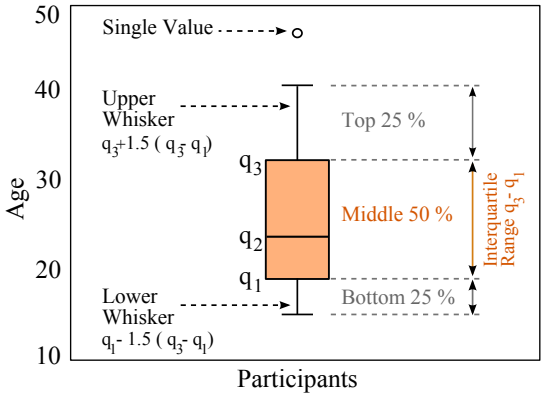
Any kind of boxplots visually represents the minimum, maximum and median value as well as the *lower* and the *upper* (75%) *quartile*, compare Figure 2.5c. Quartiles are the three values that divide the ranked measured data values into four equal groups [49]. The first q_1 or lower quartile splits off the lowest 25% of the values, and thus represents the median of the lower half. The second quartile q_2 is the median that cuts all data values in half, and the third q_3 or upper quartile splits off the highest 25% of the values (median of the upper half). If the values are normally distributed, q_1 and q_3 are symmetrically arranged around q_2. The difference between the 25% and the 75% quartile is called *interquartile range* and represents the measure of dispersion [49]. Tukey [159], who introduced the boxplot, defined values as outliers when they lie below or above the whiskers, which are defined as one and a half times the length of the box (interquartile range) from either end of the box. However, there exist several variations of boxplots. For example, the length of the whiskers is defined by the minimum and maximum data values, notched boxplots where the box narrows around the median or vio-



(a) Histogram



(b) Bar Chart



(c) Boxplot

Figure 2.5: (a) A representation of the participants' age using a histogram showing the frequency distribution. (b) A bar chart represents the calculated mean of one data category with the height of an illustrated column. The error bar's length on top of the column represents the standard deviation. (c) A boxplot also referred to as box-whisker-plot with q_1 being the upper and q_3 the lower quartile and q_2 the median value. The length of the whiskers is 1.5 of the interquartile range. Values that lie above or below are presented as circles, since they are treated as outliers.

lin plots that are a combination of a boxplot and a kernel density to additionally show the probability density at different values. For a more detailed explanation of important measurements for descriptive statistics see Field and Hole [52] or Cunningham and Wallraven [40].

2.3.2 *Statistical Analysis*

Inferential statistics provides objective criteria for testing scientific questions by following statistical procedures to infer about a larger population. Applications of inferential statistics could be testing hypotheses, correlations between variables or the validity of a model that defines measurements as a function of input variables. Since this thesis focuses on testing hypotheses, appropriate hypotheses have to be postulated in advance of an evaluation, see Section 2.1.1. An inferential statistics only determines the probability p of a hypothesis being rejected [52]. A rejected hypothesis means that it is *not probable* or *unlikely to be true* and an accepted or confirmed hypothesis means *probably* or *likely true*. It is not possible to prove that a postulated hypothesis is false or true, only evidence for the hypothesis can be provided.

Thus, there is only one set of statistical probabilities – calculation of chance effects. In detail, the tests calculate the probability that the measured results are chance results [52]. When this probability decreases, the confidence to reject the null and to accept the experimental research (alternative) hypothesis increases. Hence, instead of directly testing the alternative hypothesis H_1, the null hypothesis H_0 is tested. An inferential statistics calculates the probability that the tested samples are from the same population. The more different the sample means are, the more unlikely it is that they came from the same population. If H_0 can be rejected and extraneous factors, e.g., bias factors are under control, H_1 is likely to be true and can be accepted. The fate of the alternative or research hypothesis depends upon what happens to H_0. Therefore, the goal is to employ a statistical test on the collected data to reject the null hypothesis.

To reject the null hypothesis, the result has to be identified as being statistically significant, and thus unlikely to have occurred due to a sampling error (e.g., chance or random error) alone [49]. The thresholds for the confidence to reject the null hypothesis are typically 95% or 99%. These *confidence intervals* are limits defined such that for a certain percentage (e.g. 95% or 99%) the true mean value of the population will be within these limits. For example, only when there is a 95% confidence that the results are not a chance finding and a 5% confidence that the results are occurring by chance, H_1 is likely to be true.

The goal of the inferential statistics is to calculate a value that is directly related to the null hypothesis and decides whether or not to accept the null hypothesis. The strength of evidence in support of a null hypothesis is measured by the probability value p. A general pattern is to use either .05 or .01 as cutoff points for a confidence of 5% or 1% that the results occurred by chance. However, other cutoff points and confidence intervals are possible. The chosen cutoff point is called the *level of significance* and represents the threshold for the calculated probability value p and the level of acceptance for a significant effect when actually no difference

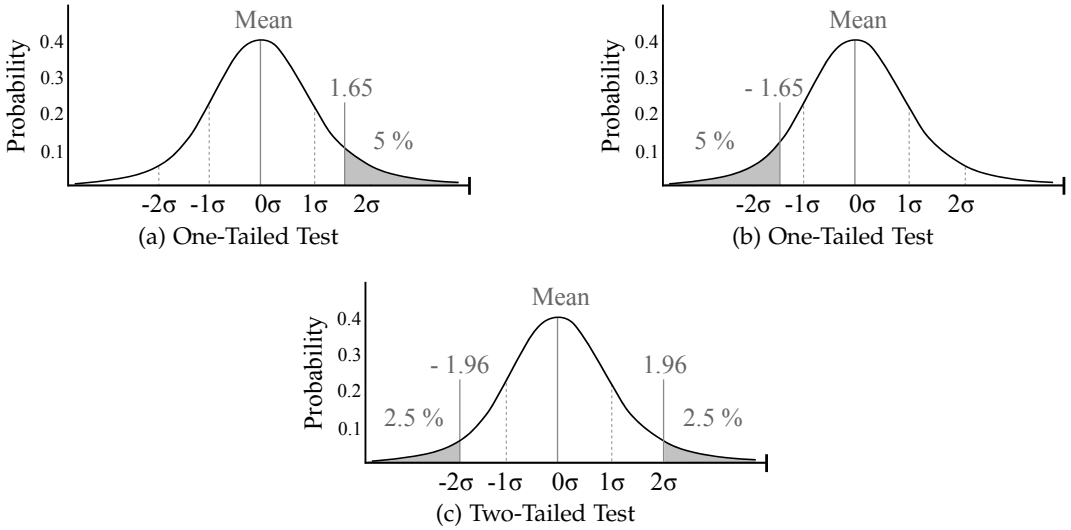


Figure 2.6: Illustration of one- and two-tailed tests for a normal distribution. The white area under the curve represents the region of acceptance and the gray area the region of rejection. One-tail testing for (a) the right and (b) the left tail. The cutoff tails represent the rejection area for the null hypothesis with each comprising 5% confidence that the results occurred due to chance. The corresponding cutoff points for a normal distribution are ± 1.65. (c) A two-tailed test cuts off 2.5% of both tails. The corresponding cutoff points for a normal distribution are ± 1.96.

exists (α level) [53]. Only when p is equal or less than the level of significance, the null hypothesis is unlikely to be true and a statistically significant effect exists at the chosen level of significance. In summary:

- $p \leqslant$ level of significance: H_0 is rejected and it is highly likely that the samples are truly different with regard to the outcome.
- $p >$ level of significance: H_0 is accepted and it is highly unlikely that the levels of the independent variable had an effect on the outcome.

A test of a one-tailed hypothesis, where the region of rejection is only on one side of the sampling distribution, is called a one-tailed test. Figure 2.6a and 2.6b illustrate a normal distribution for two one-tailed tests with .05 as level of significance for each tail according to the direction of the postulated one-tailed hypothesis, see Section 2.1.1. A test of a statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a two-tailed test and is used for testing two-tailed hypotheses. As shown in Figure 2.6c, the level of significance, therefore, is divided by the two tails to achieve the overall threshold of 5% for the region to reject the null hypothesis. However, two types of errors can result from a hypothesis test:

- **A Type I error** occurs when H_0 is rejected when actually H_0 is true. The probability of committing a Type I error is called the significance level or alpha (α) that specifies how likely the test will result in a false alarm [40].
- **A Type II error** occurs when H_0 is accepted, since the test failed to reject H_0 and actually an effect exists. The probability of committing a Type II error

is associated with the beta (β) level and relates to the power of a test which equals $1 - \beta$ [40].

Furthermore, the multiple comparisons, multiplicity or multiple testing problem occurs when one considers a set of statistical inferences simultaneously or performs multiple statistical tests with a subset of parameters (dependent variables). When a large number of statistical tests is performed, some will result in $p \leqslant .05$ purely by chance, even if all null hypotheses are really true. Several statistical techniques have been developed to prevent this, allowing significance levels for single and multiple comparisons to be directly compared. These techniques generally require a higher significance threshold for individual comparisons, so as to compensate for the number of inferences being made. The *Bonferroni correction* is one simple way to take this into account by dividing the level of significance by the number of performed tests [53]. Generally, to address both types of errors (minimize false alarms and provide a high test power) it is desirable to keep the α and the β level as low as possible [40]. This can be realized by being conservative in hypothesis testing. That means, not falsely concluding a significant difference when no effect exists (Type I) rather than missing an existing effect (Type II) [40].

TEST STATISTICS. For a statistical analysis a standardized test-specific value called test statistic is calculated. The test statistic value is used to calculate the probability value p. When the tested data exhibits a strong evidence against the assumptions of H_0, the magnitude of the test statistic becomes large and the p-value can become small enough to reject H_0 [49]. Thus, the test statistic value can also be used for hypothesis testing. If the test statistic is larger than the test-specific corresponding critical value, a statistically significant effect exists. A test statistic and corresponding critical value calculation and denotation changes from test to test based on the probability model assumed in H_0. Generally, the critical value calculation comprises the number of participants, the degrees of freedom and the test statistics. Based on the distribution of the gathered data, parametric and non-parametric statistical tests were used for the experiments analyzed within the scope of this thesis, presented from Section 4 to 7. For the applied parametric ANOVA and t-test the *F-* and *t-statistics* were determined. For not normally distributed data, χ^2 for the Friedmann test and the *z-score* for the Wilcoxon signed-rank test and the Wilcoxon-Mann-Whitney U test were calculated.

EFFECT SIZE. Only because a test statistic is significant with $p \leqslant .05$ or highly significant with $p \leqslant .01$ does not mean that the effect it measures is meaningful or important [52]. In statistics, effect size is a standardized measure of the magnitude of the observed effect. The effect size quantifies the size of the significant differences and enables the objective comparison of analysis results. However, there are several types and methods to calculate an effect size. Basically, they concentrate on group mean and variance differences or categorical variables. When two groups are compared, the most common effect sizes are the Pearson's correlation coefficient r and Cohen's d [53]. This r value is defined as a value $-1 \geqslant r \leqslant 1$ with -1 indicating a perfect negative linear relation, 1 a perfect positive linear relation and 0 no linear relation between two variables. The size of an effect is defined as

$r \geqslant |.10|$ representing a small, $r \geqslant |.30|$ a medium and $r \geqslant |.50|$ a large effect. r can be calculated in different ways depending on the applied statistical test and calculated test statistics. In this thesis, the Pearson's correlation coefficient is calculated for the pairwise comparisons using the *t-statistics* and the *z-score*.

$$r = \sqrt{\frac{t^2}{t^2 + df}} \tag{4}$$

For the *t-statistics*, r is calculated using Equation 4 where the degrees of freedom df are defined as the number of participants minus one [52].

$$r = \frac{z}{\sqrt{N}} \tag{5}$$

For the *z-score*, Equation 5 is used to calculate r. z represents the determined *z-score* value and N the number of participants.

$$\omega^2 = \frac{SS_b - (k-1)MS_w}{SS_t + MS_w} \tag{6}$$

If more than two groups or factors have to be compared, more complex methods such as omega-squared ω^2, Cohen's f or eta-squared η^2 are recommended [52]. In this thesis, ω^2 is calculated to determine the effect size when the results of a between-participants design with several levels of the dependent variables are analyzed. ω^2 is calculated as shown in Equation 6. SS_b is the squared sum of the between-groups mean, k the number of groups, MS_w the mean variance of the within factor and SS_t the total sum of squares. ω^2 values of $\geqslant .01$ represent small, $\geqslant .06$ medium and $\geqslant .14$ large effects.

Overall, the effect size is intrinsically linked to the sample size, the α level at which an effect is accepted as being statistically significant and to the power of the statistical test to detect an effect of that size. If two of these three are known, the third can be calculated. For example, the required number of participants can be calculated when the desired effect size, the α level and the power of the test are determined in advance. Field and Hole [52] stated that the power of a test should be at least .8, which corresponds to an 80% probability of not detecting the effect, otherwise more participants are required. A power analysis focuses on the effect rather than on the calculated value p. Anything that enhances the accuracy and consistency of measurement can increase the statistical power and thus the probability of rejecting a false H_0.

TEST SELECTION. Choosing the correct statistical test is difficult. Several methods exist and are hard to differentiate. Thus, decision trees were constructed to support the selection process. Generally, the number of tested sample groups, the dependency, e.g., independent or repeated measures, the type of independent variables (e.g., ordinal, nominal or interval), the number of the dependent variables and the underlying distribution are used to specify an appropriate statistical analysis test. Moreover, the frequency distribution has to be determined to chose the appropriate significance test and consequently achieve correct and valid results. Even though psychologists will assume a normal distribution if at least 30 subjects attended an experiment, it is recommended to verify the measured values [53]. If

the results are normally distributed, parametric tests can be used, otherwise it is recommended to chose non-parametric significance tests. The Shapiro-Wilk or the Kolmogorov-Smirnov test are well-established distribution tests. Since the Shapiro-Wilk test has more power to detect differences in normal distributions, this test is used for the normal distribution analysis in this thesis.

In summary, an inferential statistics is used to test the postulated hypotheses by calculating the probability that the results occurred due to chance. If this probability is very small, the postulated alternative hypotheses are highly likely and can be accepted. Besides the p-value, the individual test statistic and the effect size should be reported to illustrate the existence and the importance of the detected effect. Even if the result is statistically non-significant it can be informative to present the effect size, since it can indicate whether the non-significant findings could be due to an inadequate sample size.

2.4 SUMMARY

In psychophysics, user studies are carried out to examine aspects of human perception. Laboratory or controlled experiments try to accurately measure the human behavior and perception induced by a presented stimulus. The major part of the experimental design process is the goal definition. A clearly and detailed definition of the research questions including the specification of the concrete research goals and intentions facilitates the identification of the measured dependent variables as well as the specification of the required tasks and stimuli. The more specific the research questions or hypotheses are defined, the more obvious will be the required experimental design. However, it is recommended to restrict the evaluation to the most important questions or hypotheses to guarantee a proper test and to minimize the experimental complexity.

As it has been mentioned repeatedly within this chapter, one of the major issues for experimental evaluations is control. The goal is to isolate the causal factors to achieve valid, reliable and generalizable results. Theoretically, all relevant parameters of the experiment have to be under control, so that non-relevant and bias factors can be ruled out and no systematic error influences the experiment. Practically, it is not possible to control everything and therefore it is desirable that potential non-controlled factors are distributed randomly. Since a systematic variation of the independent variable is expected, all other possible bias factors or influences will remain unsystematic, and thus can be treated as noise. Therefore, randomization is very important, e.g., for the recruitment and allocation of participants as well as for the stimuli presentation. Moreover, to prevent extraneous factors from having an effect and thus influencing the results it is important to keep them as constant and as balanced as possible. The most frequent error in evaluation design is to compare results of differently sized groups or present a different number of stimuli. It is necessary to equalize and balance as much as possible to enable the comparison of results, e.g., a same amount of stimuli for each participant group, a same amount of stimuli representing the dependent variables and equally sized groups for each condition. Otherwise, the validity of the comparison of results is not guaranteed.

Experimental evaluations are designed and performed to analyze the relationship between physical stimuli and the behavior or perception they effect. The descriptive and inferential statistics are the two major analysis approaches that give an insight into the gathered results and enable a conclusion about the functional relationship between stimuli and perception. The descriptive analysis summarizes and presents what has been measured, while the inferential statistics generalizes from the collected results about the population. To enable a statistical analysis and thus the determination of inferential statistics, hypotheses are required. Since hypotheses have to be postulated in advance, it is necessary to define and specify the targeted result analysis at the beginning of the experimental design process during the goal definition step. If the research goal is more exploratory or qualitative, a descriptive analysis is appropriate. Assumption and hypothesis are tested using the inferential statistics by performing quantitative experiments and applying a statistical analysis. A statistically significant effect can be accepted when the null hypothesis is highly unlikely. For a statistically significant effect or difference, it is recommended to report the calculated p-value, the individual test statistics and the effect size that quantifies the magnitude of the effect.

The aim is to maximize the measurement's reliability and validity to produce generalizable results. The maximization of the measurement's reliability is realized by measuring the dependent variable as accurate as possible, which can be realized by a precise, unambiguous and objective definition of the research goal. Maximizing the measurements' validity is achieved by producing representative results that are not only valid for the specific situation in which they were obtained. Finally, a generalization has to be confirmed across participants, experimental design, methods, apparatus, stimuli and experimental environments to enable an inferential statistics to conclude about a larger population and further application domains.

3

VISUAL PERCEPTION AND 3D MEDICAL VISUALIZATION RESEARCH

Nowadays, computer assistance for the clinical workflow is an important aspect and thus a growing research domain. Starting with computer-aided diagnosis, surgery planning, navigation during the surgery through to training and learning systems. Computer-aided surgery planning and navigation systems are especially useful for difficult cases, e.g., oral and maxillofacial surgery, plastic surgery, radiation therapy, orthopedic surgery or the removal of elusive tumors in the head or neck region. Besides 2D visualizations, patient-specific 3D visualizations and derived segmentation information are integrated to promote the diagnosis and therapy planning as well as the intra-operative navigation workflow [126]. These patient-specific visualizations enable a customized therapy planning and intra-operative navigation that maximizes the therapy success and at the same time minimizes the injury risk and thus lowers the complication rates as well as the surgery stress for patient and surgeons. Since individual medical visualizations enable an extended and more realistic exploration of different disease patterns, characteristics and treatment options compared to artificially generated datasets or illustrations from anatomic atlases, medical training systems benefit from a patient-specific dataset illustration as well.

The demand for appropriate representations has grown considerably during the last decade, due to the increasing amount of 3D medical datasets and possibilities to interactively visualize large and highly detailed volumetric data by the acceleration opportunities of modern GPUs. Visualization techniques have a great potential to convey and to show the anatomy and pathologic structures realistically, and to reveal their spatial relations to adjacent risk structures. Thus, the development of new and the improvement of existing techniques is a huge research domain. Besides that, in- and output devices such as haptic devices, stereoscopic and holographic displays are developed that integrate stereoscopic cues and motion parallax as the most significant sources of depth information to enable more intuitive and realistic 3D visualizations, navigations and interactions. However, it is difficult to decide which techniques should be used, how they should be combined or how parameters should be adjusted to generate medical illustrations that effectively communicate the required information. Visual perception and attention plays an important role in the area of visualization, since an understanding of perception can significantly improve the quality and the quantity of the illustrated information [172]. An understanding of visual perception and attention is essential to guide the user's attention and enable an efficient information transfer. Thus, perceptual theories and experimental evaluations are used to analyze and verify

the suitability and the potential of medical visualizations, techniques and devices for an effective information communication and exploration.

This chapter starts with a short overview of medical tasks and areas of application in Section 3.1 followed by a description of the major requirements to generate customized 3D medical visualizations. Section 3.2 comprises fundamentals of visual perception including preattentive and attentive perception as well as important depth cues that enable humans to perceive depth and spatial relations based on 2D images that were initially projected onto the retina. Perception-guided research and evaluations investigating illustration techniques and stereoscopic views generated for the exploration and presentation of 3D medical visualizations are introduced and discussed in Section 3.3.

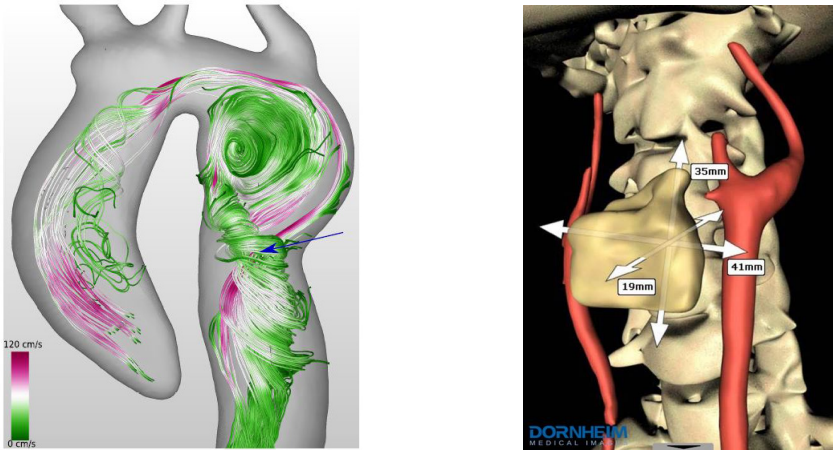
3.1 3D MEDICAL VISUALIZATIONS

A visual representation of datasets derived from measurements or simulations of real-world phenomena is called *scientific visualizations* [126]. Generally, a visualization can be defined as a visual representation that is generated to display, explore or analyze the underlying data. An important visualization goal is to generate a mental image and promote the human perception of the illustrated data to support the exploration process and guide the viewer's attention. *Medical visualizations* are visual representations of medical image data, e.g., CT, MRI or PET volume as well as X-ray or ultrasound data. Besides 2D visualizations, patient-specific 3D visualizations of segmented structures and derived segmentation information are visualized using computer graphics techniques. Pathologies and other anatomical structures such as risk or surrounding structures are segmented to provide morphological and quantitative information (e.g. distances, volumes or access paths). To promote the data representation, exploration and analysis, further information that are recorded by medical imaging techniques, e.g., functional information (fMRI), information related to metabolism (PET/SPECT), and blood flow (Phase-Contrast Angiography) can be integrated as well [126].

3.1.1 *Medical Tasks and Questions*

There are several different clinical and medical application areas that integrate patient-specific image data and require appropriate visualization techniques and representations to facilitate, accelerate and improve the accuracy and visual perception of the illustrated medical structures and represented information. Preim and Botha [126] refer to *diagnosis*, *treatment planning*, *intra-operative support*, *documentation* and *educational purposes* as the main areas of applications for medical visualizations to explore and analyze medical datasets.

In the clinical routine, patient-specific 2D or 3D image data are acquired as a part of the diagnostic process. Diagnosis comprises the identification and determination of diseases or injuries by the examinations of symptoms, the evaluation of patient history (anamnesis) and the review of laboratory and image data. Image data are acquired either with a directed hypothesis to be tested or without when

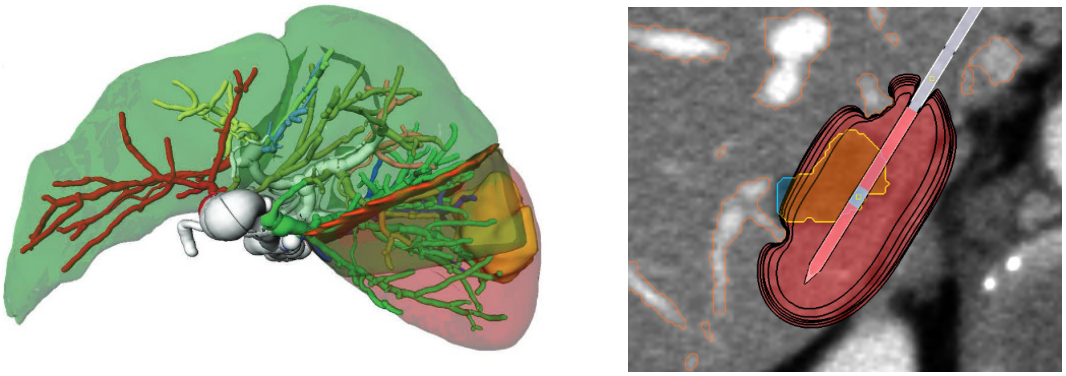


(a) (b)

Figure 3.1: (a) An illustrative visualization of steady flow features occurring in 4D MRI data of the aorta. (Image reprinted from Born et al. [18] © 2013 IEEE with kind permission from IEEE.) (b) 3D neck visualization for the tumor assessment and determination of extensions. (©Dornheim Medical Images)

a physician cannot define the disease on the symptoms or clinical examinations [126]. The imaging modality is selected based on the patient's symptoms and the diagnostic process, e.g. a suspected diagnosis, a diagnosis of exclusion or at least a differential diagnosis. Detailed 2D and 3D visualizations of the original data and derived information promote the visual disease assessment including severity determination. Additional important information for the diagnosis are, for example, quantitative measures, e.g., distances or extensions or time-varying parameters, e.g., blood flow for the analysis of flow patterns (see Figure 3.1a) However, 3D visualizations are rarely used for the daily diagnostic process, since radiologists are trained and used to 2D images and they infer depth information and spatial relations from these cross-sectional images. Thus, interactive 3D medical visualizations are generated for rather complex, rare or very unusual anatomic variations, diseases or injuries, e.g., brain tumors, elusive tumor resection or complex fractures. The goal is to provide an overview of the situation and to present the diagnosis results to the responsible medical expert for the treatment planning process. Furthermore, quantitative image analysis, such as distance or volume determinations of structures, can be provided by interactive 3D visualizations of segmented structures, as shown in Figure 3.1b. This enables, for example, specifying the stage and severity of a tumor disease and thus supports the diagnostic process [126]. Visualizations generated for the diagnostic process combined with the diagnosis report illustrate all anatomical and pathological structures that are required for the treatment planning process.

The treatment planning also denoted as *therapy planning* is the planning process of the appropriate treatment decision and intervention strategy, e.g., surgical interventions, radiation treatment or minimally invasive interventions including the estimation of the treatment success, result and corresponding risks. Besides the localization and spatial understanding of pathological and anatomical structures, a potential risk for life-critical structures has to be identified and assessed as well as

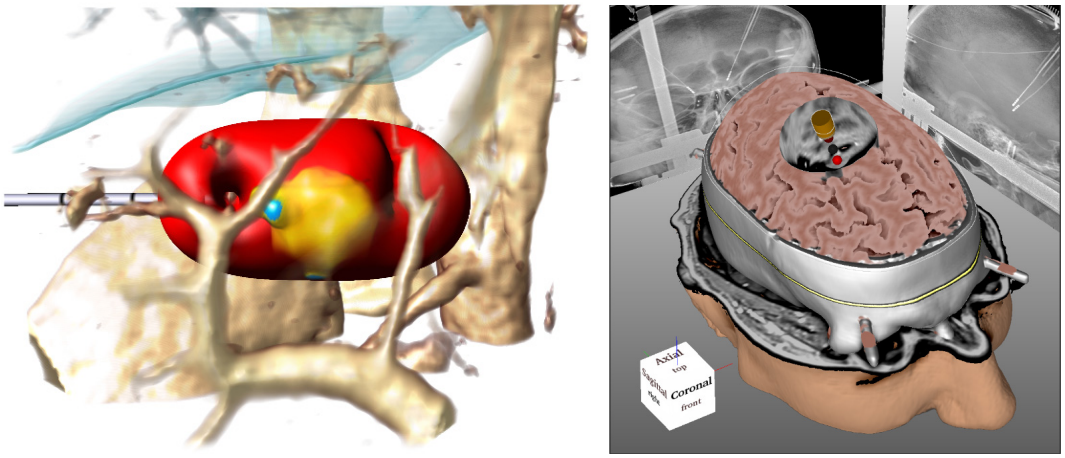


(a) (b)

Figure 3.2: (a) A 3D visualization of a liver tumor resection proposal. The tumor is located in the liver. The remaining part of the liver is green, the resection plane is orange, and the resection volume is red colored. (b) A 2D visualization of a planned radiofrequency ablation including the segmented tumor and the 2D image data as well as an applicator and approximated ablation zone visualization. (Images reprinted from (a) Birr et al. [15] and (b) Rieder et al. [136] © IEEE 2013 and 2011 with kind permission from IEEE.)

appropriate access paths for biopsies, regional anesthesia or tumor ablation [126]. Treatment planning tasks such as access, resection, radiation dose distribution or implant planning are guided by therapeutic questions that require an accurate 2D or 3D visualization of the patient-specific image data and derived information, see Figure 3.2. Therapeutic questions, e.g., "*Does the tumor infiltrate surrounding structures?*", "*Is the vortex located near the vessel wall?*", "*How many enlarged lymph nodes exist?*" or "*Is a neighbored structure affected by the radiation dose?*" are questions that help to explore the individual medical situation, to specify and to support the treatment decision. Moreover, the preoperative estimation of the intervention's extent as well as the patient's operability is supported, e.g., the determination of the remnant liver volume (see Figure 3.2a). A 2D visualization of an ablation planning process is shown in Figure 3.2b, where the patient's image data is overlaid with an applicator, the segmented tumor and the potential heat expansion to estimate the treatment result and to identify potential structures at risk. An increasing preoperative preparation and familiarization with the patient-specific data supports the minimization of injury risks and surgical stress for the patient and the surgeon. Interactive 3D visualizations are generated to visualize all relevant structures that support answering the therapeutic questions and thus the treatment planning process. The visualization goal is to improve the spatial perception and structure assessment, e.g., form, extent, location and distances to surrounding structures by illustrating 3D structures and thus enabling a 3D exploration and analysis, e.g., distance and structure measurement.

The integration of 2D and 3D visualizations of patient-specific image data acquired pre- and intra-operatively in the operating room is a growing area of application. Since the treatment planning is based on preoperatively acquired image data, the accuracy of interventions and performed treatment procedures may be improved when correlating and verifying, for example, catheter or applicator positions (see Figure 3.3a), properties and parameters such as quantitative measure-



(a) (b)

Figure 3.3: Two intra-operative 3D visualizations (a) one for a radiofrequency ablation to control the applicator position and combined with the ablation zone (b) and one for a deep brain stimulation surgery that comprises a 3D reconstruction of the skull, a volume visualization of the brain, the CT dataset and intra-operatively acquired X-ray scans. (Images reprinted from (a) Rieder et al. [136] and (b) Bock et al. [16] © IEEE 2011 and 2013 with kind permission from IEEE.)

ments that were preoperatively determined with intra-operatively updated image data. For example, during interventions on the brain (neurosurgery), the location of the brain tissue and, combined with that, structures in the brain move after opening the skull. Thus, previously planned access paths or resection planes have to be updated to minimize damage to the risk structures. Moreover, medical visualizations are used for an instant control of the currently performed procedure, e.g., resection control or navigation support. For example, Figure 3.3b presents an intra-operative 3D visualization from Bock et al. [16], who combined different image modalities to verify the positions of the electrodes and to achieve a most accurate electrode placement by protecting risk structures and brain regions, respectively during a deep brain stimulation surgery.

Moreover, medical visualizations are used for interdisciplinary discussions, and extended with labels and annotations they are suitable for documentation reports or educational purposes, e.g., patient education or training for medical students. Planned interventions can be introduced to the patients with the help of 3D visualizations of their own image data and thus improve the understanding of the individual disease and treatment decision. Originally, anatomical atlases were used for educational purposes to describe the anatomy, to present diseases and to explain intervention strategies. Nowadays, interactive 3D visualizations integrated in education systems provide the possibility to learn interactively. Students may select individual focus structures for a detailed exploration or check their knowledge by labeling different structures on their own. Patient-individual image data visualizations increase the variety of presented anatomy, disease patterns and characteristics.

3.1.2 *Customized Visualizations*

Since computer assistance is primarily used for difficult cases, an improved and accurate anatomy comprehension is essential. Some visualizations are generated to test specific directed hypotheses, to answer therapeutic questions or to present results. Others are generated to explore image data in terms of an undirected search without specific hypotheses. However, each visualization is generated for a specific purpose. Thus, the goal of medical visualizations is to transfer the patient-individual image data into simple visualizations supporting the exploration, interpretation and decision making.

Initially, 3D visualizations of the full dataset must be generated based on the acquired 3D volume date consisting of 2D images. Therefore, individual voxels of the dataset are selected, weighted, combined and projected onto the image plane [125]. Primarily, there are two possibilities: the *direct* and the *indirect volume rendering* to generate 3D visualizations. Although there are several visualization techniques and developments for direct volume rendering for medical visualizations, this thesis focuses on indirect volume rendering and, thus, whenever the term 3D visualization is mentioned in this thesis it describes a 3D indirect volume rendering visualization. In contrast to the direct volume rendering, where 3D visualizations are generated based on transparency and coloration of all voxels of the volume data, indirect volume rendering techniques explicitly extract subsets of the data for a geometric surface visualization also referred to as *isosurface visualization* of 3D structures. Thus, relevant structures must be segmented, polygonal meshes with edge points and normal vectors are generated and determined and further methods, e.g., smoothing are applied to reduce artifacts such as staircases to finally produce a 3D surface rendering of medical volume data used for one of the above-mentioned areas of application.

Besides the required high performance and the processing of different data types for the generation of a 3D visualization, the final illustration should follow a few requirements to enable an efficient data exploration and effective information transfer:

- **Accuracy:** The structures that are segmented from the 2D image data, processed and 3D visualized have to be as accurate as possible to ensure a realistic visualization and data representation. When quantitative values are determined based on a generated visualization, accuracy is the major prerequisite. A 3D visualization must represent the segmented structures such that the distances between the original and illustrated surface of the structure are as low as possible.

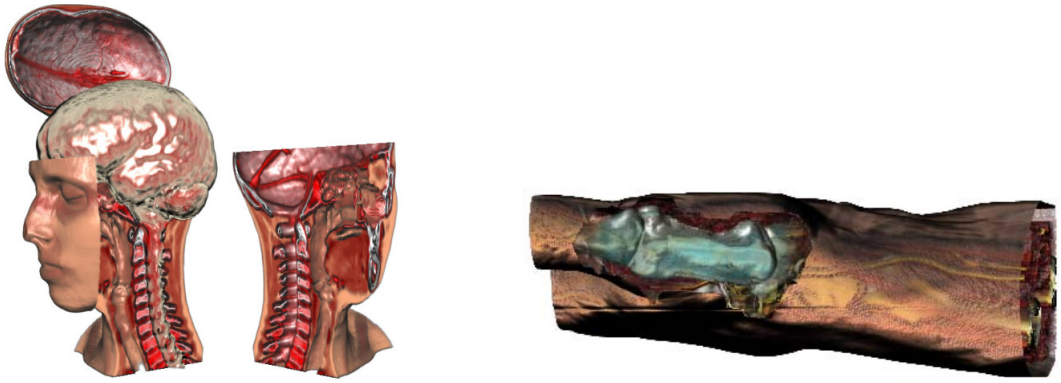
- **Application Specificity:** Since each of the areas of application requires different visualizations, this should be considered and realized during the generation process. A visualization for diagnostic purposes must show all segmented and relevant structures unfiltered to provide all information to promote the decision support. Since the diseases have to be characterized and classified, all acquired information may contribute to the diagnosis. Contrary, a treatment planning visualization is more question-driven and spe-

cific. Several questions have to be answered until the appropriate treatment is planned. Therefore, customized visualizations are required that facilitate and accelerate this question-driven process. Similar to treatment planning visualizations, documentations and training applications require question- or task-specific visualizations that document and enable to train the specific interventions, respectively. Educational visualizations combine the requirements of the previously mentioned visualizations, since they may either be more general or question-specific, depending on the field of education, e.g., domain-, disease- or intervention-specific or an anatomic overview.

- **Effectiveness:** A 3D visualization should promote an efficient exploration and effective communication of the essential information, respectively. The viewer's attention shall be guided to important structures and structure relations without being disturbed or prevented by full structure occlusions. Moreover, the perception of the illustrated structures and information must be supported. Since a patient-specific dataset is very complex, including lots of structures tightly located, an application-specific categorization of important structures and thus appropriate illustration and presentation are required.

There are several possibilities to generate an application-specific visualization of patient-specific data. Primarily, the visualizations can be divided into: a visualization of all segmented structures and a reduced visualization based on a previous question- and thus importance-driven structure selection. To generate an effective visualization, each of these presentation options include either a modified and application-specific structure presentation, e.g., defined viewpoints or exploded views according to the specific question or application, or an appropriate structure illustration, e.g., ghosting views, silhouettes, shadows or varying illumination.

When all segmented structures and derived information are visualized, a common method to guide the viewer's attention is an adapted structure presentation to reveal occluded objects and focus on the specific task or question the visualization is targeting. Bruckner and Gröller [24] solved the structure occlusion problem by their presented *exploded view technique*, where an object is partitioned into several segments and these segments are shifted or displaced to reveal the important and otherwise occluded structures (see Figure 3.4a). Thus, this technique changes the structure presentation to generate an application-specific visualization with chosen focus structures or preferred views. Another approach is an automatic viewpoint determination combined with the camera animation. Viola et al. [163] presented an approach, where the user selects a focus object and important areas in object and image space are then defined automatically and optimal viewpoints for each scene object are determined in a pre-processing step. Their approach is applied to direct volume visualization and they integrated an automatic adaption of visualized structures to enable an unoccluded view to the focus structure denoted as *ghosted view technique*, as shown in Figure 3.4b for a visualization of a thumb bone. Since these viewpoints are only estimated by their quality for the whole scene and not for single objects, Mühler et al. [121] presented an advanced automatic viewpoint selection approach for single and multiple objects. The optimal viewpoint is guided by parameters such as importance, visible surface, pre-



(a) Exploded View (b) Ghosted View

Figure 3.4: (a) An exploded view visualization of the head, where the brain, a part of the spine and the nerve-canal is exposed. (b) A visualization of a hand anatomy with an adapted viewpoint to inspect the thump bone. Since the bone would be occluded by the skin, the transparency values of the occluding voxels are adapted and the bone is visible. This technique is called a ghosting view technique. (Images reprinted from (a) Bruckner and Gröller [24] and (b) Viola et al. [163] © IEEE 2006 with kind permission from IEEE.)

ferred region and viewpoint stability, see Figure 3.5. However, their parameters cannot directly be mapped to the visualization goals and thus the viewer is not able to integrate their subjectively preferred viewpoint. Moreover, an appropriate parameter selection for different anatomical domains was not realized. To accelerate the process of surgical planning, Mühler and Preim [120] improved their viewpoint technique and introduced a method that enables the reuse of 3D medical visualizations and 2D slice views. Keystates were introduced as a concept to describe the state of a visualization in a general manner and thus to provide the reuse of once designed visualizations for similar cases and interventions. Several more automatic view point determination approaches exist [83, 85, 116]. Besides viewpoint selection methods, exploration and navigation techniques for complex

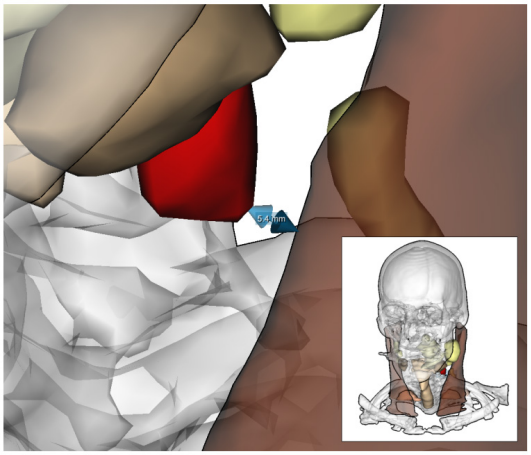


Figure 3.5: A detailed visualization showing the minimum distance between a red colored focus lymph node and a brown colored muscle to define the muscle's risk. An overview visualization representing the optimal lymph node viewpoint for the whole dataset is included. (Image reprinted from Mühler [119])

tasks, e.g., blood flow exploration, colonoscopy or sinuscopy are developed as well [45, 58, 69]. Diepenbrock et al. [45] and Hsu et al. [69] enable an efficient global and in-detail exploration, where a manual navigation is restricted and difficult, e.g., due to the structure's shape. However, their techniques are developed for direct volume rendering visualizations. Contrary, Gasteiger et al. [58] presented the FLOWLENS method as an exploration technique for an indirect volume rendering of a blood vessel and embedded flow visualization. Their FLOWLENS promotes a visual exploration by categorizing the visualization into focus and context, as shown in Figure 3.6. The applied illustration techniques are adapted with respect to this focus-and-context classification.

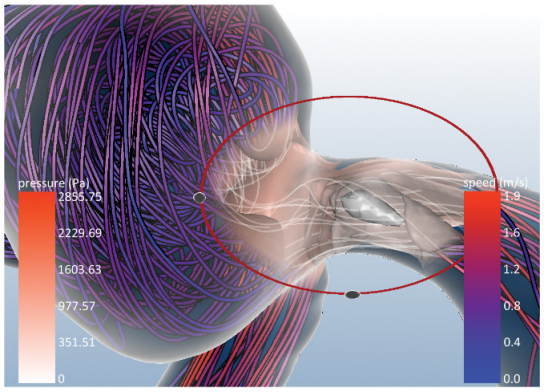
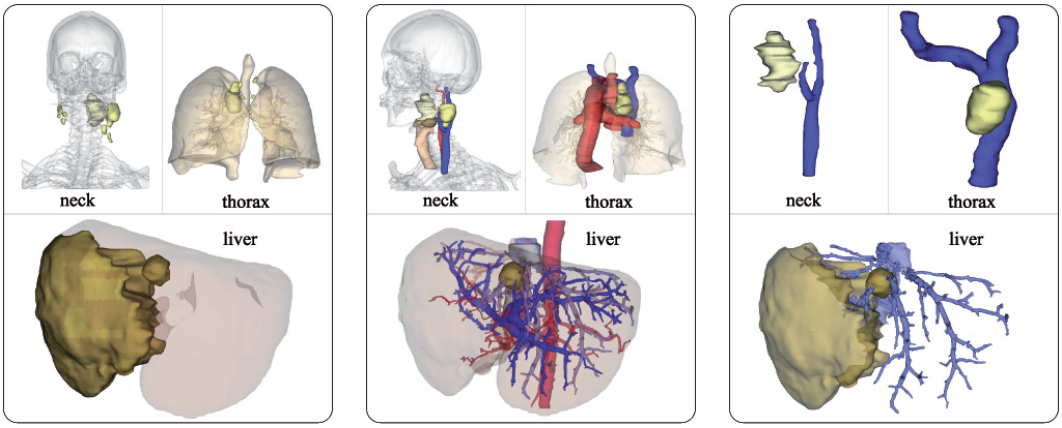


Figure 3.6: An aneurysm and parent vessel visualization, where the velocity is the context and depicted using illustrative streamlines. The focus is the flow pressure illustrated in the FLOWLENS using isosurfaces. (Image reprinted from Gasteiger et al. [58] © 2011 IEEE with kind permission from IEEE.)

The second above-mentioned strategy to generate an application-specific visualization is to minimize the number of structures and illustrate only question-specific and thus relevant structures. In medical education systems, customized 3D visualizations are generated using semantic relations between anatomic structures to identify the important structures that have to be visualized. In this way, the number of distracting and irrelevant structures as well as the number of required illustration techniques will be reduced. Viola et al. [163] already included an importance-driven technique in their approach. The structures are categorized into focus and context (less relevant) structures. The applied illustration techniques depict and emphasize this structure categorization, see Figure 3.4b. However, Viola et al. [163] defined structures such as the bones always as context and this technique is not suitable for an extensive treatment planning, e.g., bone fractures or infiltration risk determination of bones. Semantic transfer functions improve the focus and context specification [131, 142]. They use spatial focusing defined as an area-based focusing technique using geometric shapes and, therefore, reflecting the attentive focus. Salama et al. [142] used semantic transfer functions specifying the mapping of the volume attributes to the structures' visual appearance, e.g., transparency and color. Rautek et al. [131, 132] introduced a semantic layer and interaction-dependent concept for illustrative volume rendering that bases on fuzzy logic arithmetics. The user's interaction, distance to the illustration, and the data semantics define the structure classification in focus and context regions and thus their individually applied illustration technique. A semantic structure cate-



(a) Where are the pathologic structures located?

(b) Which structures are within a critical distance to the tumor?

(c) Which veins are affected by the tumor resection?

Figure 3.7: Question-specific 3D visualizations of neck, thorax, and liver for three treatment planning questions. (a) The pathologic structures are located. (b) Potential infiltrated structures (risk structures) are located and (c) veins required for the tumor resection are identified. (Images reprinted, with permission, from Baer et al. [7] © Eurographics Association 2010.)

gorization for indirect volume visualization was presented by Baer et al. [7]. Their importance-driven structure categorization is based on specific questions required for diagnostic and treatment planning. Baer et al. [7] presented an approach where structures are automatically categorized and only structures that are important to answer the question are visualized. Structures are divided into focus (structures of highest importance for the question), focus-relevant (structures that are semantically related to the current focus structures and question) and context (all other segmented structures) structures. This categorization is based on pre-calculated information for each structure such as meta information, geometric properties and determined pathologic risk information. Internally, a question is transformed to parameter values and weights. The final categorization is performed as thresholding encapsulated within a categorization pipeline. Figure 3.7 shows three different questions applied to three different datsets and anatomical regions.

Generally, a focus-context categorization is required to generate a customized 3D visualization. This categorization is either represented by customized viewpoints, animations and navigations or by a selection and illustration of relevant structures. However, all methods require appropriate visualization techniques to effectively communicate the information, guide the viewer's attention to the focus and support the visual perception. Techniques are used to either visualize one specific structure in detail or to visually discriminate into focus and context and thereby guide the viewer's attention to elements of importance. A visual abstraction and information reduction is realized. A few local regions are depicted in detail and emphasized and the surrounding contextual structures are illustrated with less detail for avoiding distraction from important structures. Context structures serve as additional information, orientation aid or shall transfer context information to facilitate the exploration. This abstractive way of illustration has a

great potential to depict the human anatomy, to show pathologic structures, to reveal their spatial relations to risk structures or to outline and describe potential treatment options. A good representation of shape and spatial relations including visual depth cues, e.g., shading, shadows, highlights and depth attenuation, promotes the visual perception of the patient-specific datasets. Moreover, the application of stereoscopic cues improves the 3D perception and facilitates, for example, the shape perception or the inquiry of quantitative 3D measures, e.g., extent or 3D distances while exploring 3D visualizations. However, developers have many choices with respect to the use of visualization techniques. This is aggravated by the fact that each visualization technique has a variety of parameters and that techniques may be combined with each other in a flexible manner. For example, surfaces of anatomic structures may be colored, textured, rendered semi-transparently, smoothed with a variety of techniques and finally be combined with silhouette and other feature line renderings. Therefore, the line style and color have to be defined as well. Since visualization and interaction techniques need to be carefully combined and integrated for clinical applications, experimental evaluations including knowledge of visual perception and attention are essential to promote the individual guidance ability and verify the perceptual assistance.

3.2 VISUAL PERCEPTION AND ATTENTION

Visual perception belongs to cognitive psychology, where psychologists study internal processes including perception, attention, language, memory and thinking. Cognitive psychology investigates the variables that mediate between stimulus and response by focusing on the way how humans process information [2]. To receive information from the environmental world, the human body has specialized organs called *sense organs*, e.g., eyes, ears, tongue, skin and nose, where sensory neurons are concentrated and operate as receptors to transmit these information to the brain. Visual perception is one of the five senses that allows the brain to intercept and interpret visible light within the visible light spectrum (wavelength between $380\text{nm} - 780\text{nm}$) to create the ability to see and perceive visual sensations in nature, e.g., brightness, color, contrast, shape or movement. This is provided by the light-sensitive sense organ known as the *eye*.

Visual perception starts with light entering the eyes, passing through the cornea and the pupil where the amount of light passing through is controlled by the iris and, finally, through the lens. An inverted image is then projected onto the light sensitive retina in the back of the eye. The most important structures in this layer of nervous tissue are the functional photoreceptor cells (rods and cones) that contribute to visual perception by being responsible for detecting color and light-intensity. The retina processes the information gathered by the rods and cones by converting it into neuronal signals that are sent to the brain via the optic nerve. Figure 3.8 illustrates the neural pathways from the eyes to the brain. The optic nerves on the outside of the retina pass through the optic chiasma, while the nerves on the inside of the retina cross over. Thus, the right part of the brain (hemisphere) receives and processes information about the left halves of the visual field and the left hemisphere about the right halves of the visual field. The optic nerves synapse onto cells in subcortical structures that are connected to the primary visual cortex from where the visual information travels along "what" and "where" pathways [2].

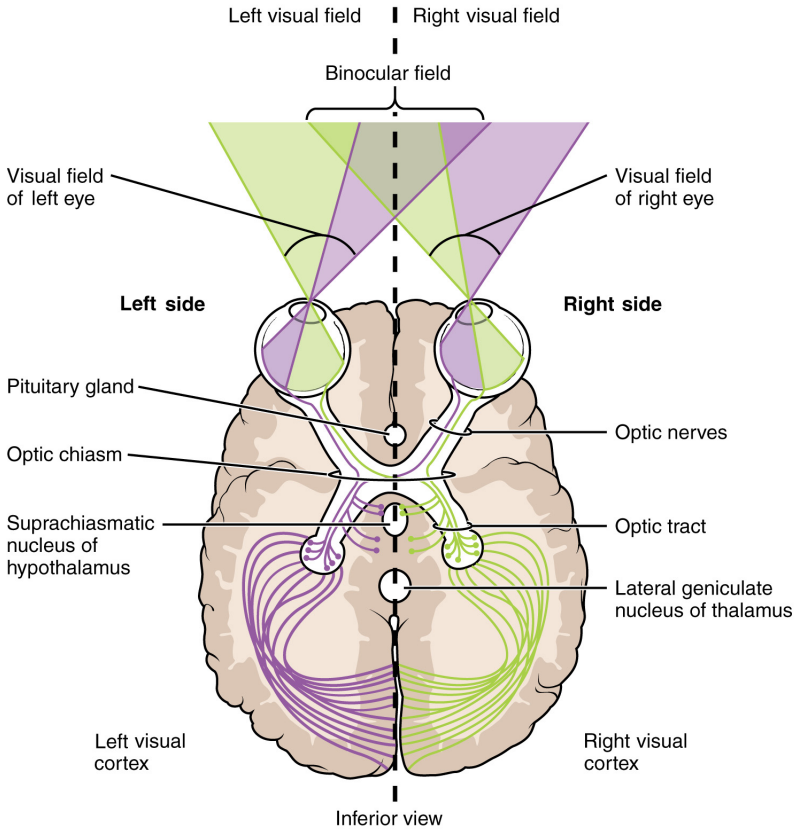


Figure 3.8: The neural pathways from the eyes to the brain. The optic nerves on the outside of the retina pass through the optic chiasma, while the nerves on the inside of the retina cross over. Thus, the right hemisphere receives information about the left halves of the visual field (green) and the left hemisphere about the right halves of the visual field (purple). (Image reprinted from OpenStax CNX® [161] *Sensory Pathways* © Creative Commons Attribute 4.0 License.)

The "what" pathway leads to brain regions that are specialized for identifying objects and the "where" pathway leads to regions that investigate spatial information and coordinate vision with action.

In summary, different processes are involved in visual perception to convert sensory input into perception. Some are physiological, caused by the reaction of the eye to light, which converts light into signals that can be understood. Others are cognitive such as *visual attention*, allowing the brain to interpret and understand the visual information, since what humans' see is not simply a translation of retinal stimuli, i.e., the images on the retina. Thus, for many years researchers' interest in perception have long focused on an explanation what visual processing does to create what humans' actually see. Primarily, what humans' see depends heavily on "where" the attention is focused and "what" knowledge exists before viewing the image or scene [66].

3.2.1 *Low- and Higher-Level Perception*

Basically, visual images are automatically and rapidly categorized into regions and properties that can be parallelized across the image [66]. Several factors influence where the attention is guided when observing an image or scenario. The human *visual attention* comprises various mechanisms and theories that help to define and identify the regions of an image that are selected for a more detailed analysis. A general and influential theory for visual perception and attention guidance is the *feature integration theory* (FIT) introduced by Treismann and Gelade [158]. According to the FIT, visual perception is characterized by two processing stages

1. **the preattentive stage**, where different features of an image, e.g., color, brightness, orientation, curvature, spatial frequency, and movement direction are analyzed in parallel and preattentively.

2. **the attentive stage**, where the low-level features are interpreted to recognize and locate the objects.

The first stage enables the perception of a limited set of visual features rapidly, automatically and in parallel by the low-level visual system. This preattentive processing enables the viewer to perceive certain properties of a presented scene in less than $200 - 250$ milliseconds. Initially, these properties were called preattentive, since their detection seemed to be without focused attention. During the past years, researchers found out that even at this early stage attention is included and plays an important role [66]. Simple search tasks, like determining whether there is a single red dot (target object) within a number of blue dots (distractor objects), can be performed preattentively. A target object that differs from distractors in a unique visual feature, allows the target to "pop out" and to be perceived preattentively (feature search). Thus, a target object can be easily detected regardless of the number of distractors. This so-called *feature search* is characterized by the detection of a target that differs usually in one feature from the distractors. Preattentive features are often classified into four categories: *color, movement, spatial localization, and form* [172]. In contrast to that, the combination of two or more visual properties, e.g., color and shape cannot be detected preattentively. This so-called *conjunctive search* requires selective attention and is performed in the second attentive stage. The second stage is controlled by visual selective attention. A serial search is required to confirm the presence or absence (combination) of features to enable a correct location of objects [158]. The visual search performance is dependent on the presented display size characterized by the number of distractors.

Since psychologists indicate that the first stage may also depend on expectations and attention, it is often referred to as *low-level processing*, whereas in *higher-level processing* (second stage) objects are recognized and classified and their spatial relations are derived. Besides that, there are further theories such as textons [82], similarity [130] and Boolean maps [70] that try to explain preattentive visual perception. For example, Wolfe et al. [178] introduced the *guided search theory* and showed that information from the first stage could be used to guide deployments of selective attention in the second stage. The viewer's attention is guided by stimulus-driven bottom-up and expectation-driven top-down processes. For a detailed discussion of preattentive features and theories see Healey and Enns [66].

3.2.2 *Depth Cues and Perceptual Factors*

Depth perception is the ability to perceive 3D images based on the 2D images that were projected onto the retina. Together with other senses such as sound, touch, and smell, humans asses the distance to objects and their layout through visual cues acquired with the eyes [1]. The visual system has to interpret various numbers of monocular and binocular cues to enable spatial and 3D depth perception, e.g., for estimating distance, depth and shape of objects. Figure 3.9 illustrates an overview and classification of depth cues according to Reichelt et al. [134], who proposed that the visual system enables depth perception based on *oculomotor* and *visual depth cues*.

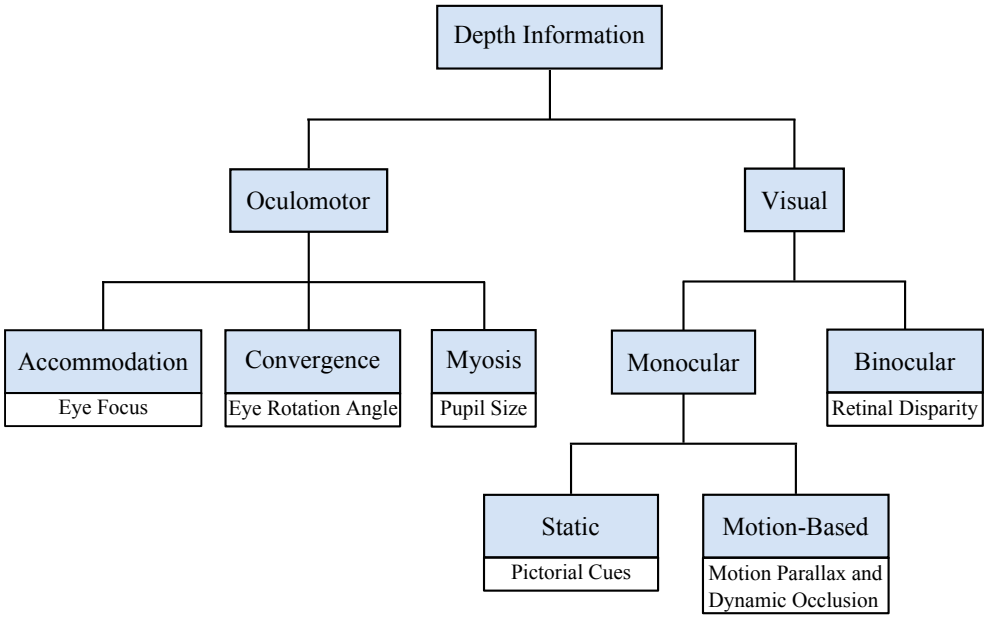


Figure 3.9: An overview and classification of depth cues. (Table reprinted, with permission, from Reichelt et al. [134].)

OCULOMOTOR DEPTH CUES are primarily based on the anatomical ability to change the rotation angle of the eyes and the tension of the eye muscles when focusing on near target objects [134]. Accommodation, convergence and myosis are occulomotor responses caused by this fixation of focus objects. Accommodation is the change of the eye's lens shape (optical power), in order to maintain an object sharp on the retina as its distance varies. The tolerance of distance differences along the optical axis without loosing sharpness of the focused object is called the *ocular depth of focus* or *depth of field*. Convergence is the difference of the direction of the eyes, as they have to converge to focus on near objects. Since pupil size, affected by the luminance level, influences the depth of focus, Reichelt et al. [134] include myosis as an important oculomotor cue. Myosis defines the pupillary constrictions when focusing on near focus objects.

VISUAL DEPTH CUES are depth information that are either monocular or binocular. Monocular depth cues are cues that enable a depth information extraction based on only one image and with one eye, respectively. They can be divided

into *static* (pictorial) and *motion-based cues* such as motion parallax. Static classic pictorial depth cues such as occlusion, shadows, texture, shading, relative height, relative and familiar size, atmospheric and linear perspective are powerful cues that provide relative depth information (e.g., relative ratios or ordering information) and therefore enable a depth perception even for static 2D images. Contrary, motion-based cues comprise 3D depth perception induced by shifts on the retina through relative movements between the viewer and the focused objects [134]. Kinetic depth, motion parallax and dynamic occlusion are cues where motion is used to perceive 3D structural information about objects. Depth information are either perceived by the movement of an object (structure from motion) or a distance-depending speed difference of moving objects with respect to a stationary viewer, or the extent of dynamic accretion and deletion of object occlusions are used to perceive depth information. Binocular vision describes vision with two eyes and the main cue for depth perception is retinal disparity. Binocular depth cues use the fact that the eyes receive two adjacent views of the world projected slightly shifted onto the left and right retina, recall binocular field in Figure 3.8. This horizontal difference caused by the difference between the left and the right eye is called *disparity*. The amount of disparity depends on the object's depth and thus is a cue that the visual system uses to infer depth. The term *binocular parallax* or *stereopsis* is most often used to refer to depth perception derived from binocular disparity. However, it is an important visual cue only for short distances. The sensitivity of stereopsis in representing small disparity differences is denoted as stereoacuity and is influenced by factors such as luminance, spatial frequency, observation time, and contours [1]. A detailed discussion of depth cues in real world and virtual realities can be found in Preim and Dachselt [127].

Overall, depth perception involves a consolidation of the oculomotor and visual depth cues. Visual depth cues include different monocular and binocular cues to form a 3D representation of the world from the 2D images projected onto the retina. These received information is then matched with structural descriptions in memory and interpreted to perceive, recognize and localize 3D objects.

3.3 EXPERIMENTAL EVALUATIONS OF 3D MEDICAL ILLUSTRATIONS

A large variety of illustration techniques have been developed to generate interactive 3D visualizations of medical image data derived from imaging modalities, such as CT and MRI. Any visualization technique that illustrates the given structures or structure information conceptually and thus produces a meaningful, expressive, and simplified representation can be denoted as *illustration technique*. Generally, techniques exaggerate or suppress visual features of scene elements, e.g., illustration style, hue, luminance, sharpness, or size to achieve visual guidance. However, in complex interactive visualizations, modifying these features may not be sufficient to reliably attract the user's attention. In addition, some techniques visually alter structure properties and thereby encourage a misunderstanding of the scene, e.g., fisheye distortions aggravate a relative distance judgment and blurring of context can hamper the identification of contextual details [168]. As mentioned in Section 3.1.2, the techniques must accurately illustrate the image data and facilitate the exploration by generating application-specific and effective

illustrations. Medical visualizations and illustration techniques are evaluated with respect to their ability to illustrate the important information, to improve shape and depth perception and to guide the viewer's attention.

THE SHAPE PERCEPTION of 3D illustrated objects can be supported and influenced by factors such as texture and shading. The accuracy of surface perception can be measured and the influencing factors can be modified in order to decrease the error in shape perception [165]. The most common experiment tasks described by Todd [157] and Koendrink et al. [87] for probing perceived surfaces are:

- **the relative depth probe task**, where participants were asked to estimate the distance of two points on the surface that are marked with different colors.
- **the gauge figure task**, where participants were asked to orient gauge figures to coincide with the perceived surface normal vector. The gauge figure technique is the most common technique, since it is the most natural and reliable task, and it is easy to understand.
- **the depth-profile adjustment task**, where participants have to assess a shaded surface overlaid by dots using a second view where the dots are presented over a blank background. Participants have to adjust the dots of the second view so that they fit the perceived height profile of the first view with the shaded surface.

THE DEPTH AND SPATIAL PERCEPTION of structures in a 3D visualization can be supported by cues such as perspective, global illumination, shadow and occlusion. Common evaluation tasks used to investigate depth or spatial perception are [165]:

- **depth judgment tasks**, where participants were asked to estimate the distance of structures compared to the user's viewpoint and inter-surface distances (depth ordering), respectively, see Figure 3.16a and 3.16c. This task is an adapted and modified version of the well-established relative depth probe task for shape perception evaluation.
- **tracking tasks**, where participants were asked to follow structures or describe the spatial pathway of structures, see Figure 3.16b.

THE VIEWER'S ATTENTION as well as the technique's ability to attract the attention and to emphasize the focus-and-context categorization are either evaluated with visual search tasks based on findings from low- and higher-level perception (see Section 3.2.1) or comparatively using rating scales to analyze the subjective opinions of newly developed techniques compared to common approaches or informal by only recording comments.

A perception-guided evaluation is important to ensure accuracy and contribute to an effective and expressive communication of the essential information. Besides that, findings from experimental evaluations are increasingly used to improve virtual and augmented environments and to analyze and improve the perception of

3D stereoscopic views. Participants' performance, subjective preferences and work practices are analyzed to evaluate visualizations, establish techniques and support further refinements by identifying existing limitations and reporting requirement analyses.

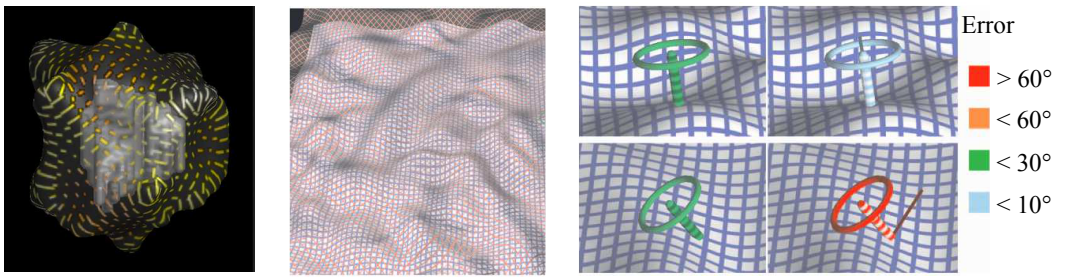
In the following sections, selected evaluations of 3D medical illustrations and important findings will be presented and discussed. They analyzed point-based and line drawing illustrations, smart visibility illustrations as well as stereoscopic views in terms of the above-mentioned visual perception aspects. Since it is difficult to recruit medical experts, the recruited number of experts will be mentioned explicitly, if it was noted in the individual publication.

3.3.1 *Point-Based and Line Drawing Illustrations*

Originally, conceptional and abstract illustrations, e.g., anatomical, archaeology or technical drawings were hand-drawn, mainly with a black pen on white paper, after which they were named as pen-and-ink illustrations. One of the first anatomic and scientific so-called *pen-and-ink illustrations* were published by Gray [61] and Hodges [67]. Due to the former available manufacturing and reproduction options, these illustrations consist of simple but well-defined and well-placed lines and dots, e.g., hatching, feature lines, silhouettes, and stippling. In principle, such *non-photorealistic illustrations* focus on point- and line-based graphics derived from traditional pen-and-ink methods.

An early evaluation investigating the effect of line drawings for architectural images was presented by Schumann et al. [147]. Their questionnaire-based study with 150 architects and architectural students asked for quantitative ratings and qualitative feedback. The non-photorealistic sketch was rated significantly better on affective and motivational criteria than the corresponding shaded images. A second important aspect - the guiding of the viewer's attention - was studied by Santella and DeCarlo [143] with an eye-tracking experiment where 74 students participated. The results showed that the local treatment of abstraction does have an effect on where people look, with the salience-based and fixation-based adaptive abstractions receiving fewer fixation clusters than the other more detailed images. Further evaluations studying the potential and effectiveness of dots and lines representing 3D surfaces and shape followed [35, 71, 77].

Saito and Takahashi [140], Interrante et al. [72, 73], Kim et al. [86] and Bair and House [11] showed that depth perception of medical surface models can effectively be improved by hatching along curvature directions. Texture features in terms of 3D shape perception were analyzed using the gauge figure task. Participants had to draw normal vectors onto the surface or orient existing vectors, as shown in Figure 3.10c. Interrante et al. [72] and Bair and House [11] assessed renderings of transparent surfaces with sparsely-distributed discrete, opaque textures containing lines systematically oriented at an oblique angle to the principal directions. The transparent surfaces surrounded an internal structure or a transparent organic-shaped layer overlaid an opaque shaded surface, see Figure 3.10a and 3.10b. This task is derived from medical visualizations, where multiple superimposed layers



(a) (b) (c)

Figure 3.10: (a) A transparent isointensity surface applied with a principal direction texture representing a radiation dose distribution, surrounding the opaque treatment region. (b) A transparent organic-shaped surface with a grid texture lies on top of an opaque surface and (c) a visualization of normal vector probe results, where green and blue represent results that deviate less than 30° from the computed normal vectors and red illustrates a deviation of more than 60°. (Images reprinted from (a) Interrante et al. [72] and (b)(c) Bair and House [11] © 1997 and 2007 IEEE with kind permission from IEEE.)

representing different anatomic structures or information have to be displayed, e.g., an organ as outer layer and its vascular supply as inner layer. The accuracy of normal vectors drawn or oriented by the participants turned out to be a good criterion to judge the different techniques and to measure how clearly the rendering method conveyed shape information. The results showed that a textured transparent surface compared to an untextured surface improves the shape perception statistically significant. However, all participants recruited by Interrante et al. [72] saw the images using a glasses-based stereoscopic display and, thus, the results were achieved as a combination of texture and stereoscopic view and the caused effect cannot be traced back uniquely to neither texture nor stereoscopic view.

Later on, Cole et al. [34] tried to assess where artists draw lines to convey the shape of a surface and examined how well these lines can depict the surface shape. Cole et al. [35] presented a study comprising images of 14 different objects illustrated with six rendering styles. 90, 180 or 210 randomly selected gauge figures had to be oriented by the 560 participants. Gauge figures are represented as small discs with a line representing the normal vector and were placed by Cole et al. [35] on each painting. Participants were asked to orient them according to an imaginary normal vector and the deviation of the oriented gauge figure and the computed surface normal at this point served as a measure of how well the shape was perceived. Even though the shapes of several objects were clearly perceived (gauge figure deviation of 15° on average), the gauge figure results showed that the participants had difficulties with the shape of the anatomic structure. The estimated gauge figure deviated from the calculated normal vector on average between 30° and 40°. Thus, Cole et al. [35] concluded that complex anatomic shapes cannot be fully understood by only applying one of the six rendering styles: fully shaded, occluding contours, apparent ridges, ridges and valleys, suggestive contours and a binarized human artist drawing. However, they used illustrations of a tooth, a cervical and a vertebra dataset that are difficult to recognize without any medical knowledge. A discussion of the importance of feature lines for medical visualiza-

tions is given by Lawonn et al. [103].

During the last decade, several non-photorealistic techniques were developed, refined, and different techniques were combined to generate effective interactive 3D medical visualizations. Since they enable a sparse illustration of complex structures and situations and are used for medical illustrations since hundreds of years, they are proven to be suitable to illustrate medical situations. Thus, illustrative visualizations provide useful rendering alternatives to conventional volume or surface rendering in medical visualization. For example, Csebfalvi et al. [39] visualized object contours based on the magnitude of local gradients as well as on the angle between viewing direction and gradient vector using depth-shaded maximum intensity projection. Lu et al. [110] developed an interactive direct volume illustration system that simulates traditional stipple drawings for scientific and medical datasets, see Figure 3.11. Later on, this stippling approach was adapted and an object-based isosurface visualization of medical datasets with stippling textures was presented by Baer et al. [5] and was further improved using different shading approaches by Tietjen et al. [156]. Lum and Ma [112] presented an approach for an interactive high-quality non-photorealistic direct rendering of volume data including tone shading, silhouettes, gradient-based enhancement, and color depth cueing (warmer colors for foreground and cooler colors for background). Thus, important medical visualization tasks, e.g., the visualization of liver and thorax data [155], vascular structures [65, 137], fiber tracts extracted from MR DTI data of the brain [48, 153] or endoscopic views [105] can be realized using non-photorealistic techniques. These visualization techniques were qualitatively or quantitatively evaluated in terms of depth and shape perception for medical visualizations.

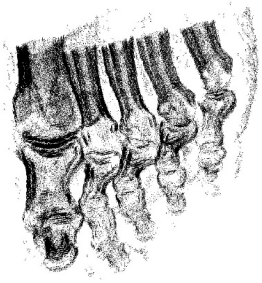
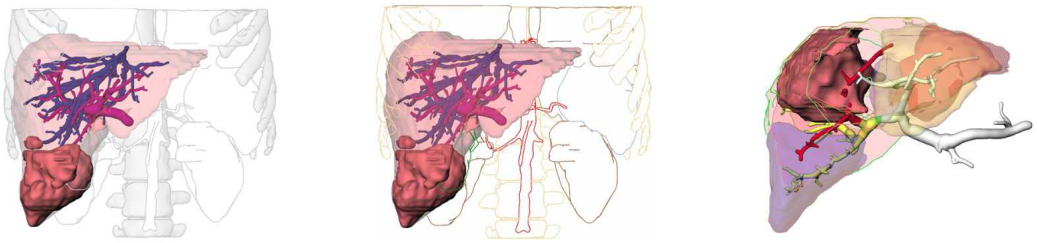


Figure 3.11: A direct stipple rendering of a foot dataset with a silhouette enhancement (Image reprinted from Lu et al. [110] © 2002 IEEE with kind permission from IEEE.)

Tietjen et al. [155] presented an approach to effectively convey medical shape information and to emphasize features while sparsely presenting context information. These visualizations can be used for diagnosis, documentation and education purposes. A combination of silhouettes, surface shading and direct volume rendering was introduced and evaluated using segmented patient-specific datasets of the liver and the thorax region, see Figure 3.12. The general concept of their work was to derive the importance of a structure (e.g., from user input), display the most important structures and apply the rendering styles. Thus, the illustration techniques were used to categorize and illustrate the structures as focus, near focus and context structures. As explained in Section 3.1.2, this classification is suitable to sup-

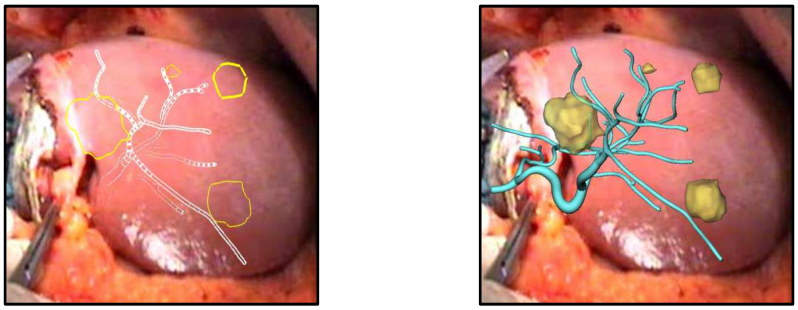


(a) (b) (c)

Figure 3.12: The favored hybrid visualizations. The focus structure (liver) is visualized using a colored surface rendering. (a) The near focus structures (bones) are rendered with a surface shading and the context structures (kidney and milt) with silhouettes. (b) When all structures, besides the focus, are classified as context, colored silhouettes facilitate the visual distinction of the structures. (c) A hybrid visualization of the liver, where the regions that would be affected by a tumor resection are emphasized with colored silhouettes. (Images reprinted, with permission, from Tietjen et al. [155] © Eurographics Association 2005.)

port the question-driven diagnostic and treatment planning process and to generate customized visualizations. Tietjen et al. [155] evaluated hybrid illustrations that combined the three rendering styles and analyzed whether line rendering is a meaningful extension to the surface shading and direct volume rendering. 33 participants including eight medical experts were asked for personal preferences and had to assess the visualizations with respect to specific questions used in typical liver surgery planning. Overall, for the comparison of a direct volume rendering with a transparent illustration no significant difference was registered. Contrary, almost all participants with low medical knowledge (approximately 80%) favored a surface shading with additional silhouettes (see Figure 3.12a) and colored silhouettes that visually enhance a structure distinction (see Figure 3.12b). Six of the eight medical experts favored the silhouette illustration to highlight vascular territories in the liver visualization, shown in Figure 3.12c. However, an exclusive silhouette visualization without further shading or color was not preferred compared to a hybrid visualization.

VASCULAR VISUALIZATIONS. Ritter et al. [137] used texture to illustrate spatial properties (e.g., distances between or to other structures) and combined them with color to illustrate vessel branch properties (e.g., branching level and vessel diameter). Moreover, they integrated the hatching technique to generate distance-encoded shadows and thus improve the depth perception. Distances at vascular branch intersections, to other structures and distances between vascular branches without intersection are encoded by the number, the texture and width of strokes. To evaluate their approach, they conducted a web-based questionnaire with 160 participants (38 being physicians or medical students). Participants were asked to order the branches starting with the closest, to track specific branches and to determine the branch order at marked branch intersections. Ritter et al. [137] compared their approach with a commonly used Gouraud shading. An improvement of the shape perception could not be confirmed by Ritter et al. [137]. Thus, Chu et al. [33] combined an enhanced silhouette drawing to improve the vascular shape vi-

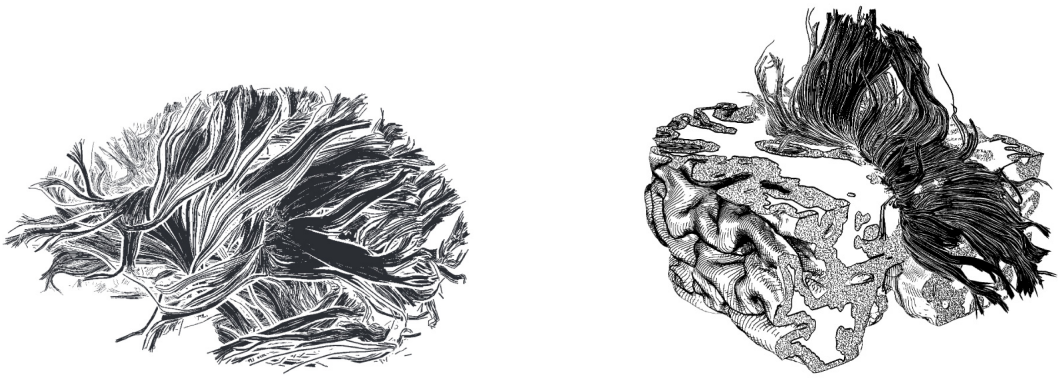


(a) (b)

Figure 3.13: (a) This illustration approach is then integrated into an open liver surgery by projecting this non-photorealistic rendering and the illustrated tumors onto the liver organ and compared (b) to a classical rendering overlay. (Images reprinted from Hansen et al. [65] © 2010 CARS with kind permission of Springer.)

sualization. Later on, Hansen et al. [65] embedded the technique of Ritter et al. [137] into an augmented reality environment for the intra-operative use in liver surgery (see Figure 3.13) and evaluated this approach as well. They recruited six liver experts, who were asked to order vessel branches and to estimate distances. The answers were recorded similar to a think-aloud protocol. Initially, they saw video frames and photos from a projector-based augmented reality visualization. Later on, preliminary studies in the operating room were preformed. Laparoscopic and open liver interventions as areas of application and three different scenarios (overview, focus on the tumor and focus on a resection plane) were used. Intraoperative overlay visualizations generated with the adapted distance-encoding surface and silhouettes (see Figure 3.13a) were compared with a standard colored surface rendering (see Figure 3.13b). The results of the quantitative study of Ritter et al. [137] and the qualitative evaluation of Hansen et al. [65] showed that this approach enables a more accurate and accelerated depth and spatial perception for vessel trees and branches compared to a Gouraud shading and colored surface rendering.

VISUALIZATIONS OF FIBER TRACTS. A depth-dependent halo technique for dense line bundles was presented by Everts et al. [48]. They created halos that do not overlap around tight line bundles illustrating diffusion tensor imaging (DTI) fiber tracts and thus use occlusion as one depth cue (recall Section 3.2.2). An additional depth cue is integrated by applying a distance-depending line width adaption (see Figure 3.14a). This technique is used to emphasize important dense line bundles and to diminish less structured line bundles to achieve a focus-and-context bundle categorization. They performed an informal evaluation with four medical experts investigating the technique's potential to show detail and depth relations and to produce high-quality fiber tract illustrations. Everts et al. [48] reported that all participants were impressed by this technique and would prefer this visualization compared to tract visualizations they were used to (e.g., colored lines or tubes based on direction). As shown in Figure 3.14b, Svetachov et al. [153] additionally visualize the brain with hatching and a stippling technique that is based on an ambient occlusion calculation. Thus, another depth cue (illumina-



(a) (b)

Figure 3.14: (a) A subset of DTI fiber tracts visualized with depth-dependent halos (Image reprinted from Everts et al. [48] © 2009 IEEE with kind permission of IEEE.) (b) and the surrounding brain visualized with stippling and hatching. (Image reprinted from Svetachov et al. [153] © 2010 John Wiley & Sons with kind permission of John Wiley & Sons, Inc.)

tion) is integrated to support the depth perception. However, Svetachov et al. [153] conducted only an informal evaluation with two medical experts that were also involved in the technique development. Thus, the results are very restricted to two participants and the potential of this approach is still not carefully analyzed due to the visual depth perception support. Eichelbaum et al. [46] presented a novel ambient occlusion approach that retains global and local structural and spatial information. This approach might improve the visual quality and supports the perception of the single fiber tracts as well as the fiber bundles. However, a perception-guided evaluation was not conducted.

ENDOSCOPIC VISUALIZATIONS. Line drawing illustrations of interior cavities were considered by Lawonn et al. [105] for endoscopic image data, see Figure 3.15. They examined real-time hatching from Praun et al. [124], high-quality hatching from Zander et al. [182] and the contour- and feature-based illustrative streamlines method (ConFIS) by Lawonn et al. [103]. These three line drawing concepts were qualitatively evaluated to assess their ability to represent interior branches and specific anatomic features. Seven medical knowledgeable participants were asked to rate four different endoscopic datasets visualized with each of the three techniques using a 7-point rating scale. Initially, they saw each dataset visualized with each technique and had to assess the perception of different features and branches. Then, the shaded visualizations of each dataset were presented. At the end, the participants saw all illustrations again and were asked to compare them with regard to the technique's ability to capture salient regions and illustrate depth and spatiality. The recorded ratings and accuracy of detected branches were used to analyze the technique's expressiveness to illustrate endoscopic views. The real-time hatching and the ConFIS method had the highest ratings and best accuracy results and thus provided the best spatial impression, while the high-quality hatching was insufficient for a 3D visualization of interior cavities.

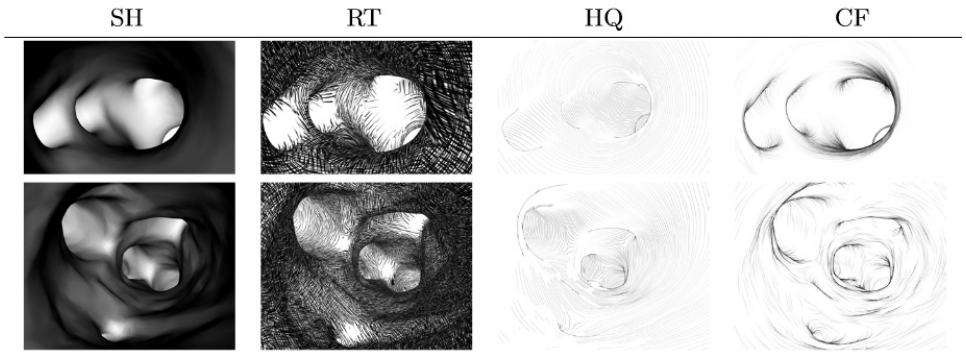


Figure 3.15: Two endoscopic datasets visualized with shading (SH), real-time hatching (RT), high-quality hatching (HT), and ConFIS (CF). (Image reprinted from Lawonn et al. [105] © 2014 Springer-Verlag Berlin Heidelberg with kind permission from Springer.)

DISCUSSION. Overall, there are several methods to illustrate patient-specific datasets using non-photorealistic and hybrid rendering techniques, respectively. Since computer-generated point and line drawings are difficult to produce, it is essential that the generated visualizations support the exploration by guiding the viewer's attention and facilitating depth and spatial perception. Current approaches integrate different depth cues, e.g., occlusion, saturation or perspective to improve traditional pen-and-ink methods. On the one hand, the existing evaluations are primarily informal with a few selected experts, who are often deeply involved in the development process and thus their positive feedback is unsurprising. This approach does not solve the problem of selecting or identifying appropriate techniques. In particular, it turns out that intra-individual differences are large and that even the preferences of one user often depend on specific datasets. On the other hand, a participant group with less than ten people combined with limited medical knowledge is not sufficient as well. However, these developments and evaluations show the potential of non-photorealistic rendering techniques.

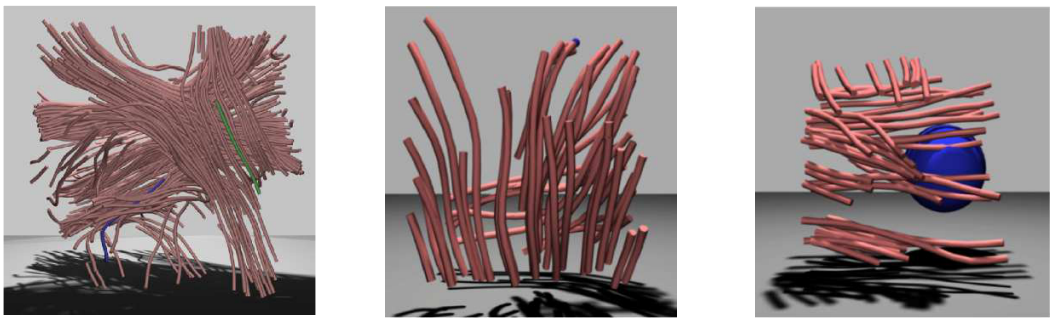
3.3.2 *Smart Visibility Illustrations*

Techniques that aim at a higher level of abstraction, e.g., importance-driven [162] or semantics-driven illustrations [131, 132] to generate expressive visualizations and, thus, to enhance the visual comprehension are called *smart visibility techniques*. They integrate relevance information to generate a focus-and-context visualization, e.g., interactive cutaways [24], close-ups [23], exploded views [24], peel-aways [36] or ghosted views [57]. The basic strategy of smart visibility techniques is to emphasize the most relevant visual information of an object by means of local modifications of visual attributes or changes in spatial arrangement. These techniques are known from technical illustrations and have been successfully applied to medical visualizations to reveal tumors and vascular structures in organs [95]. Predominantly, they are developed by combining already existing techniques (e.g. color, opacity, sharpness, shadows or fog) or by improving and refining parameters and parameter combinations. To avoid the misleading interpretation of structures and spatial relations, perceptual experiments have examined the effectiveness of several techniques that contribute to the mental reconstruction of the 3D struc-

tures from visualizations displayed on a 2D screen. For example, transparency [72, 20, 31], texture [75, 86, 11], depth of field [90], shading [27] and illumination [54, 27] were investigated and analyzed with respect to task performance, accuracy and subjective preference.

Caniard and Fleming [27] and Fleming et al. [54] used the gauge figure technique to study the effect of illumination and specular reflections on shape perception. Gauge studies can document not only shape interpretation but also the priors, bias and information used by the human visual system. While Fleming et al. [54] recruited only two participants, Caniard and Fleming [27] presented a larger evaluation with 21 participants (13 female and 8 male) and 20 different shapes. Even though they tested 3D visualization techniques, static images of 3D models were presented. However, they found a significant effect of illumination on shape perception and showed that shape from shading with directional local illumination is very sensitive to the position of the light source. Thus, it is essential to integrate illumination such that the participants can estimate the light source or direction to support an accurate shape from shading perception. Based on these findings and evaluation experiences, Šoltészová et al. [165] introduced a shading model and rendering pipeline that updates the rendering algorithm based on the results of a shape perception experiment. Thus, a shape enhancement of visualization which is driven by an experimentally-founded statistical model is developed. They studied Lambertian-shaded surfaces with 40 participants (19 female and 21 male) orienting 160 gauge figures on surfaces rendered with two shading conditions. Based on the results obtained in the experiment, a second model of correction was created and applied. Their results showed that the human ability to estimate surface shape is best on surfaces where normal vectors point upwards and worst where normal vectors point downwards.

Weigle and Banks [173] and Penney et al. [123] investigated global illumination, texture and motion using depth judgment and tracing tasks and showed 3D streamtube visualizations. Weigle and Banks [173] focused on linear perspective

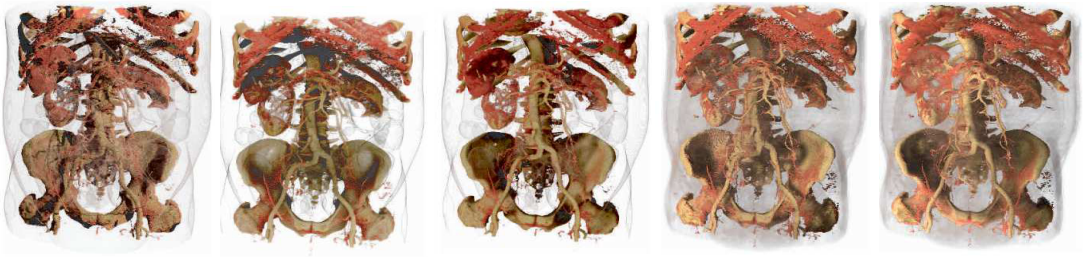


(a) Depth Judgment (b) Visual Tracing (c) Contact Judgment

Figure 3.16: A 3D streamtube visualization where the participants had to (a) estimate which of two differently colored tubes was closer to the viewer, (b) trace a selected tube and mark its endpoint, and (c) judge the relationship between the sphere colored in blue and the tubes around the sphere. (Image reprinted from Penney et al. [123] © 2012 IEEE with kind permission from IEEE.)

and physically-based (global) illumination and Penney et al. [123] on global illumination combined with texture and motion. Participants had to perform a depth judgment and a shape description task. Two experiments (between-participant design for the depth judgment and shape description task) were performed by Weigle and Banks [173]. In the first participated five (1 female and 4 male) and in the second 12 participants (4 female and 8 male). Their results for shape perception were not significantly different, while the depth judgment task results showed that the illumination provides improved perception of relative depth and that the strongest perceptual cue is achieved by combining linear perspective and physically-based illumination. One experiment (within-participant design) with 26 participants, who performed three tasks with 16 stimuli each (four stimuli for each technique) was conducted by Penney et al. [123]. The study of Penney et al. [123] investigated the visualization of 3D tensor field streamtubes and therefore used a tensor field sampled from a full-brain diffusion tensor magnetic resonance imaging (DTI) dataset as stimulus, as shown in Figure 3.16. Their experiment analyzed task completion time, error rate (accuracy) and subjective preference and showed that motion, global illumination and texture are very strong cues but have to be used with caution.

Lindemann and Ropinski [109] investigated the influence of seven volumetric illumination models on the spatial perception of volume rendered visualizations of medical data, see Figure 3.17. A within-participant study with 55 recruited participants (56% male) performing four tasks for each of the seven techniques using 21 or 42 stimuli (depending on the task) was conducted. They analyzed the relative and absolute depth, relative size and subjective preference with the help of depth judgment, depth approximation and size ordering tasks. Since the stimuli are very complex visualizations, Lindemann and Ropinski [109] tried to isolate the measured effect by keeping other visualization parameters constant, except for the illumination model. Their results were very different and it was not possible to define one illumination model as the most suitable model. Participants achieved the most accurate result for the relative depth and size perception when a stimulus with the directional occlusion shading was presented. However, the

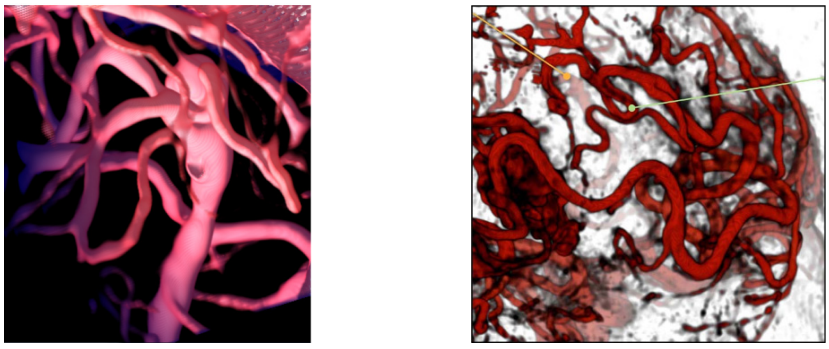


(a) (b) (c) (d) (e)

Figure 3.17: The five techniques of the seven investigated illumination models applied to renderings of a CT dataset of a human body. (a) The half angle slicing, (b) directional occlusion shading, (c) multidirectional occlusion shading, (d) shadow volume propagation, and (e) spherical harmonic lighting showed the best task performance and preference ratings. (Image reprinted from Lindemann and Ropinski [109] © 2011 IEEE with kind permission from IEEE.)

subjective results for this model were below 50%. The half angle slicing model supported participants to perceive a more accurate absolute depth and was preferred by 68%. Overall, they showed that depth and size perception in volume-rendered smart visibility illustrations is statistically significantly improved when an advanced lighting model is applied, compared to the Phong technique.

Further experiments investigated the potential of transparency [31], the depth of field [63, 138], color [79], occlusion [170] or fog and kinetic depth [84] to support the depth and spatial perception of medical volume data. Grosset et al. [63] performed an eye-tracking experiment, where the 25 participants (6 female, 19 male) were asked to perform a depth sort evaluation and found out that a depth of field improves the depth perception if the focus is close to the viewer. If the feature is far away, the user performs worse. Depth cues for the visualization of angiography image data were analyzed by Joshi et al. [79], Kersten-Oertel et al. [84], Ropinski et al. [138], see Figure 3.18. Ropinski et al. [138] recruited 14 participants, who performed real-world tasks derived from the diagnostic workflow and had to rate the techniques using a 6-point Likert scale. They focused on isolating the effect caused by the different techniques and at the same time tried to prevent model recognition. Thus, they either presented different datasets or the same dataset from a different perspective. Moreover, static images were chosen to prevent motion as another depth cue influencing the investigated cues. They stated that the visualization of angiography data benefits from color depth information. Especially the pseudo-chromadepth technique combined with a depth of field illustration improves spatial perception. This result was confirmed by Joshi et al. [79] (see Figure 3.18a) performing a comparative evaluation with domain experts and Kersten-Oertel et al. [84] conducting an extensive quantitative evaluation. Joshi et al. [79] recruited 12 medical experts, who were asked to chose one of two presented visualizations. Kersten-Oertel et al. [84] confirmed that both novices (19 participants) and experts (3 participants) performed significantly better when the pseudo-chromadepth and fog cues were presented, as shown in Figure 3.18b.

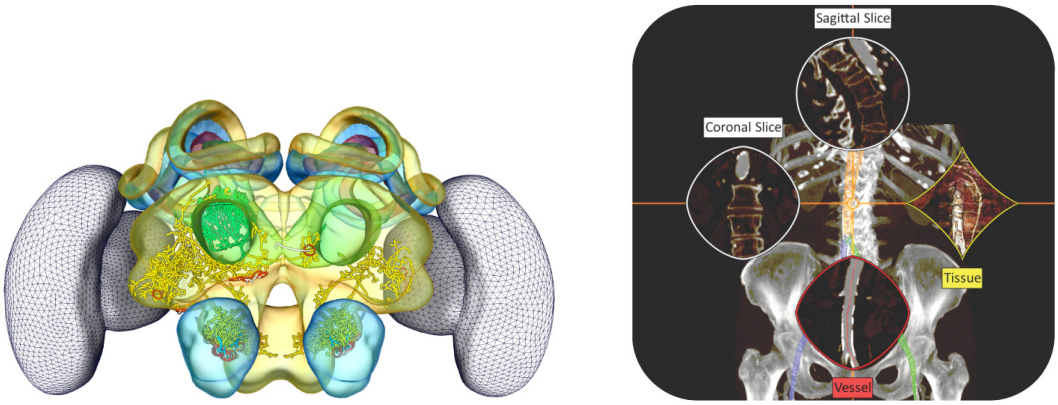


(a) (b)

Figure 3.18: (a) A 3D vessel visualization with a distance color blending technique. The vessels that are further away are colored in a shade of blue. (b) A vessel visualization including fog and edge enhancement. Participants were asked to determine which vessel (green and orange points) is closer. (Images (a) and (b) reprinted from Joshi et al. [79] and Kersten-Oertel et al. [84] © 2008 and 2014 IEEE with kind permission from IEEE.)

Thus, distance-encoded color represents a very strong cue for depth and spatial perception of vessel structures. The effect of five visual cues, e.g., stereo, local occlusion via depiction of edges, aerial perspective, kinetic depth, and pseudo-chromadepth was evaluated by Kersten-Oertel et al. [84]. Participants were asked to assess vessel pairs in terms of their depth order. Moreover, the depth ordering task was improved by choosing stimuli with vessel pairs that were either near, medium or far apart both on the screen and in depth [84]. Due to accuracy and response time, color and edge enhancement improves the task performance in case of experts only. In contrast, stereoscopic and kinetic depth cues did not yield good results, neither in time nor in correctness [84]. Šoltészová et al. [164] investigated soft shadows and designed an evaluation with three different tasks to validate the shadow's benefit for surface, contrast and depth perception. A gauge figure task was used for the shape perception, a color estimation task for the contrast perception and a depth judgment task of three prepositioned points for the depth perception. Thus, they analyzed each of the three properties of the shadow technique individually with customized tasks.

Eye tracking in order to quantify the effects of medical visualization in the perception process and to evaluate the attention guidance potential was performed by Krupinski [94] and Burgert et al. [26]. They analyzed scanpaths, areas of interest, attentional landscapes to interpret the inspection and search strategies of experienced and young medical doctors. Krupinski [94] analyzed scan patterns of experienced and unexperienced radiologists, who were asked to detect lung nodules in mammograms. Experienced users were faster, since they picked up suspicious features earlier, restricted the search region and stopped their visual search earlier



(a) (b)

Figure 3.19: (a) A 3D visualization of three nerves (filaments) in a bee brain and a surrounding structure. The nerves are colored by their surrounding surfaces, ring-shaped glyphs are placed at filament-surface intersections and a halo effect to improve depth perception. (Image reprinted from Kuß et al. [97] © 2010 John Wiley & Sons with kind permission of John Wiley & Sons, Inc.) (b) A smart super view visualization technique, where focus structures were enlarged in individual regions of interest for a detailed exploration. (Image reprinted from Mistelbauer et al. [115] © 2012 IEEE with kind permission of IEEE.)

due to their diagnostic experience. Medical students searched more thoroughly and processed more image information. Kuß et al. [97] used a complex scene consisting of three volumetric objects (bee brain) and one transparent filamentous structure (nerves) as stimulus, as illustrated in Figure 3.19a, which they presented for a few seconds and then asked the participants whether a filament runs through a transparent structure. 48 participants had to perform a conjunctive search and the different visualization techniques were evaluated with respect to accuracy, preferences and response time to choose the answer. The majority of evaluations investigating smart visibility techniques are comparative and questionnaire-based or informal. Li et al. [107] and Mistelbauer et al. [115] demonstrated their system for authoring and viewing interactive smart visibility illustrations (see Figure 3.19b) to medical educators, physicians or illustrators to gather informal feedback for the preference and guidance effectiveness. Gasteiger et al. [57] presented ghosted view visualizations pairwise and asked the participants to decide which they preferred regarding a specific criterion (e.g., depth perception). These findings are, however, strictly bound to particular applications and it is difficult to generalize from them. There is considerable evidence that such preference or meta-tasks – where one's beliefs and opinions are surveyed – does not always correlate well with actual task performance measured quantitatively [169]. Controlled experiments can be performed such that the results are more general and enable a quantitative analysis of specific visualization techniques and properties.

3.3.3 *Stereoscopic Views*

In order to explore and to navigate through a 3D visualization of the patient's anatomy, the visualization has to provide very accurate depth and spatial information. Beyond the depth cues integrated and produced by visualization techniques, binocular and motion parallax are the most significant sources of depth information [166]. Both cues were investigated using depth estimation and positioning tasks [19, 22, 114]. The results showed that depth acuity from binocular parallax is better than from motion parallax. However, a combination of both leads to an increased task performance compared to binocular parallax alone [19]. Thus, stereoscopic views combined with motion parallax represent a potential alternative compared to manifold developed visualization techniques for 3D medical visualizations. To provide binocular parallax, a special stereoscopic display technology is required, since a left- and right-eye image of the 3D scene with correct geometric properties, e.g., varying binocular disparity depending on the viewer's distance to the objects, must be generated and transferred to the viewer [21].

Over the last decade, various 3D display technologies that enable intuitive 3D visualizations, navigations and interactions with 3D scenes have been developed and evaluated, e.g., volumetric displays, stereoscopic displays or augmented reality systems. Most of them use the conventional binocular parallax combined with motion parallax to provide a viewer-customized and view-dependent stereoscopic visualization and interaction. Holliman et al. [68] refer to a display technology that combines both, and is therefore the most complete display type, as being *full-parallax*. Two images acquired from two slightly shifted camera positions with geometrically adjusted frustums produce an important depth perception of the

presented scene. This binocular parallax is combined with motion, e.g., derived from head-tracking, and therefore, provides the possibility to look around objects in 3D visualizations. Thus, view-dependent views for each eye position are presented. The more depth cues are supported by the display technology, the more realistic is the perception of the virtual reality and leads to an increasing sense of presence. This perception of being physically present in a non-physical world is called *immersion* [21]. In computer graphics, perceptually driven research has the longest tradition in immersive virtual reality, where the users' response to specific interaction and rendering techniques is evaluated using a variety of methods [13]. Focus and context perception [78], visual cues [135], spatial judgments [74] or scene complexity [149] as well as stereovision or motion parallax [113] are a few important aspects that are analyzed in virtual reality environments.

Since a lot of systems are available, it is necessary to analyze which are suitable for medical visualizations and thus for integration in the clinical routine. A clinical workplace and thus a clinical usage of a display technology is restricted by the required space and expenses. Additionally, an application in the operating room requires the possibility of sterilization and sterile control of the system. Thus, volumetric displays such as a CAVE [38] are not suitable for clinical purposes.

One display technology that is small, affordable, full-parallax and immersive is a head-mounted display such as the Oculus Rift from OCULUS VR®, where a pair of micro displays (binocular) and matched enlarging optics (e.g., lenses or mirrors) are used to generate a finite distance virtual image [21, 68]. A 3D stereoscopic visualization is directly presented in front of the viewer's eyes using two small screens, see Figure 3.20. Several experiments were performed using those head-mounted displays or see-through displays for intra-operative support by projecting the patient's image data directly onto the glasses and prevent a frequent switch between patient and image data [129, 175, 180]. The display was

Figure 3.20: The Oculus Rift as an example for a binocular head-mounted display.

used to provide patient-specific image data while the physician performed the intervention, e.g., an ultrasound-guided peripheral nerve block and an MRI-guided puncture simulation. Wendt et al. [175] investigated a binocular head-mounted display (HMZ-T2, resolution: 1.280×720 pixels) and their developed integrated image data monitoring system as a navigation support for transurethral resection of the prostate under transrectal ultrasonography.Multiple information can be presented on the head-mounted display and shared between multiple head-mounted displays in the operating room. The results showed that the displays represent a direct, intuitive guidance for interventional procedures and that the intervention

completion time was statistically significantly reduced compared to conventional procedures.

Wagner et al. [167] compared the EndoSite 3Di Digital Vision System[1] that couples a 3D view with a head-mounted display consisting of three liquid crystal displays per eye (HDI-SDI, 1080i monitor) with a 2D laparoscopic system from Storz[2] and the DaVinci Surgical System[3]. 34 participants with different levels of laparoscopic experience were recruited and had to perform three tasks using these three techniques (open, laparoscopic and robotic surgical). This evaluation followed a within-participant design, since all participants performed the three tasks in an identical sequence under identical conditions in 3D and then in 2D with all techniques. While the first and the third task used a scene consisting of simple geometric objects as stimulus, the second was a real laparoscopic task (sewing). The chosen task required two-hand coordination and ambidexterity that is a good indicator for testing depth perception. Immediately after each task, the participants were asked to rate the task's difficulty using a VAS scale from 1 (very easy) to 10 (extremely difficult to perceive). Thus, task completion time and personal preferences were analyzed. Their quantitatively measured results confirmed the recorded qualitative opinions that the task is more difficult, and thus the performance slower using 2D than 3D display systems. Moreover, the increased time in 2D is correlated to the task's degree of difficulty and not to the used technique or laparoscopic experience. An effective 3D optical system would facilitate advanced laparoscopic surgery and increase performance by 60–70% [167].

However, 3D visualizations of the individual image data were neither generated or integrated nor evaluated. Furthermore, it is somewhat difficult for such devices to obtain acceptance, since they deviate from the well known display and mouse-keyboard interaction principle. The real world and all additional information are seen through an unnatural bulky glasses-alike device. Moreover, if complications arise, the surgeon has to take off the glasses and valuable time is lost. Thus, further extensive research is required.

Stereoscopic displays are another technology that fullfills the requirements of a clinical setting. They are small, affordable, semi-immersive, provide binocular and motion parallax and can use the known mouse-keyboard interaction principle. These displays are either *glasses-based* (stereoscopic displays) or *glasses-free* (autostereoscopic displays).

GLASSES-BASED stereoscopic display systems can be divided into *active* and *passive systems*:

- **Active glasses** called *shutter glasses* are synchronized to open and close their shutters coordinated to the image display rate of the screen, where the images generated for the left and the right eye are shown rapidly alternated [21]. For each rendered image, the glasses block one eye's view, so that the image generated for the left eye can only be seen by the left eye and vice versa. To maintain this synchronization, infrared signals are used.

1 Viking Systems Inc.
2 www.karlstorz.com
3 Intuitive Surgical Inc.

- **Passive glasses** use either *spectral* or *polarization multiplexing filters* [21]. Probably the best-known spectral multiplexing display technique are anaglyph glasses and 3D images. Two different colored images, commonly red and cyan, are generated and slightly shifted displayed on a 2D display. When these images are viewed using the anaglyph glasses that use a red and a cyan-colored filter for the left and the right eye, the same color as the filter's color is washed out and each eye only receives the image colored in the opposite color as the filter's color. Polarization multiplexing glasses filter the two overlaid images that are presented on a 2D display using oppositely polarized filters that can be linearly or circularly polarized [21]. Each eye sees a different image, since each filter passes only light that is similarly polarized and blocks the light polarized in the opposite direction.

All these different glasses and systems enable the viewer to receive and filter one image for each eye out of two presented images on one 2D display. Since disparity requires real depth or two images generated as if from different positions like our eyes, the glasses enable the viewer to see images slightly shifted and thus emulate disparity. The human visual cortex of the brain receives these images and merges them together to perceive a 3D scene. If markers are attached to the glasses, head-tracking is possible and thus motion parallax is integrated as an additional source of depth information.

AUTOSTEREOSCOPIC displays are glasses-free systems that generate stereoscopic 3D images without the need of specific glasses. Besides volumetric and holographic displays, the here discussed autostereoscopic displays use parallax barrier or lenticular lens arrays as the directional optical element, placed in front of the display. This enables each eye to see a different set of pixels and thus depth is perceived through parallax. The slits (parallax barrier) and gratings or arrays (lenticular lens) are aligned vertically or cylindrically to allow the viewer to see only left image pixels from the position of the left eye and right image pixels from the right eye [21, 68]. Therefore, the two rendered and presented images have to be displayed within different pixel sets. Slits and vertical gratings enable one eye to see odd and one eye to see even pixel positions. A cylindrical lens array approach directs different images to display subzones that are projected to the viewer at different angles [21]. Thus, half of the display is visible from the left and half from the right eye.

Theoretically, an autostereoscopic display is more appropriate for medical applications, especially for integration in the clinical routine. It is more natural, since no device is necessary to perceive the stereoscopic visualization. However, the head position in front of such a display is very restricted to receive the correct image for each eye and since the two images are displayed pixelwise, only half of the screen resolution is achieved for the visualization. To generate an optimal stereoscopic perception, the software, hardware and mechanical components (e.g., the parallax barrier, the lens grating or array) have to work accurately. In the past, 3D displays suffered from negative side effects of stereopsis. Viewers either complained about perceptual deficiencies such as the perception of double vision (effects of crosstalk) and loss of contrast or physical effects like nausea and headache [160, 133]. The latest 3D imaging systems provide improved image quality and resolution, sim-

ilar to 2D monitors. In contrast, a glasses-based system enables a very accurate stereoscopic perception and head-tracking. However, a glass is required to view the 3D visualization. Thus, in several experimental evaluations 3D displays including shutter and polarized glasses as well as autostereoscopic systems were investigated and compared for medical applications.

Laparoscopic tasks, such as suturing and knot tying (psychomotor skills) were studied many times, since the development of 3D laparoscopic monitors is an increasing research area, starting with the first 3D video system in 1992 [25, 80, 152, 176]. Assistance during the surgery is accomplished by navigation systems that enable the surgeon to perform more precisely and support safer interventions by indicating structures or regions at risk. The growing improvement is based on the technology development to show and magnify a small body region on a 3D display and enable an intuitive 3D orientation during surgery. In several evaluations quantitative and qualitative aspects of 3D compared to 2D vision systems for minimal-invasive interventions were investigated. Already Buess et al. [25] found that surgeons performed faster and made 43% fewer errors using a 3D vision system compared to a 2D vision system. They concluded that tasks can be performed faster and safer using 3D vision, especially for more complicated surgical procedures. However, 3D displays are still not widely accepted in surgery. Even though their technology enables binocular vision without specific visualization techniques and, thus, supports depth perception that is an essential aspect especially for minimally invasive interventions where two-hand coordination is required while at the same time the scope and field of view is very restricted. In the past, negative side effects of stereopsis were observed, e.g., blurring, headache or nausea that outweighed the advantages. Thus, evaluations investing new technologies followed to analyze the potential and the negative effects [96].

Feng et al. [51] compared the impact of a conventional laparoscopic 2D monitor system, a high-definition monitor system (HD), and a stereoscopic display on the task performance in laparoscopic training. Their 27 participants (15 female, 12 male) including six medical experts and 21 non-physicians had to move surgical instruments to a set of targets using all display systems. Even though a strong preference for HD systems was expressed by the participants, the actual quantitative analysis indicated that HD displays offer no statistically significant advantage and may even worsen the performance compared to standard 2D or 3D laparoscopic monitors. Recently, Storz et al. [152] and Wilhelm et al. [176] presented experimental evaluations investigating 3D vision and displays for laparoscopic tasks. Both used a laparoscopic 3HD video system consisting of a laparoscope, a stereoscopic camera and a stereoscopic display. These image data were presented to the surgeons on different stereoscopic 3D displays and compared to commonly used 2D displays. As shown in Figure 3.21, phantoms were used as stimuli and realistic laparoscopic tasks were performed by participants with different levels of laparoscopic experiences, e.g., medical experts or surgeons-to-be. Thus, the stimuli and the selection of participants corresponds to the target medical application and is a fair reflection of the target population.

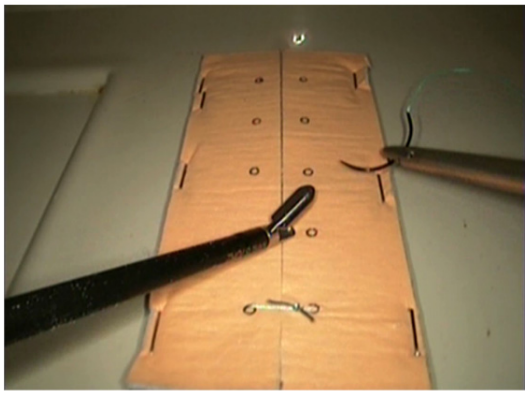


Figure 3.21: Participants had to place parallel stitches through predefined marks on a phantom model under laparoscopic conditions. (Image reprinted from Wilhelm et al. [176] © 2014 Springer Science+Business Media New York with kind permission from Springer.)

All evaluations follow a within-participant design and thus, the participants had to perform the task using each of the provided apparatus. 20 students and 10 laparoscopically experienced surgeons performed five tasks in the evaluation presented by Storz et al. [152]. All participants performed the task in 2D and in 3D with a resting of 48 hours between each modality. They were instructed using a video instruction and initially had to perform a stereo vision and a stereo adaption task (to ensure a full stereoscopic perception) before the evaluation started. Three of the five tasks were simple positioning and movement tasks and only two tasks were realistic laparoscopic tasks encountered in typical surgical procedures [152]. In contrast, Wilhelm et al. [176] presented a more extended comparative evaluation of four different display technologies. 48 participants were recruited, with half of them being experts and half being less experienced. A polarized glasses-based[4] and a mirror-based 3D display[5], an autostereoscopic display[6] (lenticular lens technology) and a 2D HD display[7] were compared with respect to task completion time, stitching accuracy and personal preferences. The accuracy was differentiated between stitch position in a predefined center, within a region or outside. Additionally, the movements and required pathlengths for the surgical instruments were recorded. The results listed in Table 3.1 show that the participants made fewer errors in 3D and required less time to complete the tasks in both evaluations. Thus, 3D displays are superior to comparable 2D displays [176]. 3D stereoscopic visualizations improved the task performance by almost 20%. The qualitative evaluation was divided into display usability and visual comfort. The results showed that the participants preferred the stereoscopic visualizations and Wilhelm et al. [176] showed that a glasses-based display was preferred, followed by a mirror-based and an autostereoscopic display.

Storz et al. [152] found that the experienced surgeons saved more time on the difficult tasks, which was confirmed by Wilhelm et al. [176], who found that the

4 Sony LMD 2451MT, Sony Corporation
5 custom-built mirror by Fraunhofer HHI, Heinrich Hertz Institute, Berlin
6 Fraunhofer HHI, Heinrich Hertz Institute, Berlin
7 24″ display, WideView, SC-WU24-A1511, Karl Storz

System	Procedure time (s)	Suturing precision (pts.)	Index score (pts./min)	Pathlength needle holder (cm)	Pathlength grasper (cm)
S1 (2D HD)	330/387	32.6/31.1	6.8/5.6	1,042/1,236	1,100/1,223
S2 (3D glasses)	**278*/330**	33.3/32.2	**8.0*/6.7**	931/1,167	1,047/1,112
S3 (autostereoscope)	309/412	32.4/30.7	6.8/5.0	955/1,226	1,030/1,147
S4 (custom mirror)	**247*/308***	**34.4*/33.2***	**9.3*/6.9***	**720*/921***	876/1,115

[a] Values are presented for experts/novices

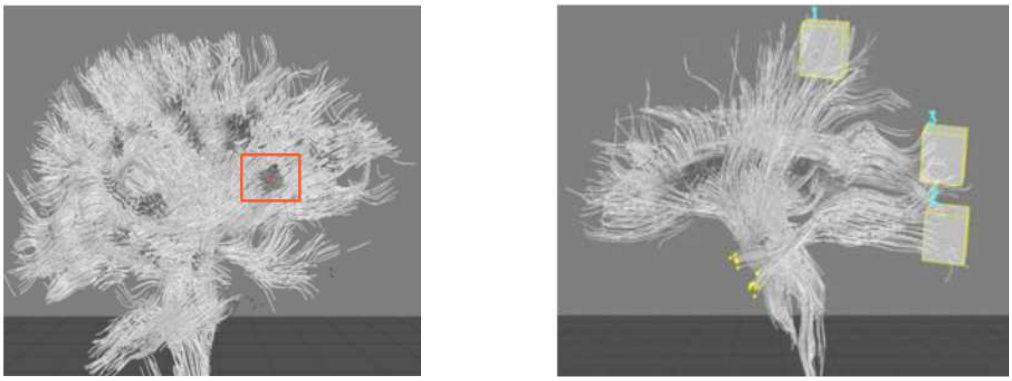
* Significant results. Analysis for significance was performed with respect to the experience level

Table 3.1: A few results for experts and students (novices). All participants were faster with the glasses-based 3D display (shorter procedure times), made fewer errors (high achieved index score) and required shorter paths. (Table reprinted from Wilhelm et al. [176] © 2014, Springer Science+Business Media New York with kind permission from Springer.)

results for inexperienced surgeons using 3D were similar to the results of experienced surgeons using 2D. However, virtual training with 3D visualizations and stereoscopic displays will not make a medical novice an expert surgeon. Overall, Wagner et al. [167], Wilhelm et al. [176] and Storz et al. [152] found in their experimental evaluations that the improvement is between 19% and 88% in performing different expert levels when changing from 2D to 3D vision, yield to an important assistance in the operating room as navigation support and for education purposes to virtually train and gain surgical skills. Although these studies provide important details, they exhibit limitations in the study design, conduction and analysis. For example, the evaluations of Wagner et al. [167] were designed such that the participants performed with the same stimuli presented in an identical sequence on each device without rest between the devices. Thus, learning and recognition effect are present and bias the results. Wilhelm et al. [176] provided no training with the individual 3D displays and Storz et al. [152] compared the 20 results of a student group with 10 surgeon results, which is a poor comparison.

Overall, all authors described statistically significant results without describing or presenting postulated hypotheses. Their design description sounds more like an exploratory evaluation.

Although binocular and motion parallax have been studied extensively, the effect of those depth cues and illustration techniques individually and in combination on the visual perception of medical visualizations has not yet received enough attention. Chen et al. [32] evaluated the effect of stereo and screen size (24″ and 72″ display) on the task performance in diffusion magnetic resonance imaging with the 12 medical experts (6 female, 6 male). The results showed that a larger display neither accelerated nor improved the performance accuracy. While participants saw four datasets, they had to perform five tasks such as locating, tracing and naming fiber bundles and the required time, the achieved accuracy and afterwards the subjective workload was recorded (see Figure 3.22). Although all participants were overwhelmed by the large display combined with the stereoscopic visualization, no advantage due to the size could be quantitatively confirmed. Moreover, they achieved more accurate and faster results with the monoscopic view on the small and large display compared to the stereoscopic view. Even though,



(a) (b)

Figure 3.22: 3D DMRI visualizations for which the participants had to (a) locate a fiber bundle by right-clicking or (b) trace the bundles. The yellow spheres mark the beginning and the participants had to specify the box in which the corresponding ending points lay. (Images reprinted from Chen et al. [32] © 2012 IEEE with kind permission from IEEE.)

this difference was statistically not significant. Chen et al. [32] concluded that this result is due to some intrinsic drawback of the stereoscopic display (e.g., a darker visualization) and that such negative stereo performance may be due to cue conflicts. There is a mismatch between the natural convergence and accommodation caused by the difference between the perceived position of the structures (in front of or behind the display) and the real origin of the structures. A closer structure is perceived larger and the more distant internal lesion areas were perceived smaller with increasing distance in the stereoscopic view and hampered the visual search compared to the monoscopic view. One solution might be to alter the illustration technique to facilitate the visual perception of the internal structures.

Another evaluation investigating medical visualizations and stereoscopic displays qualitatively and quantitatively will be presented in this thesis in Section 7. Based on the overwhelming results of the used zSpace technology[8], Saalfeld et al. [139] presented an approach for a detailed cervical vertebra inspection in a 3D cervical spine visualization based on an indirect volume rendering and qualitatively evaluated this using the zSpace technology. They used the *semantic depth of field* methodology that was previously investigated by Kosara et al. [90], who concluded that this is a very effective depth cue and method for guiding the viewer's attention. Saalfeld et al. [139] extended this approach by implementing an animated structure categorization into focus and context. Spine surgery requires difficult path planning tasks to prevent injuries of the spine canal and at the same time to achieve the best vertebral or intervertebral disc access. As shown in Figure 3.23, a selected cervical vertebra yields an increasingly blurred visualization of the cervical spine (context) until the focus vertebra is replaced animated. Adapting the illustration technique combined with an animation is one approach to divide and present the semantically defined focus and context and guide the viewer's attention. A qualitative analysis using a questionnaire with 5-point Likert scales

8 A 3D virtual imaging display (1920 × 1080 pixels full HD) with a passive circular 120 Hz stereo 3D polarization technology including an optical tracking developed by zSpace Inc.

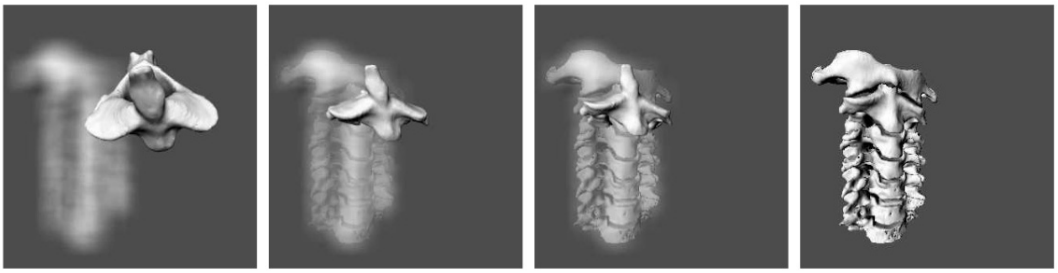


Figure 3.23: An indirect volume rendering of a patient-specific dataset of the cervical spine with single cervical vertebra. A selected cervical vertebra is the focus and explored in detail. The context (cervical spine) is visualized blurred. The blurred visualization of the cervical spine (context) is simultaneously reduced when the focus vertebra is replaced. (Image reprinted from Saalfeld et al. [139] © 2015 Springer-Verlag Berlin Heidelberg with kind permission from Springer.)

(from $--$ to $++$) was performed with nine stereoscopic display experienced participants. The analysis was divided into stereo, head-tracking, stylus interaction, blurring and animation impression and the results for each category were between $+$ and $++$. Especially the stereo impression was rated very good and when the participants got used to it, they liked the head-tracking exploration technique that provides a motion parallax cue. The enormous personal preference exhibits the high potential of the zSpace system.

3.4 CONCLUSION

2D and 3D visualizations based on 2D image, 3D volume and 4D (time-varying 3D data) data are used for several areas of medical application. These visualizations enable a customized and optimal diagnosis and disease characterization, treatment planning and assessment of the therapy success. Moreover, documentation, medical education and training benefits from illustrative visualizations to gain and improve medical and surgical skills. Since each visualization must be generated for a special medical task, customized visualizations are required that support the data exploration, interpretation and decision-making process. The major requirements for those customized visualizations are the accurate data representation (accuracy), the application specificity and the effective information transfer. Many illustration methods and techniques have been developed and refined for the visualization of medical volume data and segmentation information. However, to depict the essential information and facilitate the exploration process, depth and spatial perception have to be supported and the viewer's attention must be guided to the essential information. Without perceptual guidance, it is difficult to decide which techniques should be used for particular applications, how they should be combined and how parameters should be adjusted. Thus, human visual perception plays an important role for the development of 3D medical visualizations either for the integration of depth cues into 3D visualizations or for the development of stereoscopic views. Besides the development of sophisticated illustration techniques that enable a focus-and-context visualization and facilitate the explo-

ration, stereoscopic views provide another possibility to illustrate patient-specific data. Depth and spatial perception is achieved by binocular parallax and improved with motion parallax.

Overall, a perception-guided evaluation is important to ensure accuracy, visual guidance, to support the viewer's perception and analyze the visualization's effectiveness for a specific task. Task performance, subjective preferences and work practices are analyzed to evaluate visualizations, established techniques and support further refinements by identifying limitations. However, only a few existing evaluations adapt psychological guidelines to design evaluations that minimize bias factors, maximize the isolation of the measured effect while still analyzing medical data and applying real-world tasks to achieve valid, reliable and reproducible results. Additionally, complex visualizations require individual and detailed quantitative and qualitative measurements to identify advantages and limitations and combine objectively measured criteria with subjective preferences.

Part II

MAIN CONTRIBUTION

4

A QUANTITATIVE EVALUATION OF THREE ILLUSTRATION TECHNIQUES FOR THE EMPHASIS OF LYMPH NODES

This chapter is based on the following publication:

"Perception-Based Evaluation of Emphasis Techniques Used in 3D Medical Visualization". Alexandra Baer, Friederike Adler, Daniel Lenz and Bernhard Preim. In *Proceedings of Vision Modeling and Visualization*, pp. 295-304, 2009

A widespread method to investigate the relationship between physical stimuli and the perceptions they affect are visual search task experiments, see Section 3.2.1. Eye-tracking experiments integrate search tasks to analyze the users' scanpaths and viewing strategies. Viewing strategies of medical experts exploring patient-specific data [101] and the resulting attentional landscapes [26] during the diagnostic and the decision-making process [93] are perceptually analyzed. Besides eye-tracking experiments, visual search tasks are appropriate to perceptually evaluate illustration techniques used for focus and context visualizations. The technique's effectiveness to guide the users' attention to the focus structures or objects can be quantitatively analyzed. Kosara et al. [89, 90] examined the semantic depth of field illustration technique to illustrate that this technique is preattentively perceived and thus suitable for focus and context visualizations. Waldner et al. [168] compared a flicker, a spotlight and a halo technique within a search task experiment. In both presented studies the stimuli scenes consisted either of simple geometry (dots with different colors) or were easy to understand (chess board, landscape maps, and text editors).

Baer et al. [6], the author of this thesis, performed a visual search evaluation combined with the signal detection theory to compare three emphasis techniques applied to lymph nodes in 3D patient-specific neck visualizations derived from CT datasets. The required anatomic and pathologic structures are visualized to explore the patient-specific datasets and to answer all relevant therapeutic questions required for the diagnostic and therapy planning process. Since there are a lot of relevant structures to assess the focus structure and the risk structures, an appropriate visualization is required to ease and fasten the exploration process of such complex medical scenarios. Thus, guidelines and findings from psychophysical experiments are adapted to a common therapeutic question of enlarged lymph node detection within patient-specific data visualizations. The major challenge using medical visualizations as stimuli is the structure's individuality and the arrange-

ment of the structures. Anatomic structures are located very close and occlude each other. Moreover, there are several similarities between shape and representation of focus and context objects. Therefore, a few modifications are necessary to generate an appropriate stimulus visualization for an experimental evaluation and to perform an adequate conjunctive search evaluation. A trade-off between rather complex realistic clinical visualizations and rather simple appropriate psychological stimuli is required, as described in Section 4.4.2.

This chapter introduces a visual search task experiment to analyze and compare the effectiveness of three illustrative emphasis techniques. Therefore, we investigated the techniques' ability to guide the users' attention. Further on, to validate the first evaluation results regarding generalization for lymph node detection, the evaluation presented by Baer et al. [6] was extended within this thesis primarily by a follow-up study using thorax datasets and a detailed analysis.

4.1 MEDICAL BACKGROUND

Neck dissection or thoracic surgeries are frequent parts of treatment of patients with malignant tumor in the neck or lung region. The extent of the interventions depends on the existence and location of malignant lesions. It is necessary to specify the patient-specific tumor node metastasis classification (TNM) to decide about operability and to define the individual surgical procedure. The TNM classification system describes occurrence and the extent of the malignant tumor, suspicious lymph nodes and distant metastasis to provide a cancer staging and thus to enable a description and categorization as well as a prognosis [177]. Besides distant metastasis, the tumor and the occurrence of enlarged lymph nodes as well as their location in relation to risk structures need to be assessed. Enlarged nodes might be malignant and thus have to be removed, since tumor cells primarily metastasize in lymph nodes that cause a lymph node enlargement.

Baer et al. [6] focused on the enlarged lymph node detection as a part of the TNM classification and therapy planning process. Three relevant nodal stages indicate a pathologic risk for lymph nodes. Nodes larger than 1, 3 or 6 cm are suspicious and have to be analyzed [177]. Especially the detection of nodes between 1cm and 3cm is very difficult, since a lot of distracting structures occur in the neighborhood of other normal-sized lymph nodes smaller than 1cm. To support the preoperative planning and thus the lymph nodes' identification, a supportive emphasize visualization of the suspicious ones is necessary.

4.2 VISUAL SEARCH AND SIGNAL DETECTION THEORY

Several factors influence where the attention is guided when observing an image or scenario. As introduced in Section 3.2.1, a general and influential theory for visual perception is the feature integration theory (FIT) introduced by Treismann and Gelade [158]. According to the FIT, visual perception is characterized by two processing stages. The first stage is called the *low level processing* stage, where different features are analyzed preattentively and the second stage is called the *higher*

level processing stage, where attention is required to recognize and classify the perceived features and objects.

Search task experiments are conducted to evaluate visual search performance and thus to analyze the attention guidance ability of specific image features or in the evaluation of [6] the effectiveness of emphasis techniques. A visual display comprising a number of elements to be searched, called *targets*, was shown to the participants. They had to determine whether a target element is present or absent in a field of background distractor elements that were more or less similar to the target. All images are shown randomly, and in half of the images a target is present alongside one or more distractor elements. Such experimental studies focus on *feature search tasks* and *conjunctive search tasks*. The feature search is characterized by the detection of a target that usually differs from the distractors in one feature. In contrast to that, a conjunctive search experiment requires the combination of features, for example of color and shape, to detect the target. Thus, the performance is dependent on the presented display size characterized by the number of distractors.

However, such visual search experiments require motivated, attentive and concentrated participants over the entire time. As participants are often moody, subjective and differently motivated, Green and Swets [62] introduced the *signal detection theory* to validate their response tendency as well. They suggested to differentiate between *the sensitivity* and *the psychological bias* of the participants by examining the correct responses and the false alarms. The sensitivity or discriminability refers to how hard or easy it is to detect that a target stimulus is present from background or distractor elements. To apply the signal detection theory, target and noise stimuli are required where the target is present and absent and the participants have to categorize each stimulus. A false alarm occurs once the participant reacts on a signal which was not given when presenting a noise stimulus. A miss occurs when the participant does not react on the target stimulus that was shown. Otherwise, the participants correctly rejected when a noise stimulus was presented or correctly detected the target within the target stimulus. Further on, the theory can be used to assess the detectability of a target from a background of distractor elements.

4.3 EMPHASIS TECHNIQUES

This evaluation was designed to perform a controlled perceptual evaluation using realistic visualizations and applying psychophysical theories. Besides that, a real-world task – enlarged lymph node detection – from the clinical workflow was used. A common motivation for the illustration technique development is the technique's ability to guide the users' attention to a specific focus region. This region might be a specific part of the visualization or a specific focus structure. Furthermore, illustration techniques are developed as a kind of *focus-and-context* visualization. The viewer's attention is guided to the focus to enable a detailed focus inspection. The context is sparsely visualized to provide additional information and for orientation purposes. Emphasis techniques modify the object's

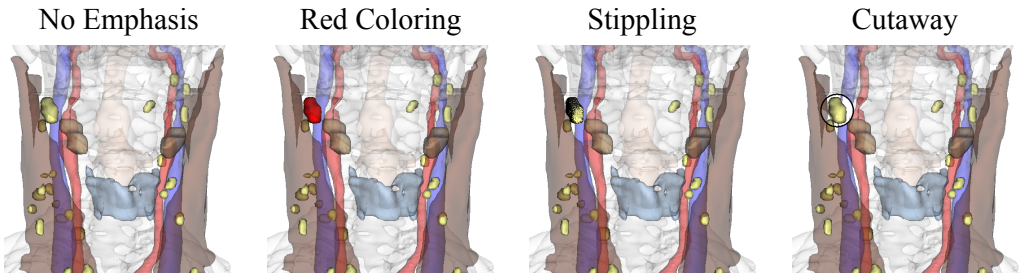


Figure 4.1: Three different visualization techniques were applied to one lymph node and their attention guidance effectiveness was analyzed compared to no emphasis.

appearance such that the object can be clearly recognized and its location in the overall anatomical region becomes obvious [128].

RED COLORING. The most popular visualization technique to emphasis important structures is coloring. A color contrary to the other used colors is applied to emphasize special regions and thus achieve an attention guidance. Since red is a signal color with a high attention guidance potential, red coloring was favored and investigated within this evaluation as an emphasis technique for the lymph node detection (see Figure 4.1).

STIPPLING. Pen-and-ink illustration techniques, e.g., hatching, stippling or feature lines are inspired by traditional medical applications and anatomic atlases. These techniques are rather used for context structures and often investigated within evaluations on sparsely presenting information without disturbing the viewers attention [155]. Several experimental studies focus on the illustration potential for the presentation of anatomical structures [141, 154, 103]. Stippling as a representative pen-and-ink surface visualization technique was chosen to be the second emphasis technique. A stippling visualization of an enlarged lymph node was generated with the object-based stippling method of Baer et al. [5].

CUTAWAY. The third emphasis technique is based on the work of Krüger et al. [92], who presented illustration techniques to facilitate lymph node detection and classification for pathologic lymph nodes in the neck region. We applied cutaway as a smart visibility technique. The cutaway view was defined as a circular region that is aligned with respect to the enlarged lymph node to enable an unobstructed view. The size of the region was defined according to the projected lymph node size (maximum node extension = circle diameter). All occluded structures or structure parts, respectively were removed within this circular region.

These three representative techniques were used as emphasis techniques for the enlarged lymph nodes, as illustrated in Figure 4.1. While all other structures were rendered with predefined standard colors and transparency values, enlarged lymph nodes were either visualized with one of the emphasis techniques or similar to the other lymph nodes without being emphasized. Thus, the individual illustration technique is systematically varied to measure the caused perception and behavior, compare Section 2.2.4. We were able to quantitatively compare cutaway views as a kind of smart visibility technique, stippling as a pen-and-ink technique,

red coloring as a typical emphasizing technique, and the normal yellow lymph node visualization with no special technique.

4.4 EXPERIMENTAL DESIGN

Two visual search task experiments were performed to analyze and compare illustration techniques in 3D visualizations of patient-specific anatomy and their attention guidance effectiveness. An enlarged lymph node is the focus object and serves as the target visualized with one of four illustration techniques. All other anatomic structures are the context. In our visual search task experiments, the context objects were distractor elements and can be treated as noise corresponding to the FIT, respectively. Since the participants had to search for a specific structure (lymph node) with a specific size (enlarged), a *conjunctive search* for the two features structure $\times$ size was performed. The results were validated using the signal detection theory.

In detail, we performed two conjunctive search experiments for rather spherical structures in medical visualizations. Both evaluations followed a *within-participant* design (compare Section 2.2.2). They were *one-factorial* (illustration technique) with four factor levels (no emphasis technique, red, stippling, and cutaway). Illustration technique was the *independent variable* and the two measured *dependent variables* for each individual technique were response time and accuracy. According to the techniques' attention guidance capability described in Section 3.2.1, the following one-tailed hypotheses were postulated:

- $H_{Emphasis}$: Emphasized enlarged lymph nodes are detected more often and faster than those without emphasis.
- $H_{Cutaway}$: Cutaway views will be more effective regarding accuracy and response time than stippling and red coloring.

Accuracy was defined as the number of correctly detected focus structures and the number of "hits", respectively. This visual search experiment is an evaluation based on decisions. In detail, there were two possible answers: yes there is an enlarged lymph node or no there is none. Based on the signal detection theory, target and noise stimuli were required that will be described in Section 4.4.2.

Additionally, to validate the response tendency, the number of false alarms that indicate the technique's usability was measured. A false alarm occurs if a participant detects a target and thus an enlarged lymph node, when only normally sized nodes are present. Moreover, the target detection capability of the illustration technique was analyzed. To validate whether the achieved accuracy correlated to the individually preferred and subjectively perceived most effective technique, each participant was asked to order and thus rank the presented techniques.

4.4.1 *Participants*

A pilot study with seven participants for the neck (three women and four men aged between 25 and 35 years with $\bar{x} = 29.14$ years) and three subjects (two women

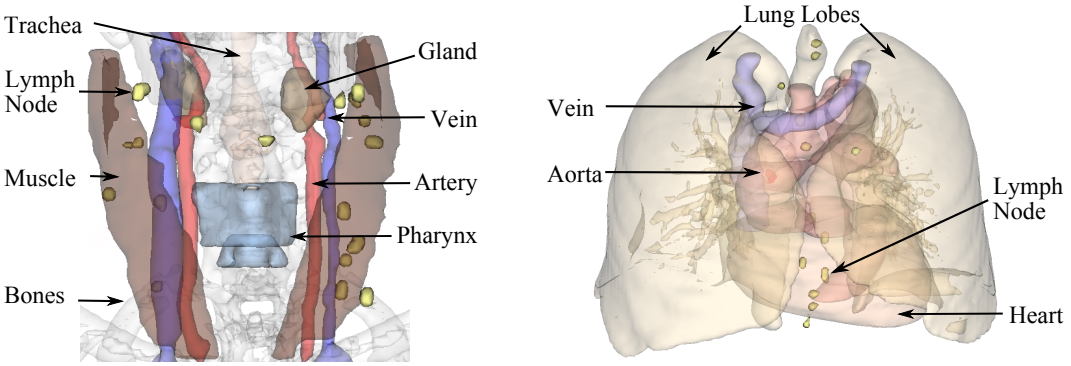
and one men aged between 20 and 28 years with $\bar{x} = 23.66$ years) for the thorax experiment was conducted in advance. We recruited 33 participants from various parts of the university, like psychology and engineering students, designers and a few medical experts for the first evaluation. In detail, 18 women and 15 men aged between 19 and 51 years ($\bar{x} = 31.90$ years) participated in the first evaluation and were asked to detect enlarged lymph nodes in neck visualizations. Four of them were experienced users (three medical experts and one a medical student), nine with a passing knowledge and ten participants without experience in medical visualizations. Eleven participants performed the follow-up evaluation and had to find enlarged lymph nodes in the thorax visualizations. Five of them were women and six were men aged between 23 and 31 years with an average of $\bar{x} = 25.72$ years. Four of them were experienced with 3D medical visualizations and had advanced medical knowledge (one medical expert and two students), five with medium and passing knowledge and two without experience.

4.4.2 *Stimuli*

Since the goal of Baer et al. [6] was to perform a controlled experiment which enables a wider set of conclusions, realistic 3D patient data was employed. 3D patient-specific anatomy visualizations derived from CT neck and thorax datasets are used in the clinical routine to determine the size of the lymph nodes. Common psychological user studies present simple scenarios as stimuli containing letters or basic geometric shapes that differ strongly and that are equally distributed over the display [158, 91]. A major condition of controlled experiments and especially search tasks is that stimuli are similar to reduce errors, see Section 2.2.4. Hence, a trade-off between the visualizations used in the clinical routine and the psychological conditions was necessary to generate appropriate stimuli for our conjunctive search. Restrictions were made with respect to (1) the presentation, (2) the visualized structures and (3) the field of view to achieve stimuli that were as similar as possible and still realistic and representative illustrations of patient-specific anatomy from the clinical routine. Based on that, static images were generated derived from visualizations showing a restricted number of structures with a defined field of view.

1. Static images with orthogonal projection were used as stimuli, since rotation was not required and unintended rotation performed by the participants should be prevented. Due to that, the stimuli similarity was increased. Moreover, according to the display time of each stimulus, transformations like translation or rotation were not possible. A stimulus was a rendered image of 3D patient-specific visualizations with a displayed size of 512×512 pixel.

2. To enhance the similarity of the stimuli, the patient-specific visualizations were restricted to a representative number of structures. As shown in Figure 4.2a, each neck stimulus showed lymph nodes, muscles, glands, trachea, two veins, two arteries, the pharynx and bones as orientation structures. The thorax stimuli included lymph nodes, heart, vein, the aorta and two lung lobes (see Figure 4.2b). Bones were not required, since the lung lobes enabled an anatomic orientation. Lymph nodes were always on both sides of the bones in neck stimuli and on each lung lobe side in the thorax stimuli.



(a) Neck Stimulus (b) Thorax Stimulus

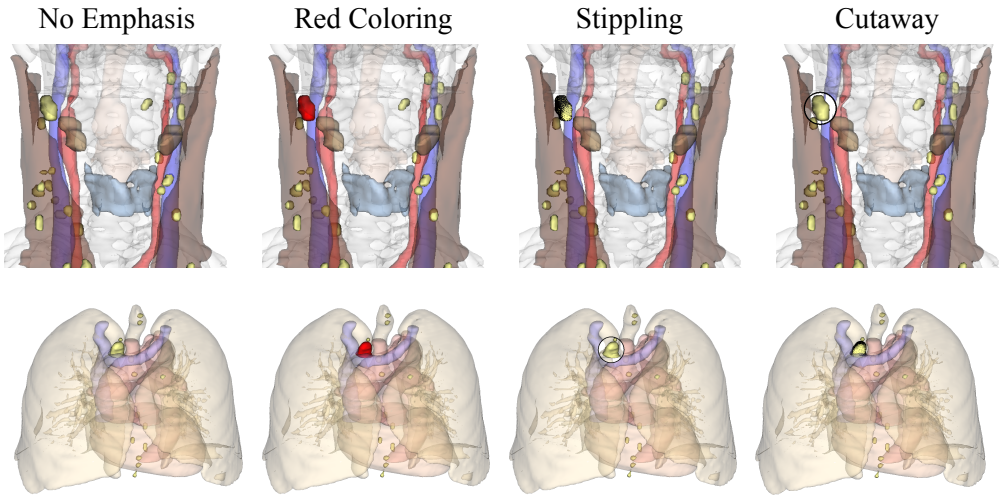
Figure 4.2: Patient-specific neck and thorax stimuli with a representative number of structures. (a) The field of view for the neck stimuli was defined by the muscles and illustrates lymph nodes, muscles, glands, bones, trachea, two veins, two arteries and the pharynx. (b) Thorax stimuli were defined by the lung lobes and show lymph nodes, two lung lobes, heart, vein and the aorta.

Each stimulus showed 12 to 20 lymph nodes. However, they were not equally distributed on both sides due to the patients' individuality.

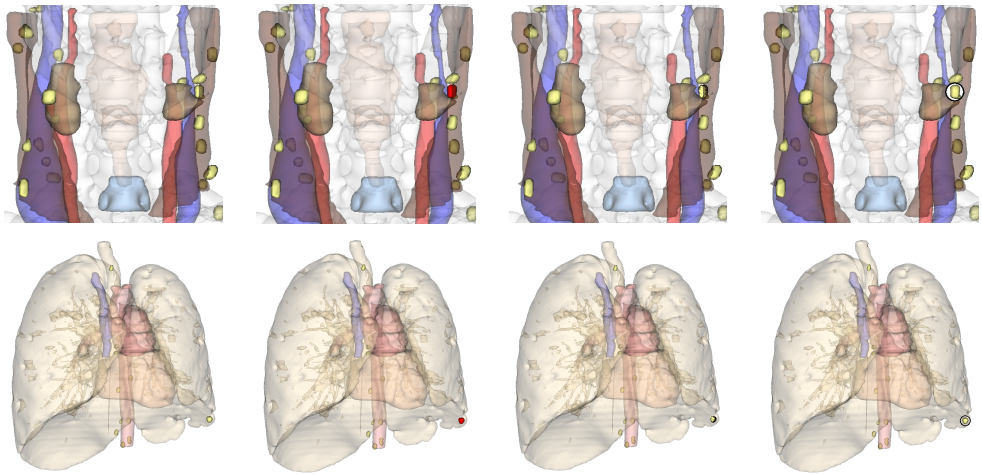
3. The third restriction affects the applied field of view. Neck stimuli were generated using a field of view that was defined by the muscles, while the field of view of thorax stimuli was defined by the lung lobes. These three restrictions enabled results that were traceable to the emphasis techniques and not caused by the stimuli appearance.

To ensure that the lymph nodes were not occluded by other opaque structures, we analyzed the neck and thorax visualizations with respect to the visibility of lymph nodes. If necessary, we adapted the transparency or the saturation of occluders. In the neck stimuli a lymph node is occluded by one translucent structure at the very most. Thus, it was not required to change the appearance of the structure. Moreover, this visibility limitation was accepted for this evaluation, since the stimuli should be realistic, too. In contrast, the thorax visualization required an illustration technique parameter modification. Caused by the anatomic location of the lymph nodes in the thoracic region, they were occluded by several structures like lung lobes, heart and vessels, as shown in Figure 4.2b. As the aorta is very dominant in this anatomic region and occludes the lymph nodes located alongside the bronchial structures, they were visualized using lower saturated red color and higher transparency than the artery vessels illustrated in the neck stimuli. Besides the transparent lung lobes, the heart was visualized translucent, too. Thus, the identification and detectability of the enlarged lymph nodes (target) was guaranteed for each stimulus.

NOISE STIMULI were rendered images of neck or thorax visualizations including the above-mentioned anatomic structures without any enlarged lymph node but normal-sized lymph nodes, as shown in Figure 4.3b. Moreover, to generate noise for the illustration technique, a healthy lymph node may be emphasized with cutaway, stippling or red coloring to provoke false alarm and to evaluate the target detection capability of the illustration techniques. The noise stimuli can be



(a) Target Stimuli



(b) Noise Stimuli

Figure 4.3: (a) Eight target stimuli (neck and thorax) with the enlarged lymph node positioned on the right side visualized with no emphasis, with red coloring, with stippling and with cutaway. (b) Eight noise stimuli (neck and thorax) where no enlarged lymph node was included. The visualized stimuli were generated using four datasets.

considered as permanent noise, since a target stimulus was based on this noise stimulus and included one further additional structure.

TARGET STIMULI included one enlarged lymph node – the target structure – besides several normal-sized lymph nodes that were also included in the noise stimuli. The minimum display size of the enlarged lymph node was > 30 pixels, which represents an appropriate minimum size of an enlarged lymph node referring to the illustrated size of healthy lymph nodes and was comparable to the first nodal stage $> 1\text{cm}$ of the TNM classification. As illustrated in Figure 4.3a, the enlarged lymph node (target) was either visualized equal to the normal sized lymph nodes (yellow colored) with no emphasis technique or contrary with one of the three emphasis techniques. Since each dataset contained more than one enlarged lymph node and more than one lymph node size, several target stimuli were gen-

erated for each dataset showing one enlarged lymph node and the position of the target varied within the stimulus image, too. This increased the randomization of lymph node position and shape.

In summary, this evaluation comprised noise stimuli with no enlarged lymph node and target stimuli with one enlarged lymph node. Both stimuli types included images showing emphasis techniques. In both evaluations, 50% target and 50% noise stimuli were presented. Moreover, in 50% of the stimuli, the target structures were located on the left side and in 50% on the right side of the cervical spine. Overall, 1160 neck stimuli were generated derived from 16 different datasets, to provide a representative sample of the anatomic variety and to avoid that the results are strongly influenced by the peculiarities from one specific patient. The follow-up evaluation comprised 240 thorax stimuli generated of two different datasets.

4.5 APPARATUS AND PROCEDURE

All participants of both studies were tested under the same conditions. The experiment was performed alone by daylight on a 26″ monitor. The evaluation was performed and the data was recorded using the program Presentation® from NeuroBehavioralSystems.[1] This is a stimulus delivery and experimental control program for neuroscience. No other processes were run on the computer during the experimental session.

Initially, the instruction was performed in written form to provide the same initial conditions for every participant. A first practice session followed to ensure that the subjects understood the experimental task and were able to identify and to detect enlarged lymph nodes. This practice session consisted of ten stimuli that were not included in the final evaluation. During the evaluation, each stimulus was presented for 1.1s in the neck and for 1.2s in the thorax evaluation. The participants were asked to perform a conjunctive search and to press the left mouse button if an enlarged lymph node was present or to press the right mouse button if no enlarged lymph node was found. The stimuli were shown for a fixed duration independent of whether the participant pressed a button or not. As conceptually illustrated for three neck stimuli in Figure 4.4, a fixation cross followed each stimulus for a varying time of 0.75s − 1.25s. This is a well-established psychological method to avoid expectations and to promote the subjects' attention.

The presented stimuli were arranged into trials. Eight trials with each presenting 145 stimuli for the neck and – due to the small number of datasets – four trials with 60 stimuli each for the thorax evaluation. Individual rests between the trials were integrated to avoid becoming fatigued during the trials. Participants decided on their own how long they required resting between the trials. The presentation order of target and noise stimuli was random but both stimulus types and the illustration techniques were presented equally often. Afterwards, each participant was asked the same set of questions to gather demographic data (age, gender,

1 www.neurobehavioralsystems.net

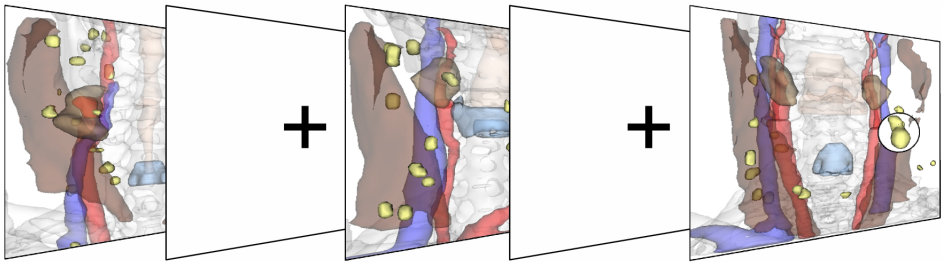


Figure 4.4: The stimuli arrangement for three neck stimuli. The presentation was characterized by a random stimuli order. Each stimulus was presented for a fixed duration. A fixation cross followed and was displayed for a varying time of 0.75s – 1.25s (randomly distributed). Then, the next stimulus is presented.

medical knowledge) and to check for visual impairments, e.g., color blindness or other potential bias factors, e.g., handedness, since all participants used the mouse as input device. Additionally, each participant was asked to order and thus rank the illustration techniques from 1 to 4.

The specific display duration was determined during the pilot study. Durations of 0.9s - 1.4s that are used in common psychological studies with rather simple stimuli were tested. If the stimulus was displayed 0.9s, the participants did not have enough time to search for the target structure and react to the stimulus. If the stimulus was displayed too long, the number of hits was almost 100% and the results were not expressive. Since the thorax visualization included several translucent lymph nodes, occluding and distracting structures, e.g., the transition area of the bronchial structures and the lung lobes, a higher subject attention was required.

4.6 ANALYSIS AND RESULTS

Since target and noise stimuli were shown and possible answers were "yes" (target present) and "no" (target absent = noise), the following discriminations can be derived:

- target stimulus: "yes" = *Hit*; "no" = *Miss*
- noise stimulus: "yes" = *False Alarm*; "no" = *Correct Rejection*

Accuracy (number of hits), response time for the hit results and false alarm results were analyzed to evaluate the techniques' effectiveness for a lymph node detection within neck and thorax visualizations. The number of hits was the number of correctly detected enlarged lymph nodes and thus targets in the presented target stimuli. The number of false alarms was the number of falsely detected target lymph nodes in noise stimuli with no enlarged lymph node included. Initially, the reliability of the recorded data concerning the hits and the false alarm results were descriptively analyzed. A statistical analysis including a Shapiro-Wilk test for normal distribution followed. Based on that, the non-parametric Friedman and Wilcoxon signed-rank test and the parametric ANOVA and t-test were applied and thus the postulated hypotheses were tested (recall Section 4.4). For each participant assessing neck stimuli, there were 580 results for target and 580 recorded

results for noise stimuli. 145 gathered results for each illustration technique presented in the target stimuli and 145 results per technique in the noise stimuli. Additionally, 120 target and 120 noise stimuli results with 30 for each technique were recorded for each participant in the follow-up study. We used the software package IBM SPSS STATISTICS (Statistical Package for the Social Sciences) for the statistical analysis.

4.6.1 *Descriptive Analysis*

We defined a reliable number of hits as a hit result above 50%. Results below were considered as outliers and were discarded, since they deviate widely from the average number of hit results and thus, the participant either did not understand the task or were unmotivated, bored or tired. Due to that, five participants had a bad physiological sensitivity for the neck and one for the thorax experiment. Their hit results were $< 50\%$ combined with a false alarm result of $< 3\%$. Those results were neglected. The measured data of the remaining 28 participants for the neck and 10 for the thorax experiment were used to evaluate the techniques and to examine whether there was a statistically significant difference or not. The achieved false alarm result of the other 28 participants was on average 8% for the neck stimuli that represented a very good response bias and indicated that all results are valid. In contrast to that, the average false alarm result of the 10 reliable thorax participants was higher (16%), but still acceptable.

First of all, the results were summarized as frequency distribution to perform an initial descriptive analysis. The average number of hits and the average response time results are listed in Table 4.1. Participants assessing neck stimuli detected on average $\bar{x} = 53.73\%$ hits when the enlarged lymph nodes were visualized similarly

	Neck			Thorax		
	$\bar{x}$	σ	Shapiro-Wilk	$\bar{x}$	σ	Shapiro-Wilk
No Emphasis	53.73	15.21	$p > .05$	65.17	14.42	$p > .05$
Red	72.10	12.58	$p \leq .001$	93.09	10.38	$p \leq .05$
Cutaway	83.19	10.87	$p \leq .001$	96.66	4.37	$p \leq .05$
Stippling	74.07	12.35	$p \leq .001$	89.17	5.99	$p > .05$

(a) Accuracy

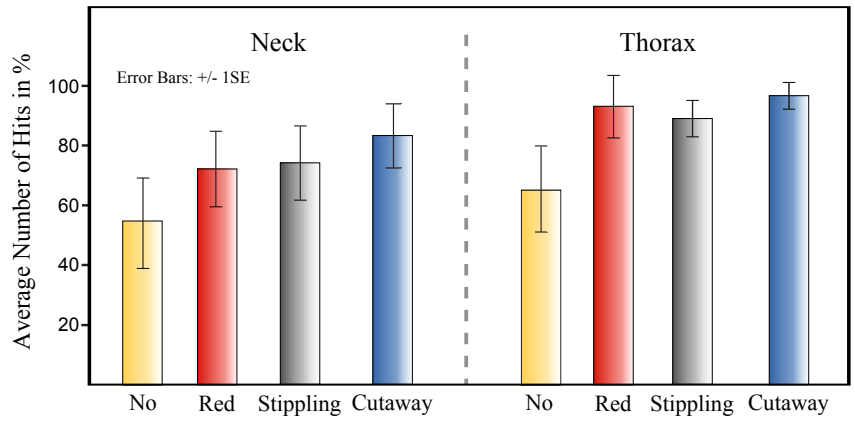
	$\bar{x}$	σ	Shapiro-Wilk	$\bar{x}$	σ	Shapiro-Wilk
No Emphasis	800.09	91.33	$p > .05$	808.69	166.97	$p > .05$
Red	725.43	72.84	$p > .05$	757.51	107.34	$p > .05$
Cutaway	705.11	77.94	$p > .05$	732.25	120.74	$p > .05$
Stippling	726.94	75.48	$p > .05$	788.34	94.11	$p > .05$

(b) Response Time

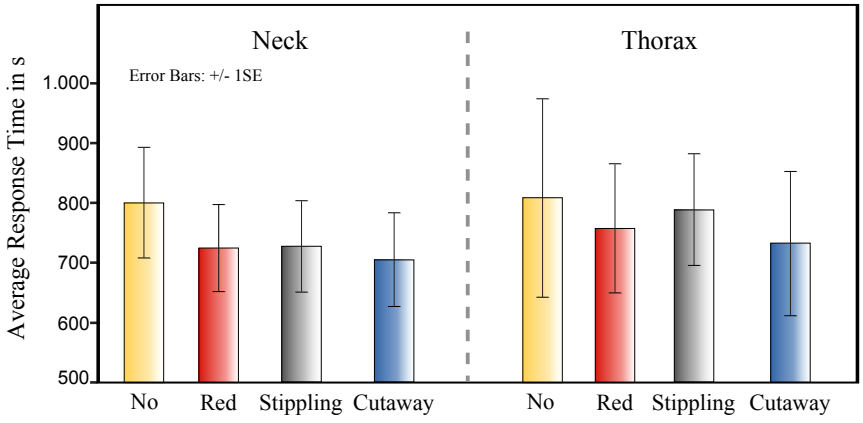
Table 4.1: This table covers the mean value ($\bar{x}$), standard deviation (σ) and the Shapiro-Wilk test results (a) for accuracy in % and (b) for response time in ms. Shapiro-Wilk results with $p > .05$ represent normally distributed results while $p \leqslant .05$ describe a statistically significant difference from a normal distribution.

to the normal-sized nodes. $\bar{x} = 72.10\%$ of all enlarged lymph nodes were detected for the red coloring, $\bar{x} = 83.19\%$ for the cutaway and $\bar{x} = 74.07\%$ for the stippling technique. They required on average $\bar{x} = 800.09$ ms for no emphasis technique, $\bar{x} = 725.43$ ms for red coloring, $\bar{x} = 705.11$ ms for cutaway and $\bar{x} = 726.94$ ms for stippling. Participants of the follow-up evaluation achieved better hit results compared to the neck results. For stimuli with no emphasis technique $\bar{x} = 65.17\%$, for red coloring $\bar{x} = 93.09\%$, for stippling $\bar{x} = 89.17\%$ and for cutaway $\bar{x} = 96.66\%$ of all enlarged lymph nodes were detected. The participants required on average $\bar{x} = 808.69$ms for no emphasis technique, $\bar{x} = 757.51$ms for red coloring, $\bar{x} = 732.25$ms for cutaway and $\bar{x} = 788.34$ms for stippling. However, the thorax results were based on 10 participants, which is less than a half of the number of neck participants.

Figure 4.5a illustrates the accuracy and Figure 4.5b the response time results using bar charts to enable a visual comparison between the techniques. Each same colored bar pair represents one technique. Bars on the left hand side are the neck and on the right hand side the thorax result. The black bars on top of each tech-



(a) Accuracy



(b) Response Time

Figure 4.5: (a) The average number of hits in % and thus the accuracy and (b) the average response time results in ms illustrated as bar charts including standard error bars for the neck and the thorax evaluation.

nique bar are standard error bars that may indicate a statistically significant difference. If two error bars do not overlap, there will be a statistically significant difference between those bars and the represented techniques. If they overlap, it does not mean that there is statistically no significant difference.

4.6.2 *Statistical Analysis*

Since the appearance of scaled bar presentations can be deceiving, a statistical significance test is necessary and was applied to accept or to falsify the hypotheses. The applied Shapiro-Wilk test showed that all hit results for the emphasis techniques of the neck and the results for red coloring and cutaway for the thorax evaluation were statistically significantly different with $p \leqslant .001$ and $p \leqslant .05$ compared to a normal distribution (see Table 4.1). The frequency distributions were left skewed, due to the fact that emphasized enlarged lymph nodes were detected more often. In contrast to that, the hit results for no emphasis technique in both studies and the stippling hit results in the thorax evaluation were normally distributed for the neck and the thorax evaluation. Based on the not normally distributed results, the non-parametric statistical Friedman test was applied to the hit results.

A statistically significant difference exists if $\chi^2 \geqslant \chi^2_{crit}$. The critical value for χ^2 is listed in Table A.1. For $\alpha = .05$ it is $\chi^2_{\alpha,df} = \chi^2_{.05,3} = 7.81$ and for $\alpha = .01$ it is $\chi^2_{.01,3} = 11.34$. The Friedmann tests confirmed the existence of statistically significant differences for the hit results with $p < .001$ (for neck: $\chi^2 = 74.70$; for thorax: $\chi^2 = 23.22$). Since the response time results of both evaluations were normally distributed ($p > .05$), we applied the one-factorial ANOVA test that confirmed the existence of statistically significant differences with $p < .001$ and $F(1, 27) = 57.150$ for the neck and with $p \leqslant .05$ and $F(1, 9) = 6.118$ for the thorax results. A statistically significant main effect exists for $\alpha = .05$ with $F_{crit}(1, 27) = 4.21$ for the neck and with $F_{crit}(1, 9) = 5.12$ for the thorax (compare Table A.2). Due to that, the hit results were compared pairwise using the Wilcoxon signed-rank test and the response times using the t-test with Bonferroni correction. The effect size r was calculated using the Equation 5 for the *z-score* and for the *t-statistic* values Equation 4 from Section 2.3.2. All results for each technique comparison are presented in Table 4.2.

Pairwise significant differences for accuracy were confirmed with $p \leqslant .01$ for the neck hit results and with $p \leqslant .05$ for the thorax hit results. Differences between red coloring compared to stippling ($p > .05$) were not statistically significant. Additionally, red coloring compared to cutaway showed statistically no significant difference of achieved hits for the thorax evaluation. Similar to the accuracy results, the response time for red coloring was statistically not significantly different to stippling ($t(27) = -.245, p > .05, r = .04$) for the neck experiment. With $\alpha = .05$ the critical values for t are $t_{crit}(27) = 1.70$ for the neck and $t_{crit}(9) = 1.83$ for the thorax (ignoring the minus sign). All other pairwise comparisons exhibited statistically significant large effects with $p \leqslant .001$ and $r \geqslant .68$ or for red coloring compared to stippling with $p \leqslant .05$ and $r = .53$. The thorax response time results exhibited statistically significant differences for all compar-

Comparison	Neck		Thorax	
	Accuracy	Response Time	Accuracy	Response Time
No Emphasis - Stippling	z = - 4.623 p ≤ .001; r = -.87	t (27) = - 8.085 p ≤ .001; r = .84	z = - 2.803 p ≤ .05; r = -.88	t (9) = - .637 p > .05; r = .20
No Emphasis - Red	z = - 4.623 p ≤ .001; r = -.87	t (27) = - 7.220 p ≤ .001; r = .81	z = - 2.803 p ≤ .05; r = -.88	t (9) = - 1.834 p ≤ .05; r = .52
No Emphasis - Cutaway	z = - 4.623 p ≤ .001; r = - .87	t (27) = - 10.612 p ≤ .001; r = .89	z = - 2.803 p ≤ .05; r = -.88	t (9) = - 2.467 p ≤ .05; r = .63
Red - Cutaway	z = - 4.623 p ≤ .001; r = - .87	t (27) = 3.312 p ≤ .05; r = .53	z = - 1.352 p > .05; r = .42	t (9) = 2.080 p ≤.05; r = .56
Red - Stippling	z = - 1.594 p > .05; r = -.30	t (27) = - .245 p > .05; r = .04	z = - .969 p > .05; r = -.30	t (9) = - 2.485 p ≤ .05; r = .63
Cutaway - Stippling	z = - 4.509 p ≤ .001; r = -.85	t (27) = - 4.851 p ≤ .001; r = .68	z = - 2.295 p ≤ .05; r = -.72	t (9) = - 3.067 p ≤ .01; r = .71

Table 4.2: The p values and the corresponding *z-scores* for the Wilcoxon signed-rank test as well as the *t-statistics* for the t-test are listed in this table for each pairwise technique comparison. If $z \notin [-1.65, 1.65]$, the Wilcoxon signed-rank test confirms a statistically significant difference with $p \leqslant .05$. Moreover, the Pearson's correlation coefficient r with $-1 \geqslant r \leqslant 1$ illustrating the *effect size* is included (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$ effect). Green colored results represent statistically significant differences. Pairwise technique comparisons with no significant difference are colored red.

isons ($p \leqslant .05$ and $p \leqslant .01$ with $r \geqslant .50$) except for no emphasis compared to stippling ($t(9) = -.637, p > .05, r = .20$).

Since the p-values of pairs with no emphasis confirmed a significant difference and the r-values showed a high effect, the postulated hypothesis $H_{Emphasis}$ introduced in Section 4.4 is highly likely for the neck experiment. All participants detected more enlarged lymph nodes and their response times were lower when an emphasis technique was applied. Cutaway achieved significantly more accurate and faster results with a medium and a large effect than no emphasis, red coloring and stippling, as shown in Table 4.2. Thus, $H_{Cutaway}$ can be confirmed for the neck experiment, too. Contrary, $H_{Emphasis}$ and $H_{Cutaway}$ is highly unlikely for the second evaluation. Concerning the first hypothesis, the response times for stippling were statistically not significantly shorter compared to no emphasis when detecting an enlarged lymph node. Furthermore, we had to reject the second hypothesis due to the comparison of the hit results for cutaway and for red coloring ($z = -1.352, p > .05, r = .42$).

4.7 QUALITATIVE RESULTS

The quantitatively measured results corresponded to the technique ranking performed by the participants. For the neck evaluation the ranking from 1 to 4 was: 1: cutaway, 2: stippling, 3: red coloring and 4: no emphasis. The thorax resulting rank for 1 and 4 was similar and only rank 2 with red coloring and 3 with stippling was different. The ranking almost corresponded to the computed results, except for the order of stippling and red coloring. The difference between stippling and no emphasis for the thorax study was very small. Achieved results for the neck

user study did not always correspond with the individual opinions. One subject preferred stippling and achieved the best results with this technique. Although 27 of 28 subjects of the neck experiment had the best performance with cutaway, eight of them favored another technique. That is a variance of 28% of their quantitatively determined most qualified technique. Similar to that, seven subjects of the thorax experiment achieved best results with cutaway or red coloring, three favored another technique (variance of 42%).

4.8 RESULT DISCUSSION

The descriptive results and the bar charts in Figure 4.5 indicate that the participants achieved better results with each technique concerning the enlarged lymph node detection than without emphasis. This first result was confirmed by the not normally distributed data. Contrary to the normally distributed hit results of the target stimuli without an emphasis, the subjects detected more lymph nodes in the emphasized visualizations. Especially for the smaller enlarged lymph nodes, participants achieved more hits.

The number of hits was higher and thus, the participants' results were more accurate when the enlarged lymph node was emphasized with one of the investigated illustration techniques. Especially for the neck experiment, the cutaway views enabled the detection of more enlarged lymph nodes in a shorter response time. Thus, this illustration technique seems to be most suitable regarding accuracy and detection time. The number of hits and the response time results of cutaway views for neck ($\bar{x}$ = 83.19 %, $\bar{x}$ = 705.11 ms) vary from the thorax results ($\bar{x}$ = 96.66 %, $\bar{x}$ = 732.25 ms). The accuracy results of cutaway compared to red coloring as well as stippling compared to red coloring were statistically not significantly different for the thorax. An extended experiment and revision of the thorax stimuli seems to be necessary to refine the individual technique effectiveness. With respect to the false alarm result of 16%, the participants' tendency for a hit reaction (detect enlarged lymph node) was higher compared to the neck experiment. Contrary to the neck, the response time of red coloring compared to stippling was significantly shorter. Since the appearance of a few structures was adapted to enable the visibility of the shown target structures (see Section 4.4.2), red coloring accelerated the node detection for the thorax region, even though the hit results are not significantly different. No statistically significant difference in response times was found for stippling compared to no emphasis.

LIMITATIONS. First of all, this experimental design tried to adapt a visual search experiment based on preattentive perception to complex anatomical structures. Therefore, an appropriate stimuli display duration was required. A trade-off between as long as necessary to enable an almost preattentive perception and as short as possible to prevent a complete serial search was chosen. The display duration, however, was not validated within an individual controlled experimental evaluation. Even though the stimuli design was restricted to fulfill psychophysical criteria and at the same time the individuality of the data needed to be preserved, there is still a difference between the used static stimuli images and the 3D visu-

alizations generated for the therapy planning process. In clinical practice, medical doctors do not rely on static rendered images. Instead, they interactively explore patient-specific 3D models by rotating them, zooming on relevant details and looking to related 2D views with CT slice data for detailed inspection. The study can not reliably predict which emphasis technique is optimal for such settings. Furthermore, there were only two thorax datasets available for the stimuli generation. Compared to the 16 different neck datasets, the two thorax datasets did not provide enough anatomy variations. Thus, it remains an open challenging task to systematically explore emphasis techniques in such settings.

Due to the fact that the thorax evaluation was performed with only 11 participants, a statistically significant difference may occur when recruiting more participants and presenting more stimuli generated from further datasets. Additionally, the participants required more detection time for stippling than red coloring. This may be caused by the worse visibility of the stippling technique. Stippling is a technique to sparsely illustrate structures. Since the enlarged lymph nodes were occluded by a lot of structures in the thoracic region, stippling did not attract the users' attention. Moreover, a detailed analysis referring to adequate transparency parameters and saturation adaption of occluders is recommended.

However, with respect to the application area, the hit results are more important than the response time. Furthermore, compared to all other tested techniques no apparent differences between stippling and red coloring were detected for the neck stimuli, neither for the hit results nor for the response time results. Besides the analysis of the postulated hypothesis, the experiment results enable a first insight into the techniques' effectiveness depending on the lymph node stages introduced in Section 4.1. Emphasized lymph nodes larger $> 1\text{cm}$ are detected more often, especially using cutaway views. Stippling and cutaway views are more supportive for the detection of lymph nodes $> 3\text{cm}$ and $> 6\text{cm}$ than red coloring or no emphasis. However, the subjects detected more lymph nodes $> 3\text{cm}$ that were illustrated with red coloring than without an emphasis. Since lymph nodes $> 6\text{cm}$ ($\geqslant 100$ pixel) are very obvious, the hit rate differences are very small. However, emphasis techniques are more effective.

4.9 SUMMARY

We presented a controlled perceptual evaluation design using 3D patient-specific visualizations and compared three different illustration techniques (red coloring, stippling and cutaway) and their emphasis effectiveness, respectively. These visualizations relate to the neck and thorax region and as a rather general task, the detection of enlarged lymph nodes as an example of a therapeutic task and rather small roughly spherical anatomic structures was investigated. To achieve a reliable result, the well-known visual search theory was used and regarding the participants' individual motivation the experimental design was based on the signal detection theory by Green and Swets [62] described in Section 4.2. Thus, both psychophysical theories were used an applied to the therapeutic task of enlarged lymph node detection. In detail, a conjunctive search of a specific structure and size was performed. The participants' task was to search for the conjunction of two features (a specific structure $\times$ size).

A major challenge was the trade-off between psychological guidelines like stimuli similarities and the patient-specific anatomy. For this purpose, we conducted two experimental evaluations using neck and thorax visualizations from clinical routine, a common therapeutic question and three simple representative visualization techniques. In detail, the techniques' capability was validated by analyzing the accuracy and the required response time to detect an enlarged lymph node rendered with red coloring, stippling and cutaway. The capability was defined via the search task performance. This evaluation gives an insight into the experimental design, conditions and implementation.

However, this may serve as an orientation for evaluating the capability of techniques concerning specific purposes in 3D therapy planning scenarios. The presented experimental setup is applicable to other medical domains, such as radiation treatment planning. Both evaluations proved that emphasis techniques guide the users' attention and thus improve and enhance the enlarged lymph node detection. Even an illustration technique like stippling that is originally developed to sparsely illustrate context structures as shown by Tietjen et al. [155], supports the attention guidance. Thus, it is neither possible nor right to generalize that all pen-and-ink techniques are preferable context visualization techniques. This evaluation documents the influence of the displayed surrounding structures or objects to the techniques' emphasis effectiveness. Every emphasis technique is better than no emphasis regarding the detection of the enlarged lymph node.

5

A COMPARATIVE EVALUATION OF FEATURE LINE TECHNIQUES

This chapter is based on the following publications:

"Comparative Evaluation of Feature Line Techniques for Shape Depiction".
Kai Lawonn, Alexandra Baer, Patrick Saalfeld and Bernhard Preim. In *Proceedings of Vision Modeling and Visualization*, pp. 31-38, 2014

"Statistical Analysis of a Qualitative Evaluation on Feature Lines".
Alexandra Baer, Kai Lawonn, Patrick Saalfeld and Bernhard Preim. In *Proceedings of Bildverarbeitung für die Medizin*, pp. 71-76, , 2015

Originally, medical and other conceptional or abstract illustrations were hand drawn and consisted of simple but well defined and well placed lines and dots. As mentioned in Section 3.3.1, such pen-and-ink illustrations have been proven to be effective according to structure cognition including the structure's shape and surface orientation [35, 73, 86]. In a recent work of Lawonn et al. [104], feature lines were quantitatively compared according to personal preferences. The author of the thesis was involved in this work and primarily designed the evaluation, provided the evaluation tool and performed the statistical analysis, while Kai Lawonn implemented the six techniques and generated the feature line illjustrations. The qualitative data analysis was a joint work of Kai Lawonn and the author of this thesis. As different existing evaluations compare some feature line methods quantitatively and qualitatively, no study compared all of the most commonly used feature line methods in one evaluation.

Therefore, Lawonn et al. [104] investigated six feature line methods and compared them in an extensive evaluation: a pilot study comprising 20 participants and a final evaluation with 129 participants. The techniques were qualitatively evaluated in the context of *realism* and *aesthetics* and the participants had to choose their favorite feature line illustration. Related line drawing evaluations investigate either just personal preferences or the potential of computer-generated illustrations imitating hand-drawn illustrations [77]. Moreover, previous works analyzed the artists' drawing techniques and aesthetic aspects to define surface regions where to draw the lines [34]. Hand-drawn illustrations aim at simplifying real objects or complex scenes to ease and support the understanding by concentrating on the most important features. That means, a realistic object or scene is visual-

ized with a few well placed lines but still illustrates the realistic situation. Thus, realism is a major topic for computer-generated and hand-drawn feature line visualizations. Additionally, illustrations should visualize the objects or scenes in an aesthetic way. Thus, aesthetics is another main aspect when comparing feature line drawing techniques.

This chapter presents the study design, implementation and analysis. The results of the evaluation were used to provide guidelines which feature line method is best suited according to personal preferences compared to the other techniques. Moreover, the first qualitative analysis was extended and quantitatively examined, which will be explained in Section 5.6

5.1 FEATURE LINE METHODS

This description and summary of feature line methods is contributed by Kai Lawonn [104]. Feature line methods are a special kind of low-level visual abstractions. They are used to give a simplified representation of an object. For this simplification, the object is illustrated by using lines, which are placed at the most salient regions such that the object's shape can be perceived without additional shading. Feature lines are a family of non-photorealistic visualization techniques that can be applied on arbitrary surfaces and thus on patient-specific data, as introduced in Section 3.3.1. This low-level abstraction of surfaces is important and has a high potential for depicting pathologies, anatomical and risk structures required for surgery planning [155] and for intraoperative visualizations [137]. Moreover, abstract illustrations of patient-specific data and therapy options support an individual patient documentation. The most commonly used feature line techniques comprise six methods.

Interrante et al. [71] introduced *ridges and valley lines* (RV) to illustrate salient regions. This method is curvature-based and therefore view-independent. Thus, DeCarlo et al. [42] presented a view-dependent approach: *suggestive contours* (SC). SC are an extension of the conventional contour definition, but fail in depicting convex structures. Therefore, Judd et al. [81] combined the advantages of RV and SC and presented *apparent ridges* (AR). This method uses a view-dependent curvature term and applies the RV definition to illustrate the shape. Xie et al. [179] introduced *photic extremum lines* (PLs). This technique determines the maximum of the variance of illumination. The user can add additional spotlights to influence the result. Kolomenkin et al. [88] presented *demarcating curves* (DC), a view-independent approach. This method is best suited to enhance furrows. Zhang et al. [183] introduced *Laplacian lines* (LL). LL extend the Laplacian-of-Gaussian for images to 3D surfaces to illustrate the shape. Figure 5.1 presents all techniques applied to different models and surfaces. The feature line drawings were generated in cooperation with two artists who are familiar with computer-generated line drawings.

5.2 EXPERIMENTAL DESIGN

In detail, we performed a comparative evaluation that was presented by Lawonn et al. [104]. Personal beliefs, opinions and preferences were analyzed to determine the most appropriate line drawing techniques according to realism and aesthetics. One possibility to realize that is to ask the participants to assess each technique. This leads to an independent technique evaluation without comparison to the other techniques. Another possibility is to ask the participants to compare the techniques. This method is more appropriate in this case, since there are various techniques available when generating an illustration and the artist has to choose which one is more suitable. Thus, the participants of this evaluation were asked to compare presented line drawing techniques to assess which one is better in terms of more realistic or more aesthetic. This is called a *comparative evaluation* or more specifically an *ordered ranking*. Different stimuli are compared along a given dimension by asking the participants to place the stimuli along that dimension in some order [40]. An ordered ranking evaluation investigates if one technique is better than another technique, instead of determining an exact value for one technique. To realize this comparison task, we asked participants to rank from 1 to 6, whereas rank 1 means the most realistic or aesthetic depiction. Additionally, participants had to define one stimulus image as the most realistic and one as the most aesthetic image. This additional task was used to double-check their previously defined ranking. The major research questions of the comparative evaluation were:

- What does a technique ranking from 1 to 6 look like for a realistic and an aesthetic depiction comprising all six line drawing techniques?
- Which feature line technique is considered to be the most aesthetic out of the six line drawing techniques?
- Which feature line technique is considered to be the most realistic out of the six line drawing techniques?

The illustration technique was the *independent variable* for this evaluation. That means, this was a *one-factorial* evaluation with six factor *levels*. Since a technique ranking had to be defined, the measured *dependent variables* for each technique were the assigned ranks. All participants assessed the six line drawing techniques and thus this evaluation followed a *within-participant* design (see Section 2.2.4).

5.2.1 *Participants*

Overall, 149 participants were recruited: 20 for the pilot study and 129 for the final evaluation. The 129 participants were gathered at the *long night of the sciences in 2014*. Therefore, openness and interest in science was above average and they had a broad spectrum of educational backgrounds. In detail, 15 men and five women participated in the pilot study and 68 men and 61 women attended the final evaluation. The age of the men in the pilot study ranged from 24 to 33 years ($\bar{x} = 26.43$ years) and for the final evaluation from 10 to 67 years ($\bar{x} = 30.53$ years). On the other hand, the women participating in the pilot study were aged between

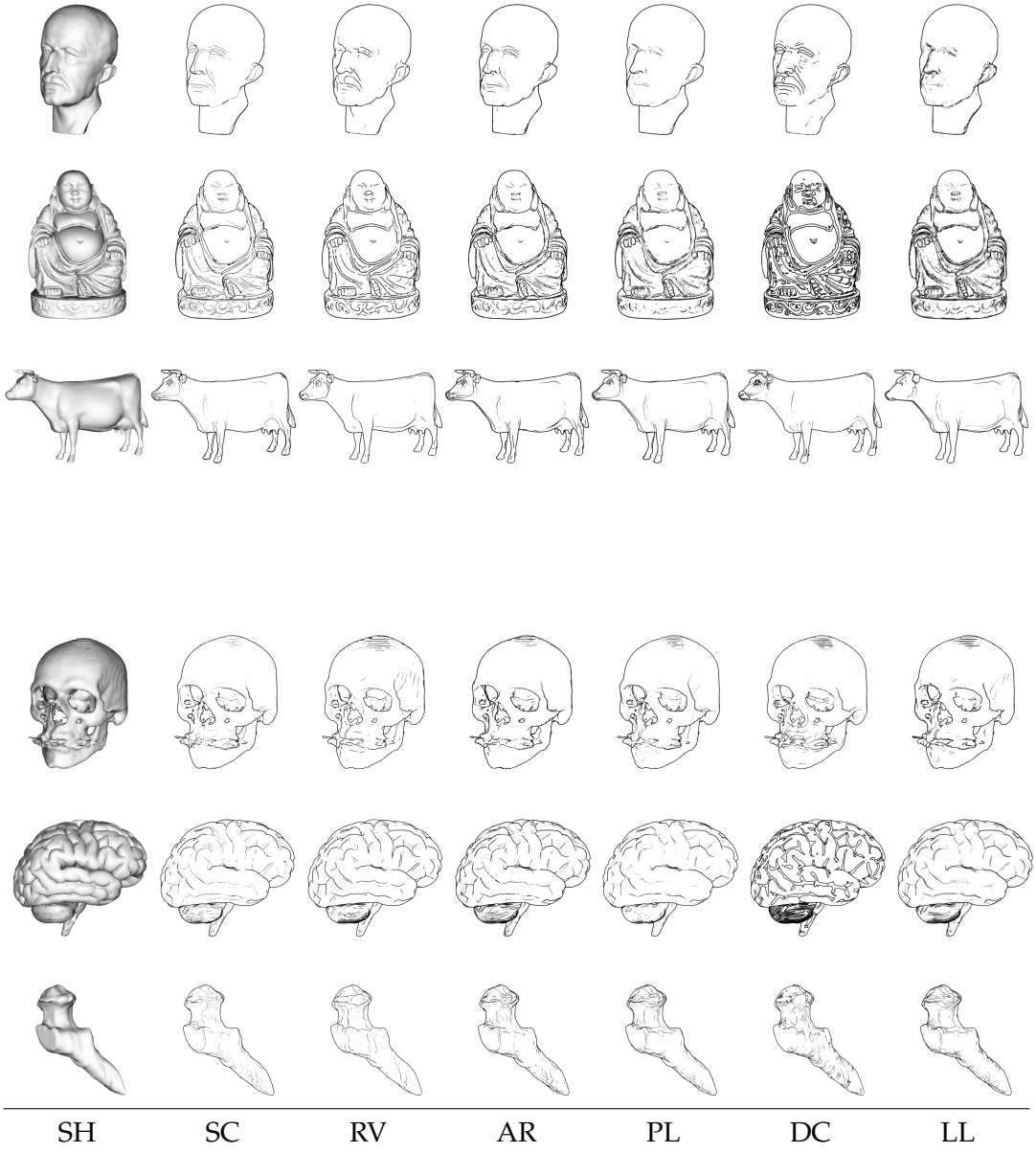


Figure 5.1: All stimuli images and the shaded visualization (SH, left) of each model. The Max Planck, buddha and cow model above and the medical models (skull, brain and femur) below visualized with each of the six feature line methods (SC: suggestive contours, RV: ridges and valleys, AR: apparent ridges, PL: photic extremum lines, DC: demarcating curves, and LL: Laplacian lines). (Illustrations generated by Lawonn [102])

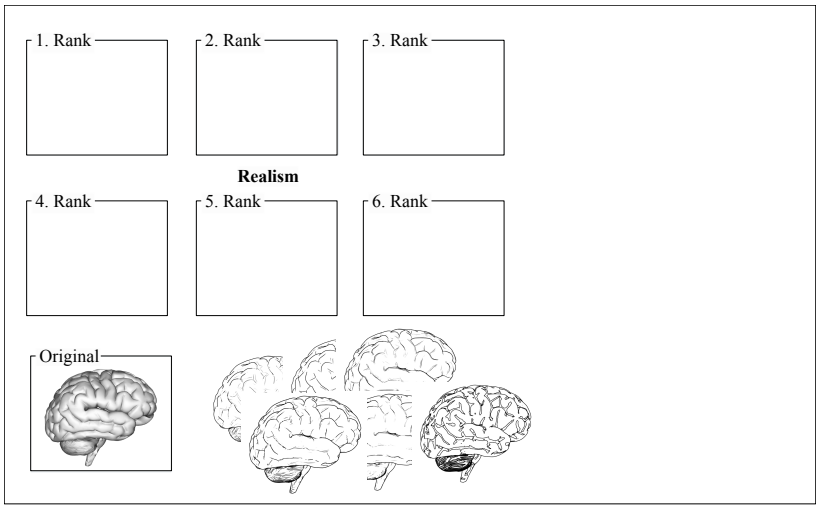
29 and 34 years ($\bar{x} = 32.53$ years) and the women who attended the final evaluation between 12 and 68 years ($\bar{x} = 30.92$ years).

5.2.2 *Stimuli*

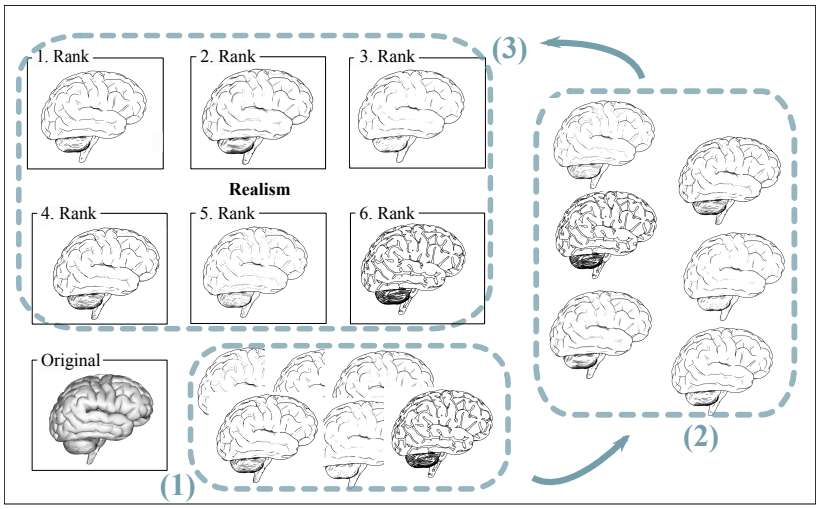
A stimulus was defined as an image of 264×214 pixels illustrating a feature line representation of one model. We decided to use screenshots, since the participants would otherwise lose their concentration after observing a few models and the evaluation focused not on the illustrating process but on the personal preferences of a finished illustration. Three medical models were derived from either patient-individual datasets, i.e., the femur and the skull model, or were artificially generated, i. e., the brain model (see Figure 5.1 below). Additionally, to provide a broad spectrum of surfaces, widespread computer graphics models such as a cow, the Max Planck and a buddha model were used as well (see Figure 5.1 above). For each model, six stimuli were generated illustrating this model with the feature line techniques RV, SC, AR, PL, DC, and LL. Each stimulus was generated in cooperation with two artists who are familiar with computer-generated line drawings. The corresponding feature line technique parameters were generated and adjusted on their own. This resulted in feature line drawings that illustrate the important features of each model and were individually generated similar to the hand-drawn illustration process. One stimulus set comprised all six stimuli of one model. Six models were used to generate six different stimuli sets. All models and stimuli sets are shown in Figure 5.1. Each row shows one stimulus set with the shaded visualization in the first place followed by the six generated stimuli.

5.3 APPARATUS AND PROCEDURE

An evaluation tool designed and implemented using C# enabled the participants to sort and to rank the presented stimuli sets and to record the personal data and ranking results. This evaluation tool was implemented by the author of this thesis and is shown in Figure 5.2 with the brain model and the corresponding stimuli set. For the realism and the aesthetics task, the participants saw the surface visualization of each model in a grey shading positioned in the lower left corner (see Figure 5.2a). Next to this surface model, the stimuli set was arranged randomly one image above another (see Figure 5.2b, (1)). This arrangement was chosen to force the participants to rearrange the images for an overview, as shown on the right hand side in Figure 5.2b (2). Thus, for each participant an individual stimuli set presentation was achieved, unconsciously generated by the participants. Six potential ranks were illustrated as labeled rectangular areas and the participants were able to assign each stimulus to one rank with drag and drop (see Figure 5.2b, (3)). For the selection of the favorite line drawing representation, the gray shaded model representation was shown again and the participants had to select one stimuli image by positioning this image to a defined rectangular area. The evaluation tool was presented on a display with a screen resolution of 1600×1200 pixels. The log files were generated automatically during the evaluation. For every participant, an ID was generated to ensure anonymity. The individual personal details, i.e., age and gender and the evaluation results (realism ranking, aesthetics ranking



(a)



(b)

Figure 5.2: (a) The evaluation tool with the brain model for the realism task. (b) For each task the shaded stimuli were visualized in the lower left corner and next to it (1) the six stimuli images (the stimuli set) were randomly arranged. (2) Initially, the participants rearranged the stimuli set on the side to get an overview of all stimuli images. (3) The six potential ranks were illustrated as labeled rectangular areas and the participants were able to assign each stimulus to one rank with drag and drop.

and favorite selection) were recorded.

Initially, the participants were instructed in written form, to ensure equal evaluation conditions for each participant. Moreover, the individual tasks were displayed at any time, to provide all required information to the participants. The evaluation was structured into:

- the personal data acquisition
- the instruction and training
- followed by the evaluation consisting of a realism and an aesthetics assessment and a technique preference task.

A training task was used to familiarize the participants with the evaluation tool and to learn the drag and drop interaction. The participants saw a final image which consists of a sentence with six words. The participants were asked to sort the words such that the final sentence occurs. We used a sentence to indicate that an ordering is required during this evaluation. After this training, the study began. For each model, the participants were asked to rank each image of the corresponding stimuli set from 1 to 6. An evaluation trial starts with the realism task followed by the aesthetics task and ends with the selection of the overall favorite stimuli image. In detail, the same stimuli set was shown to the participants three times. First, they were asked to order and to rank the stimuli images according to realism, then according to aesthetics, and finally, to select their personal favorite stimuli image and thus feature line representation. After that, the second evaluation trial started with the next model until all models and corresponding stimuli sets were ranked and assessed.

Initially, a pilot study was conducted to make an inquiry about the participants' behavior. In the pilot study (about 15-20 minutes), every participant saw all six models and stimuli sets visualized in Figure 5.1. Participants most frequently complained about the duration time and the same questions during this pilot study. They stated that it lasts too long and that this leads to reduced concentration. From this insight, the final evaluation was slightly altered. Every participant had to perform the evaluation with two models. To assess all six stimuli sets, three evaluation sets were used with two different models each. The study was conducted on three computers, all equipped with the same monitor and resolution as mentioned above in Section 5.3. Thus, the study lasts $7-10$ minutes, the participants acted more concentrated and all six stimuli sets were assessed. Even though there are three participant groups with two different models shown to each group, all participants asses the same six line drawing techniques (independent variable) and thus this evaluation followed a *within-participant* design (see Section 2.2.4).

5.4 QUALITATIVE ANALYSIS AND RESULTS

The assigned ranking positions have to be analyzed for each line drawing technique, to enable a technique assessment and comparison. Therefore, the results of each task (realism, aesthetics and favorite) were analyzed individually. Additionally, a more detailed analysis for each task and for each model was performed. Stimuli sets of six models were assessed by 129 participants divided into three groups without the pilot study results. Each group with 43 participants ranked stimuli sets of two models for realism and aesthetics. Thus, for each group 86 realism and aesthetics rankings and favorite selections exist. Overall, 258 rankings for each task were recorded.

5.4.1 *Frequency Distribution*

Since this initial analysis focused on a qualitative comparison, no quantitative, mathematical or statistical methods were applied during the analysis process.

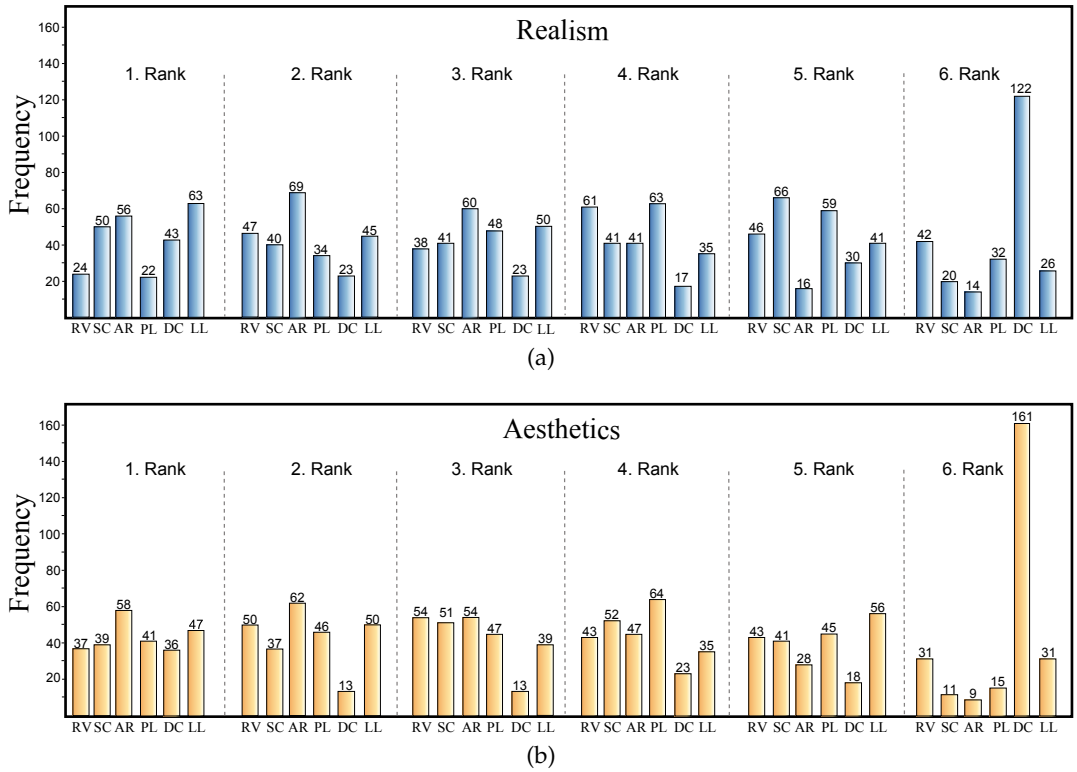


Figure 5.3: The frequency distribution of the technique rating for (a) the realism and for (b) the aesthetics assessment visualized for each rank. (SC: suggestive contours, RV: ridges and valleys, AR: apparent ridges, PL: photic extremum lines, DC: demarcating curves, and LL: Laplacian lines)

However, various analysis methods are available as described in Section 2.3.1. The first and most common way to summarize the recorded results is a frequency distribution (histogram) [52]. Since the first research question (see Section 5.2) required a ranking from 1 to 6 comprising all techniques, a frequency distribution is not sufficient enough for this qualitative comparison. A result based on the number of assigned ranks does not always result in a unique ranking with each technique in one rank.

The frequency distribution for realism indicates that LL was the most realistic line drawing technique, since this technique was ranked with the first place for 63 times (see Figure 5.3a). Compared to all six techniques, AR exhibited more ratings for the second and the third rank, PL for the fourth, SC for the fifth and DC for the sixth rank. The ranking according to the frequency distribution was LL, AR, AR, PL, SC and DC. AR was in the second and in the third rank and RV was in no rank. As shown in Figure 5.3b, AR had most ratings for the first, the second and the third rank for the aesthetics assessment task. Additionally, RV exhibited most ratings in the third rank with the same frequency as AR. The ranking for the aesthetics assessment task according to the frequency distribution was 1: AR, 2: AR, 3: AR and RV, 4: PL, 5: LL and 6: DC. Thus, a pairwise comparison to analyze how often one technique had a smaller and thus a better rank than another technique was required to achieve a unique ordering.

5.4.2 *Schulze Method*

Methods that determine a resulting ranking are used in elections where the voters order the candidates. They are called *Condorcet* methods, named after the 18th-century French mathematician and philosopher Marie Jean Antoine Nicolas Caritat, the Marquis de Condorcet. Winners are selected by majority in comparison to all pairings against the other candidates. Each candidate will be compared pairwise and it will take into account if one candidate is more often preferred against another candidate. The most widespread method is the *Schulze* method [146]. "Having a list of ranked candidates, the Schulze method determines the winner and the final rank order of the candidates. The *Schulze* method fulfills several criteria, which are important for the election. For example the *Condorcet* criterion, which means whenever a candidate is more often preferred in comparison to the other candidates, this candidate should win. Another example would be the *reversal symmetry*, which means that if an election chooses a candidate as the winner the same candidate should not be the winner if the rankings would be inverted" [104]. This method fulfills the requirements for a qualitative comparison and analysis method. The original ratings are used to determine a resulting ranking without statistical analysis methods or mathematical calculations. Thus, we used this method to determine the line drawing technique ranking for the realism and the aesthetics results.[1]

The *Schulze* method requires the generation of a rank matrix, creating and analyzing a directed graph and the determination of a minimal weight matrix to receive a final technique ranking. The following steps were applied to our ranking results for the realism and for the aesthetics assessment task:

1. **Rank List:** All recorded technique ranking orders and their frequencies were listed. Six feature line methods result in $6! = 720$ potential rankings that may occur.

2. **Rank Matrix** R**:** A rank matrix R was created such that r_{ij} denotes how often method i had a smaller and thus a better rank than method j.

3. **Directed Graph:** Based on the rank matrix a directed graph was created. If $r_{ij} > r_{ji}$ in the rank matrix, a direct edge (i, j) with weight r_{ij} was created in the graph.

4. **Minimal Weight Matrix** M**:** All graph paths from i to j were considered and analyzed to determine a new matrix M called minimal weight matrix. The matrix entry m_{ij} was set to the overall strongest edge of the weakest element in a path, i.e., having two paths p_1, p_2 from i to j and let w_1 be the lowest weight of p_1 and w_2 of p_2, then $m_{ij} = max\{w_1, w_2\}$. If $w_1 \geqslant w_2$ then $m_{ij} = w_1$. Finding the matrix M, the Floyd-Warshall algorithm [55] was applied.

5. **Technique Order:** The matrix M was analyzed according to the highest value entries that define the individual rank for each technique and result in a unique technique ranking of the six investigated line drawing techniques.

1 This *Schulze* analysis was primarily performed by Lawonn et al. [104].

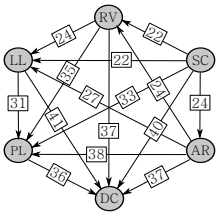
Rank

1	2	3	4	5	6
AR	SC	PL	LL	RV	DC
SC	PL	LL	AR	DC	RV
LL	SC	AR	DC	PL	RV
SC	LL	RV	AR	DC	PL
⋮	⋮	⋮	⋮	⋮	⋮

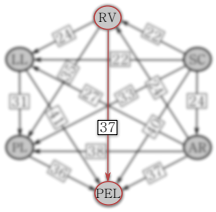
(a) Rankings

	RV	SC	AR	DC	PL	LL
RV	0	21	19	37	35	24
SC	22	0	24	40	33	22
AR	24	19	0	37	38	27
DC	6	3	6	0	7	2
PL	8	10	5	36	0	12
LL	19	21	16	41	31	0

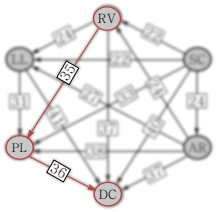
(b) Rank Matrix



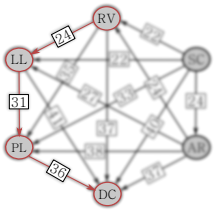
(c) Directed Graph



(d) Path 1



(e) Path 2



(f) Path 3

	RV	SC	AR	DC	PL	LL
RV	0	0	0	37	35	24
SC	24	0	24	40	33	24
AR	24	0	0	37	38	27
PL	0	0	0	0	0	0
DC	0	0	0	36	0	0
LL	0	0	0	41	31	0

(g) Minimal Weight Matrix

Figure 5.4: (a) Initially, all recorded rankings are listed. (b) A rank matrix R is created where r_{ij} denotes how often i has a better rank than j. Green marked entries indicate that $r_{ij} \geqslant r_{ji}$. (c) These values are used to set up a weighted direct graph. (d)-(f) A new matrix M is determined. The entry $m_{1,4}$ is obtained by analyzing all possible paths from RV to DC. For each path, the minimal weight is denoted (37,35,24). (g) The biggest weight is the entry for the matrix $m_{1,4} = 37$. (SC: suggestive contours, RV: ridges and valleys, AR: apparent ridges, PL: photic extremum lines, DC: demarcating curves, and LL: Laplacian lines) (Images reprinted, with permission, from Lawonn et al. [104] © Eurographics Association 2014.)

Figure 5.4 illustrates the *Schulze* method including each of the explained steps for the buddha model and the realism assessment task as presented by Lawonn et al. [104]. Initially, in step 1 all recorded technique rankings were listed followed, by the rank matrix R generation of step 2, as shown in Figure 5.4a and 5.4b. RV for example was 37 times more often preferred compared to DC for a realistic buddha visualization. Therefore, the rank matrix entry was $r_{1,4} = 37$. In contrast, DC was only six times preferred compared to RV ($r_{4,1} = 6$). Since the rank matrix entries were $r_{1,4} = 37$ and $r_{4,1} = 6$, a direct edge with weight 37 was created in the directed graph for RV compared to DC in step 3, see Figure 5.4c. Step 4 comprised the analysis of all paths from i to j. As shown in Figures 5.4d-5.4f, all paths from RV to DC were considered. Since the lowest weights of the three listed paths were 37, 35 and 24, the matrix entry was $m_{1,4} = max\{37, 35, 24\}$ and thus $m_{1,4} = 37$. Finally, in step 4 we analyzed the matrix M according to the highest value entries. The winner for this buddha example was SC, as it was preferred (placed on a smaller rank) compared to all other techniques (the row is completely green, see Figure 5.4g). The second most realistic depiction for the buddha model was achieved with AR according to the participants' opinion. The other rankings from 3 to 6 were: RV, LL, PL, and DC.

5.4.3 *Ranking Results*

Initially, we applied the *Schulze* method to the results of each task for all models (model-independent). This was followed by a model-based analysis, where we an-

alyzed the ranking results for each model and task. This was done to analyze and to verify the chosen models in detail. The result presentation and interpretation is a joint work of the author of this thesis and Lawonn et al. [104].

MODEL-INDEPENDENT. The *Schulze* method determined a technique ranking for the realism assessment task for all models from 1 to 6 with 1: AR, 2: LL, 3: SC, 4: RV, 5: PL and 6: DC. The resulting ranking for the aesthetics assessment was 1: AR, 2: SC, 3: LL, 4: RV, 5: PL, and 6: DC. These two technique rankings answered both research questions that asked for a technique ranking for realism and aesthetics (recall Section 5.2). Similar to the determined first rank of AR for realism and aesthetics, this technique was chosen as the favorite line drawing technique by 65 participants, too. In detail, 65 times AR was chosen to be the favorite technique, followed by LL (58 times), SC (49 times), PL (33 times), RV (27 times), and DC (26 times). Thus, the favorite technique ranking from 1 to 6 was: AR, LL, SC, PL, RV, and DC. Hence, the realism and aesthetics ranking is verified by this technique selection result. The preferred technique for realism and aesthetics was AR, which answers the second and third research question introduced in Section 5.2) for the most preferred technique for each assessment task.

Rank	Buddha	Brain	Cow	Femur	Max	Skull
1	SC	AR	AR	LL	AR	SC
2	AR	SC	PL	DC	LL	LL
3	RV	LL	RV	AR	SC	PL
4	LL	RV	DC	PL	RV	AR
5	PL	PL	SC	SC	PL	DC
6	DC	DC	LL	RV	DC	RV

(a) Realism

Rank	Buddha	Brain	Cow	Femur	Max	Skull
1	AR	LL	PL	LL	AR	PL
2	RV	SC	AR	AR	LL	RV
3	SC	AR	RV	PL	RV	SC
4	LL	RV	DC	DC	PL	AR
5	PL	PL	SC	SC	SC	LL
6	DC	DC	LL	RV	DC	DC

(b) Aesthetics

Table 5.1: The ranking results determined by the *Schulze* method (a) for realism and (b) for aesthetics. (SC: suggestive contours, RV: ridges and valleys, AR: apparent ridges, PL: photic extremum lines, DC: demarcating curves, and LL: Laplacian lines) (Content of tables reprinted, with permission, from Lawonn et al. [104] © Eurographics Association 2014)

MODEL-BASED. Tables 5.1a and 5.1b contain the determined rankings for each model and for both tasks. Both tables show that AR, SC and LL were more often placed in the three first ranks (AR: 10, SC: 7 and LL: 7) than in the bottom ranks. The technique rankings were similar between the tasks for the buddha, brain, cow, femur and Max model, except for the techniques RV and LL applied to the skull model. RV reached the last rank for realism, while it was assigned to the second rank for aesthetics. LL was assigned to the second rank for realism and to the fifth rank for aesthetics. Even though the other techniques and their assigned ranks varied between the two tasks, no other technique exhibited a rank difference of more than two. The favorite results per model are listed in Table 5.2. Similar to realism and aesthetics, the results confirmed the preference for AR, SC and LL. AR had the most votes for the cow and the Max model, followed by SC that exhibited the highest selection frequency for the buddha and skull model. Participants preferred LL for the brain and the femur model. Interestingly, DC was chosen as the favorite method for the femur model, too. For this model, DC was in the

	Buddha	Brain	Cow	Femur	Max	Skull
RV	8	7	3	2	6	1
SC	13	7	4	6	5	14
AR	11	12	14	6	16	6
PL	2	1	10	9	3	8
DC	1	3	8	10	1	3
LL	8	13	4	10	12	11

Table 5.2: The favorite technique results determined for each technique and for each model. The green colored values represent the favored technique for the corresponding model.

second rank for realism (compare Table 5.1a). However, mostly it did not reach a rank better than four. In summary, the *Schulze* method placed AR, SC, LL in the first three ranks for realism as well as for aesthetics, which was confirmed by the analysis of the favored techniques. AR, SC and LL had the highest frequencies, too.

Besides mentioning that they did not like the skull model, only eight participants took the opportunity to pass a more detailed comment at the end of the experiment. Mostly, the participants mentioned that "some methods were hard to distinguish", without specifying the difficult stimuli images. Furthermore, some mentioned that the realism and the aesthetics task were hard to distinguish. One participant wrote that some parts of the brain seemed a bit more realistic for some methods and other methods delivered a more realistic impression on other regions of the brain.

5.5 RESULT DISCUSSION

According to the previously presented rankings, AR (apparent ridges), SC (suggestive contours) and LL (Laplacian lines) were chosen to be the techniques that realized the most realistic and aesthetic depiction and that were preferred by the participants when selecting a favorite illustration. However, when analyzing the technique results for each model, the results for the skull model visualized with LL and RV differ more than two ranks, as presented in the previous section. Since LL and RV applied to all other models were similar to the results of the other techniques, the explanation may be found in the complexity of the skull model. Since it has too many features, it was hard to distinguish the different feature line methods. Furthermore, several participants stated that they do not like this model and based on that they had difficulties to rank the stimuli images according to aesthetics. Moreover, participants complained about difficulties with the brain model. This was identified for the cow model, too. "For instance, some feature line methods strongly depict the eyes of the cow, although they are hardly perceivable from the shaded image. Thus, one can raise the question whether it is realistic to emphasize the eyes or not. Artists would illustrate the eyes as they are an important characteristic for such a model" [104]. This decision-making process is based on a priori knowledge of the surface, which cannot be performed by modern feature line methods as they only use surface measurements. Hence, to choose the right feature line method, it is first a matter of taste and second a question of the underlying surface and the required performance.

SURFACE PROPERTIES. According to the results of the evaluation and thus the participants' taste, Lawonn et al. [104] recommended to use a method of low order derivatives on surfaces which exhibit noise. Thus, SC was recommended, as this method is of second order, while the other methods uses third-order derivatives. Beyond that, SC has two equivalent representations. This feature line method may be determined by radial curvature directions or with the headlight. If the feature lines were determined by the light representation, no pre-processing is necessary. This technique, however, cannot depict convex features or illustrate the object's contour. The feature line techniques AR and LL on the other hand are able to convey the contour, to illustrate sharp edges and depict a cube appropriately.

PERFORMANCE. AR is the slowest of the presented techniques. "As the computational effort is very high, AR reaches only 8 FPS on a mid-class PC of a model with 64k triangles. In comparison, SC reaches 45 FPS and LL 15 FPS. One disadvantage of the technique LL is the required substantial computational effort for pre-processing. First, the Laplacian of the surface needs to be calculated. As it is recommended in Zhang et al. [183] to use the Belkin weights [14], first a parameter is used to determine the Laplacian. Unfortunately, the user has to wait until the computation is finished" [104]. If the result is not satisfying, a trial-and-error loop with different parameters is required until a satisfied result is reached.

In summary, the personal preferences tend to AR, SC, and LL and according to pre-processing and underlying surface properties, AR and SC are recommended. SC is very fast and delivers satisfying results on most meshes. Unfortunately, this holds not on convex surfaces where we recommend to use AR. "Especially for organic surfaces which were derived from medical image data, we recommend to use SC, as these surfaces may exhibit surface noise. For representing industrial models, which exhibit sharp edges, AR is strongly recommended" [104]. The evaluation, however, represented the participants' preferences and provided a qualitative final rank ordering, but no statistical significance was tested. A winner and the three most preferred techniques were determined, but it was not clear if this result was statistically relevant.

5.6 QUANTITATIVE ANALYSIS AND RESULTS

Since the presented evaluation design also fulfills the requirements of a quantitative study with a defined *independent* and *dependent variables*, Baer et al. [10] postulated two-tailed hypotheses and statistically analyzed the ranking results. The hypotheses were:

- H_{Realism}: There is a difference in personal preferences between the feature line techniques for realism.
- $H_{\text{Aesthetics}}$: There is a difference in personal preferences between the feature line techniques for aesthetics.

This analysis consisted of two steps and will be presented within this section. The first step comprised of the quantitative technique analysis. This step was followed by a second, more detailed model-based technique analysis step, similar to the

two analyses presented in Section 5.4.3. The software package IBM SPSS STATISTICS (Statistical Package for the Social Sciences) was used for the descriptive and statistical analysis.

5.6.1 *Descriptive Analysis*

Initially, the first and simplest quantitative data processing and representation is the calculation of mean and medium values, compare Section 2.3.1. In contrast to the *Schulze* method that integrated all ranking results to perform a pairwise comparison, this analysis requires one ranking result for each participant and task. Since each participant assessed two stimuli sets for each task and thus two models, a mean ranking for each task was calculated per participant. Then, the mean and median realism and aesthetic ranks were calculated of these 129 rankings. As already mentioned, the first analysis step comprised a model-independent analysis and in the second step the analysis was performed for each model.

	RV	SC	AR	DC	PL	LL
$\bar{x}$	3.71	3.35	2.78	4.28	3.76	3.09
σ	1.05	1.29	0.98	1.53	1.24	1.01
md	3.50	3.50	2.50	4.50	4.00	3.00

(a) Realism

	RV	SC	AR	DC	PL	LL
$\bar{x}$	3.39	3.40	2.81	4.74	3.27	3.37
σ	1.30	1.18	0.99	1.59	1.17	1.14
md	3.50	3.50	2.50	5.50	3.00	3.50

(b) Aesthetics

Table 5.3: The mean ($\bar{x}$), the standard deviation (σ) and the median (md) results for (a) the realism (above) and (b) the aesthetics (below) assessment.

MODEL-INDEPENDENT. A ranking result based on the mean ranks for realism was: AR ($\bar{x} = 2.78$), LL($\bar{x} = 3.09$), SC ($\bar{x} = 3.35$), RV ($\bar{x} = 3.71$), PL ($\bar{x} = 3.76$), DC ($\bar{x} = 4.28$) and for aesthetics: AR ($\bar{x} = 2.81$), PL ($\bar{x} = 3.27$), LL ($\bar{x} = 3.37$), RV ($\bar{x} = 3.39$), SC ($\bar{x} = 3.40$), DC ($\bar{x} = 4.74$) (see Table 5.3). In detail, for realism the participants tended to place AR, LL and SC on average in the third rank and RV, PL and DC in the fourth rank. In contrast to that, for aesthetics AR, PL, LL, RV and SC were on average in the third and DC in the fifth rank. Table 5.3 presents the standard deviation to illustrate the variation from the calculated mean values. For both tasks, AR had the lowest ($\sigma_{realism} = 0.98$, $\sigma_{aesthetics} = 0.99$) and DC had the highest ($\sigma_{realism} = 1.53$, $\sigma_{aesthetics} = 1.59$) standard deviation.

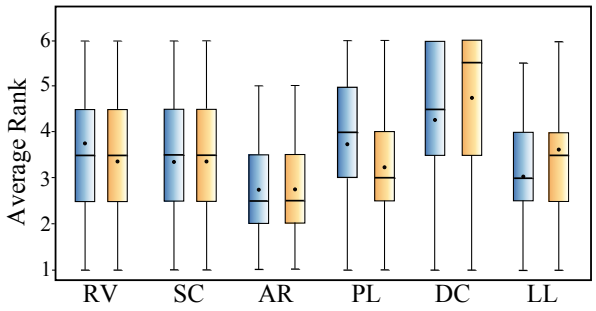
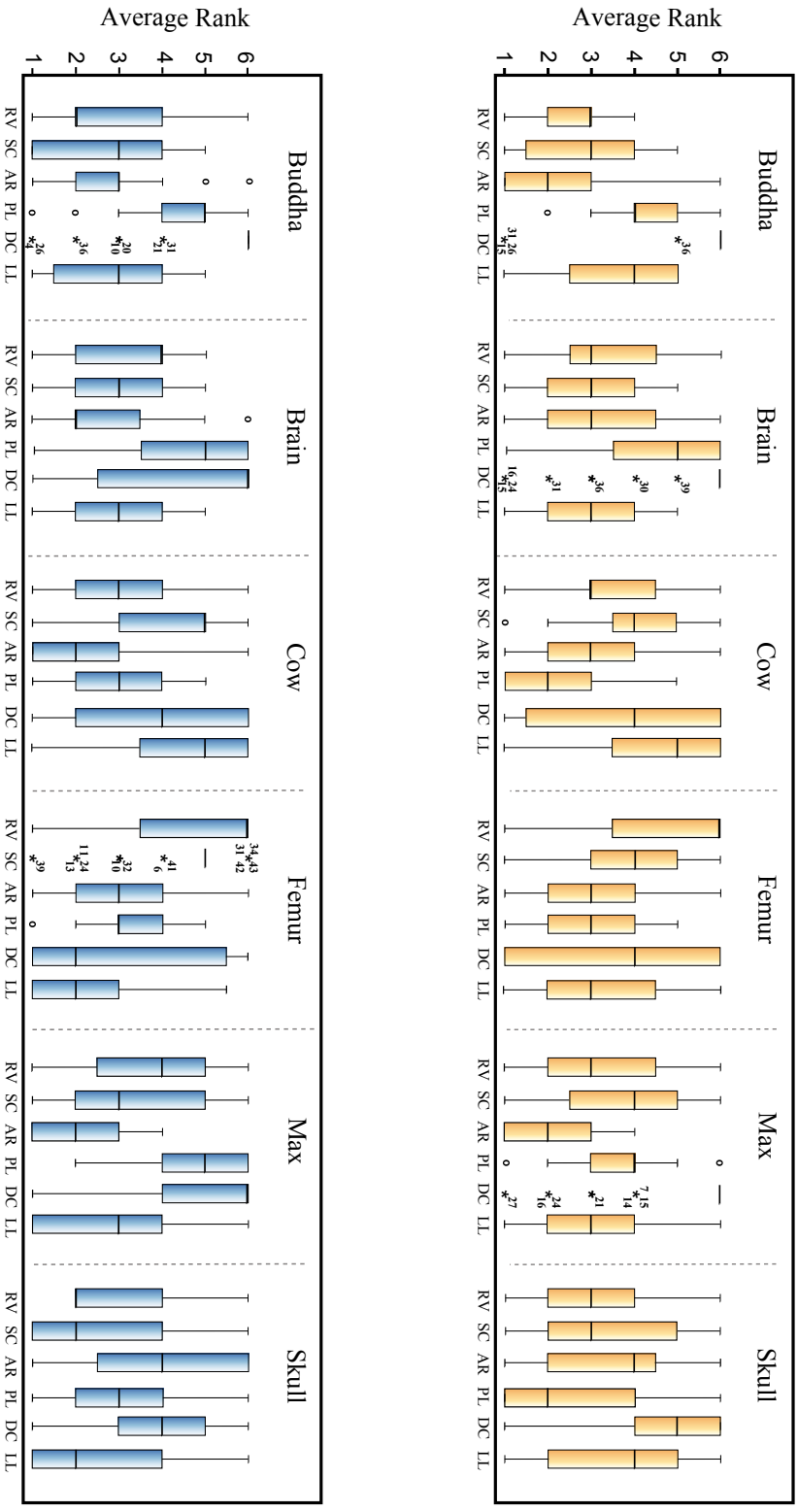


Figure 5.5: Both task results are visualized as boxplots. Blue boxplots are the realism and yellow are the aesthetics results. The mean values are integrated and visualized as black dots.



(a) Realism (b) Aesthetics

Figure 5.6: The model-based results for (a) realism and (b) aesthetics visualized as boxplots. Detected outliers are displayed as an asterisk with the according sample number. Circles are potential outliers. (Tables reprinted, with permission, from Baer et al. [10] © 2015, Springer-Verlag Berlin Heidelberg.)

	RV	SC	AR	DC	PL	LL
Buddha	2.83	2.72	2.62	5.44	4.37	3.02
Brain	3.30	2.76	2.69	4.67	4.44	3.11
Cow	3.18	4.00	2.46	3.79	2.83	4.72
Femur	4.97	4.67	3.09	3.02	3.09	2.13
Max	3.60	3.32	1.90	4.81	4.48	2.79
Skull	4.39	2.60	3.90	3.97	3.32	2.79

(a) Realism

	RV	SC	AR	DC	PL	LL
Buddha	2.67	2.79	2.16	5.62	4.18	3.55
Brain	3.39	2.81	3.13	5.18	4.02	2.44
Cow	3.48	3.95	3.02	3.67	2.25	4.60
Femur	4.62	3.88	2.97	3.65	2.95	2.90
Max	3.16	3.72	2.16	5.48	3.51	2.95
Skull	2.97	3.25	3.41	4.86	2.72	3.76

(b) Aesthetics

Table 5.4: The average (a) realism and (b) aesthetic ranks listed for each model. The green marked mean ranks were the lowest ranks for each model and thus the best ranked technique.

The mean values of SC, LL, PL and RV were close together, while the values for AR and DC were markedly below and above the other values (see Table 5.3). However, by analyzing the mean value differences, there was a technique ordering for realism (1: AR, 2: LL, 3: SC, 4: RV, 5: PL, 6: DC) and aesthetics (1: AR, 2: PL, 3: LL, 4: RV, 5: SC, 6: DC). Instead, there were several equivalent median values. Figure 5.5 presents both results with the median and the mean value (black dot). The boxplots visually convey the similar results for RV and SC and emphasize the lower ranking for AR and the higher ranking for DC compared to the other techniques.

MODEL-BASED. A more detailed analysis investigating the results for realism and aesthetics for each model was performed in the second analysis step. This enabled a model-based insight into the technique rankings. The mean ranks for each technique per model are listed in Table 5.4. The mean ranks for realism were almost similar except for the skull, the femur and the cow model (compare Table 5.4a). These three models exhibit slightly different results for the lowest and highest aestehtic rank, too (compare Table 5.4b). As mentioned in the previous section, the majority of participants did not like the skull model and thus they did not asses the techniques properly. The femur model is illustrated with only a few lines, and therefore exhibits very slight differences between the stimuli images. Potential outliers or models that were difficult to asses were visually presented when illustrating the results as boxplots, as presented in Figure 5.6. Outliers were evenly distributed and thus not eliminated for this analysis. Nevertheless, the model-dependent results demonstrated that it was necessary to offer a choice of different models to gain comprehensive results.

5.6.2 *Statistical Analysis*

For a statistical analysis, we applied the Shapiro-Wilk test to the technique ranking results presented in Table 5.3 and based on that we chose non-parametric significant tests. Not normally distributed rank results were confirmed with a significant difference of $p \leqslant .05$ for realism and of $p \leqslant .01$ for aesthetics compared to a normal distribution for the model-independent and the model-based analysis step. The non-parametric Friedmann test and the Wilcoxon signed-rank test were ap-

	RV	SC	AR	PL	DC	LL
RV		z = - 2.654 p ≤ .025; r = -.23	z = - 6.117 p ≤ .001; r = -.53	z = - .362 p > .05; r = -.03	z = - 2.666 p ≤ .025; r = -.23	z = - 4.335 p ≤ .001; r = -.38
SC	z = - .252 p > .05; r = -.02		z = - 3.327 p ≤ .001; r = -.29	z = - 2.199 p = .025; r = -.19	z = - 4.230 p ≤ .001; r = -.37	z = - 1.725 p > .05; r = -.15
AR	z = - 3.576 p ≤ .001; r = -.31	z = - 3.820 p ≤ .001; r = -.33		z = - 5.731 p ≤ .001; r = -.50	z = - 6.929 p ≤ .001; r = -.61	z = - 2.219 p ≤ .025; r = -.19
PL	z = - .673 p >.05; r = -.05	z = - .847 p >.05; r = -.07	z = - 3.026 p ≤ .001; r = -.26		z = - 2.767 p ≤ .025; r = -.24	z = - 4.083 p ≤ .001; r = -.35
DC	z = - 5.276 p ≤ .001; r = -.46	z = - 5.600 p ≤ .001; r = -.49	z = - 7.782 p ≤ .001; r = -.68	z = - 6.471 p ≤ .001; r = -.56		z = - 5.847 p ≤ .001; r = -.51
LL	z = - .223 p >.05; r = -.01	z = - .089 p >.05; r = -.007	z = - 3.621 p ≤ .001; r = -.31	z = - .610 p > .05; r = -.05	z = - 6.104 p ≤ .001; r = -.53	

Table 5.5: The p and the corresponding *z-score* values for the two-tailed Wilcoxon signed-rank results for each pairwise technique comparison for realism (upper blue triangular matrix) and for aesthetics (lower yellow triangular matrix). If $z \notin [-1.96, 1.96]$, a statistically significant difference with $p \leqslant .05$ exists. The Pearson's correlation coefficient r defines the *effect size* (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$ effect). Green colored results represent statistically significant differences. Pairwise technique comparisons with no significant difference are colored red.

plied to test for statistically significant differences. The effect size r was calculated with the *z-score* values using the Equation 5 in Section 2.3.2. When applying the Friedman test, a statistically significant difference exists, if $\chi^2 \geqslant \chi^2_{crit}$. For $\alpha = .05$ and $df = 6 - 1$ the critical value is $\chi^2_{\alpha,df} = \chi^2_{.05,5} = 11.07$ (see Table A.1). The Wilcoxon signed-rank test compared the techniques pairwise for the model-based and model-independent analysis. Thus, a Bonferroni correction was required. The easiest method to use this correction is to use a critical value for p divided by the number of conducted test [53]. In this case, a statistically significant difference exists, if $p \leqslant .05/2$. Additionally, we determined the two-tailed significance values, since our postulated hypotheses are non-directional.

MODEL-INDEPENDENT. The non-parametric Friedmann test confirmed statistically significant differences between the techniques for realism ($p \leqslant .001, \chi^2 = 77.165$) and for aesthetics ($p \leqslant .001, \chi^2 = 97.322$). All Wilcoxon signed-rank results are listed in Table 5.5. Pairwise statistically significant differences for realism were confirmed with $p \leqslant .025$ and a low effect size between the techniques RV - SC, RV - DC, SC - PL , AR - LL, and PL - DC. Statistically significant differences with $p \leqslant .001$ and a medium or a high effect size existed between RV - AR, RV - LL, SC - AR, SC - DC, AR - PL, AR - DC, PL - LL, and DC - LL. Statistically no significant difference with $p = .36$ was confirmed for RV - PL and with $p = .085$ for SC - LL. A statistically significant difference for aesthetics with $p \leqslant .001$ was confirmed for AR and DC pairwise compared with any other technique and the effect sizes varied between $r = -.26$ for PL - AR and $r = 0.68$ for DC - AR. No significant differences existed between each pairwise comparison of RV, SC, PL and LL. In detail, the best and the worst ranked technique were significantly different compared to the other techniques. $H_{Realism}$ and $H_{Aesthetics}$, however, are highly unlikely, since there were techniques with statistically no significant difference.

χ^2	Buddha	Brain	Cow	Femur	Max	Skull
Realism	80.650	44.635	42.189	73.339	71.449	31.080
Aesthetics	97.977	59.558	39.705	29.073	76.209	35.252

Table 5.6: The χ^2 results of the Friedmann test for each model and task. A statistically significant difference exists, if $\chi^2 \geqslant \chi^2_{crit}$ with $\chi^2_{crit} \geqslant 11.07$.

The analysis determined a statistically significant difference for AR placed in the first rank and DC placed in the last rank for realism and aesthetics. Besides that, a different ranking result for aesthetics compared to the qualitative results was determined. For example, the technique PL tended to be placed in a smaller rank than SC according to the mean value calculation. A mean value calculation quantifies the difference of two ranks with the value one. Different rankings for PL and SC influence the mean and thus do not represent the participants' preference when comparing SC with PL. A few smaller ranks for PL that influence the mean calculation lead to a smaller mean than SC, and thus to a smaller quantitatively evaluated rank for PL. Moreover, the descriptive analysis determined that the boxplots were insufficient and deceiving, since the mean values were close together and no clear differentiation was visible.

MODEL-BASED. The Friedmann results of the second model-based analysis step are listed in Table 5.6 and the Wilcoxn signed-rank results are shown in Table 5.7. This analysis step was performed to verify the stimuli models and to identify difficult model geometries in terms of participants' preferences. Table 5.7 contains the results for realism (upper blue triangular matrix) and for aesthetics (lower yellow triangular matrix) for each model. Green check marks illustrate the confirmed significant differences between two techniques. A statistically significant difference exists if $p \leqslant .025$. All confirmed differences for realism for the buddha, the femur, and the Max model were highly significant with $p \leqslant .001$. In most cases, starting from the resulting rank order of the feature line techniques, neighbored

Buddha	RV	SC	AR	PL	DC	LL
RV		.765	.576	✓	✓	.623
SC	.649		.790	✓	✓	.412
AR	.080	.041		✓	✓	.277
PL	✓	✓	✓		✓	✓
DC	✓	✓	✓	✓		✓
LL	✓	.068	✓	.044	✓	

Brain	RV	SC	AR	PL	DC	LL
RV		.121	.105	✓	✓	.501
SC	.052		.647	✓	✓	.264
AR	531	.384		✓	✓	.087
PL	✓	✓	✓		.469	✓
DC	✓	✓	✓	✓		✓
LL	✓	.224	.050	✓	✓	

Cow	RV	SC	AR	PL	DC	LL
RV		✓	✓	.331	.139	✓
SC	.092		✓	✓	.776	✓
AR	.161	✓		.258	✓	✓
PL	✓	✓	✓		✓	✓
DC	.622	.529	.127	✓		.069
LL	✓	✓	✓	✓	.079	

Femur	RV	SC	AR	PL	DC	LL
RV		.166	✓	✓	✓	✓
SC	✓		✓	✓	✓	✓
AR	✓	✓		.860	.870	✓
PL	✓	✓	.699		.688	✓
DC	.050	.637	.120	.127		.050
LL	✓	✓	.744	.829	.050	

Max	RV	SC	AR	PL	DC	LL
RV		.400	✓	✓	✓	✓
SC	.199		✓	✓	✓	.149
AR	✓	✓		✓	✓	✓
PL	.269	.599	✓		.323	✓
DC	✓	✓	✓	✓		✓
LL	.650	✓	✓	.096	✓	

Skull	RV	SC	AR	PL	DC	LL
RV		✓	.191	✓	.310	✓
SC	.237		✓	.042	✓	.590
AR	.165	.378		.070	.845	✓
PL	.141	.061	-1.736		.082	.201
DC	✓	✓	✓	✓		✓
LL	-1.727	.059	.154	✓	✓	

Table 5.7: Each table contains the realism (upper blue triangular matrix) and the aesthetics (lower yellow triangular matrix) results for each model. A green check mark confirms a significant difference with $p \leqslant .025$ between the corresponding two techniques. If no significant difference was found, the p value is listed. (Table reprinted, with permission, from Baer et al. [10] © 2015, Springer-Verlag Berlin Heidelberg.)

realism ranks were statistically not significantly different. For example, regarding the result of the cow model the resulting order was 1: AR, 2: PL, 3: RV, 4: DC, 5: SC, and 6: LL. In this case, the difference of AR (first rank) and PL (second rank) was statistically not significant but AR compared to RV (third rank). Analyzing the first three ranks, AR, SC, and LL were dominant. In cases where AR was on the first rank (brain, buddha, cow, Max), the results were statistically not significantly different, with the second place only on the brain and cow model. Thus, the AR method was mostly ranked on the first place or the second place. The SC technique was placed once on the first place (skull), twice on the second place (brain, buddha), and once on the third place (Max). On the skull model, SC was statistically not significantly different compared to the second place, and on the Buddha model it was statistically not significantly different compared to the first, third, and fourth place. Regarding the second and the third place of SC, it was statistically not significantly different to the third and fourth place. The first rank of LL for the femur model was statistically a significant result. Furthermore, LL was placed on rank 2 twice and once on rank 3. As mentioned, mostly there was no significant difference between neighbored ranks, but mostly between two ranks, i.e., rank two and four. In summary, although the realism results were not uniquely defined, there was a tendency to the methods AR, SC, and LL. Moreover, Table 5.7 includes a few p values with $p \leqslant .10$. A one-tailed analysis of those technique pairs will result in a significant difference with $p \leqslant .05$. However, a second analysis with re-defined hypotheses is methodically not correct and a further evaluation is required. Counting the number of how often a technique was placed on a rank between $1-3$, there were AR: 5, SC: 4, LL: 4, PL: 2, RV: 2, DC: 1.

The aesthetics results were similar to the realism results regarding the significance. Mostly, neighbored ranks were not statistically significant. The best result was obtained by AR. AR reached twice the first place (buddha, Max) with significant difference to the second rank; it reached twice the second rank (cow, femur), whereas it was statistically not significantly different even to the fourth rank. The LL technique reached twice the first place (brain, femur), compare Table 5.4b. Statistically no significant difference compared to the second place was confirmed for LL applied to the brain model and to the femur model. There was statistically no significant difference even to the third place. The PL technique reached also twice the first place (cow, skull). On the cow model, PL was significantly better compared to the second place, and on the skull model it showed statistically no significant difference to the third rank. The results for this task exhibited several potential outliers for DC, as shown in Figure 5.6b. The Buddha, brain and Max results had up to seven outliers. According to the median and other DC results, we assume that the participants misunderstood rank one and thus ranked the technique vice versa. For this task, the techniques AR, PL, and LL were placed best. Counting the number of how often a technique was placed on a rank between $1-3$, there were AR: 5, RV: 4, SC: 3, LL: 3, PL: 3, DC: 0.

The statistical analysis showed that mostly neighbored ranks were statistically not significantly different, but analyzing the techniques that differ from more than two ranks, the difference was mostly significantly different. Statistically significant differences with low, medium and high effects existed and were confirmed.

5.7 LESSONS LEARNED

A comparative evaluation is suitable to define the best or worst illustration technique. In this case, it was not required to evaluate the techniques' individual properties. This evaluation focused on a pairwise comparison strategy, which is typical when choosing an illustration technique for a specific dataset and task. Several techniques are present and the illustrator or scientist had to chose one to generate a representative visualization. First, the viewer's attention is guided to two techniques and a decision is made in terms of which one is better or worse than the other technique. If more than two techniques are available, the winner is then compared to the next technique. This natural strategy was used and re-designed to perform an ordered ranking evaluation with comparative tasks.

If only the most preferred technique is interesting, a favorite selection will be sufficient. An extended comparison similar to our evaluation provides further insights into a technique ranking. This is useful to individually rank a technique if not all six presented techniques are available. Any two techniques out of the six feature line techniques can be classified and pairwise compared in terms of realism and aesthetics based on the resulting order.

When performing an evaluation with a huge number of participants that have a broad spectrum of educational background, it is necessary to design simple tasks with unique stimuli. It is important that all stimuli can be recognized by the participants without additional explanation to avoid individual instructions and thus different information transfer and evaluation requirements. Participants should be able to assess whether or not the stimuli represent a realistic and aesthetics representation of the original model. Otherwise, they are not able to solve the assessment tasks and compare the techniques. For example, the skull model on the one hand was a realistic anatomical dataset and therefore suitable for this evaluation but on the other hand, our participants were not used to such visualizations and really had problems with the aesthetics assessment. They did not liked this stimulus and therefore found it hard to rate any of the stimuli as the most aesthetic stimulus. This resulted in a random selection of the favorite technique for the skull model, since they were forced to rank the techniques with each technique in one rank. If more than one technique for each rank is possible this problem is avoided. However, such rating option would not contribute to a ranking or technique comparison result. The skull stimulus problem was not identified during the pilot study, since all participants of the pilot study were familiar with anatomical visualizations. Hence, it is necessary to chose stimuli that represent the target research domain, that are clearly comprehensible, and that enable the participants to solve the task mentally and physically. Additionally, we recommend that aesthetic assessments require the possibility to chose a *neither nor* option as possible answer.

For the qualitative analysis, we chose a rather rare method. Even though the *Schulze* method is not a common analysis approach in psychophysics, this method compared our techniques pairwise in terms of how often one technique is better than the other without quantifying the difference like a mean value calculation does. Hence, this method corresponds to the idea of an ordered ranking evaluation and to the procedure when selecting a technique for a model or a surface; a pair-

wise comparison. A qualitative evaluation is appropriate to gain insight into the personal preferences and to postulate hypotheses. However, the results just give a tendency and should not be seen as a definite statement. According to an existing research question or concrete hypotheses, it is essential to choose an appropriate study design: qualitative or quantitative or a combination of both. This evaluation was designed to analyze the results qualitatively and quantitatively, since the techniques were quantified by the number of assigned ranks. However, it is important that potential hypotheses are postulated in advance and that they are not derived from the qualitative results, to provide a correct experimental procedure.

If the evaluation is designed appropriately a qualitative and quantitative analysis is possible. To analyze quantitatively a numerical value the *dependent* variable has to be attached to the participants' opinions. Thus, even the subjective opinions can be quantified when providing a dimension or a scale like a rating scale. Since these ranks provide only ordinal information without any numerical distance information between the ranks, the existence of a statistically significant difference between the ranks has to be analyzed by the frequency distribution of assigned ranks. The number of assigned ranks can be used to quantitatively analyze the qualitative pairwise comparisons and the techniques' individual ranks within the six feature line techniques.

A model-based analysis is appropriate if a stimulus verification is necessary or if the detection of models with deviating rankings is desired. Since we aimed at a general ranking and the model-based analysis was performed additionally, we had to deal with multiple applied significance tests to the same data and thus to halve the p value to control the familywise error rate and correct a potential type I error. An increased number of applied tests results in a decreased and very restrictive critical value for significance. Therefore, it is recommended to minimize to number of applied test and to be selective about the comparisons.

5.8 SUMMARY

This chapter presented and discussed the design, implementation and analysis of a comparative feature line technique evaluation. In the experiment, personal preferences of 129 participants were qualitatively and quantitatively analyzed to evaluate and compare six widespread line drawing techniques. We realized the techniques' realism and aesthetic assessment by a ranking task from 1 to 6. This way, participants were forced to decide and to compare which technique is better than another technique. To verify the best ranked techniques for realism and aesthetics, we also asked the participants to select their favorite technique.

LIMITATIONS. As a limitation, pre-generated images were used as stimuli. Thus, the user was not able to interact with the surface model and could not set own thresholds to fine tune the illustration techniques. Screenshots are not suitable for techniques that are view- or illumination-dependent and benefit from model-viewer interaction. The static images were used to avoid lacking in concentration. After observing a few models including rotation, it is highly likely that the participants would lose concentration and reduce the interaction with the following

models. Therefore, the number of required participants to neglect the symptom of fatigue would increase. Furthermore, the evaluation investigated not the illustration generation process but personal preferences of a finished illustration. We did not evaluate the techniques with identical relevance parameters. In detail, we did not restrict the number or length of the generated lines. Thus, some techniques generate more lines or longer lines than others, which may influence the realistic or aesthetic perception. Therefore, it is recommended to evaluate the line drawing techniques based on identical input (relevance) parameters. Additionally, a follow-up study with more anatomical structures is suggested, since this evaluation comprised only three real anatomical models. As this evaluation was targeted for participants with a broad spectrum of educational background, we chose structures that were clearly recognizable. Further evaluations should integrate anatomical and patient-specific models and recruit medical experts. However, this study design and the shown models enabled an evaluation with 129 participants.

Initially, we performed a qualitative analysis. The qualitative analysis exhibited a tendency to the feature line methods: AR, SC, and LL. These three line techniques have also been selected as the most favorite techniques. The results combined with the techniques' properties and characteristics led to the recommendation for applying SC to medical or organic structures and AR to other models with sharp edges. To demonstrate the difference to the quantitative analysis and to verify the ranking results as well as the models, a descriptive and statistical analysis were performed, too. Statistically significant differences with low, medium and high effects existed and a statistically significant difference for AR placed in the first rank and DC placed in the last rank for realism and aesthetics were confirmed. In summary, the three different feature line techniques AR, LL, or SC are strongly recommended according to the qualitative and quantitative analysis. More work needs to be done and a more precise analysis of the underlying surface is essential for a more accurate guideline. Especially a detailed analysis of patient-specific data, since the anatomical structure visualization may be affected by segmentation artifacts that influence a feature line illustration and, hence, the personal preference.

6

QUANTITATIVE AND QUALITATIVE GHOSTED VIEW TECHNIQUE EVALUATION FOR THE PRESENTATION OF VASCULAR STRUCTURES AND EMBEDDED FLOW

This chapter is based on the following publication:

"Perceptual Evaluation of Ghosted View Techniques for the Exploration of Vascular Structures and Embedded Flow". Alexandra Baer, Rocco Gasteiger, Douglas Cunningham and Bernhard Preim. In *Computer Graphics Forum (EuroVis)*, 30(3), pp.811-820, 2011

In medical research and treatment planning it is often necessary to visualize multiple superimposed layers of spatial information. Multiple layers may represent different anatomic structures, e.g. an organ as outer layer and its vascular supply as inner layer. In other situations, the anatomy should be displayed at the same time as derived information such as biomechanical simulation [44], blood flow simulation [57], or nasal airflow [181]. One of the earliest examples was radiation treatment planning [72], where isodose distribution was displayed along with anatomic structures. Since the internal structures are spatially embedded in the surrounding structure, occlusions where objects or surfaces that are nearer to the observer hide more distant ones have to be tackled. Thereby, a trade-off between the visibility of internal information and the simultaneous depiction of the 3D shape of the enclosing surface has to be found. Moreover, to avoid the misleading interpretation of spatial relationships, enhancement of depth is important.

Smart visibility techniques, like cutaway, ghosted, section and exploded views are used to generate focus-and-context and thus customized visualizations (see Section 3.1.2). When exploring blood flow data, streamlines representing the flow are the focus and the vascular surface is the context. There are, however, still no clear guidelines to decide which technique should be used in different situations. Even less clear is when – let alone how – the chosen technique should be combined with others to improve effectiveness. Smart visibility techniques have rarely been systematically evaluated, despite their huge potential for surgery and therapy planning. The majority of existing user studies are primarily informal or questionnaire-based (qualitative) and, thus reflect more the personal beliefs and preferences of the participants than the actual objective measurement of task performance with different techniques. Without an extensive, controlled perceptual evaluation, the techniques' potential remains underutilized.

Gasteiger et al. [57] developed two ghosted view techniques to illustrate the cerebral aneurysm anatomy with embedded flow. Moreover, they presented an initial survey that is primarily informal and questionnaire-based to evaluate their developed ghosted view techniques. This qualitative analysis, however, reflects more the personal beliefs and preferences. After adopting the illustration techniques according to the first evaluation results, a more detailed analysis was required. In this chapter, three controlled perceptual experiments quantitatively and qualitatively evaluating the two developed smart visibility techniques will be introduced. Additionally, Rocco Gasteiger implemented a software framework to conduct the experimental study. The author of this thesis primarily designed and performed the evaluation and analyzed the recorded results.

The illustration techniques were evaluated in the context of the visualization of cerebral aneurysms, where the vascular anatomy should be displayed along with the internal flow to promote the exploration of flow-vessel correlations. The common semitransparent visualization technique is compared with the ghosted view and the ghosted view with depth enhancement technique from Gasteiger et al. [57]. The techniques' ability to facilitate and assist the shape and spatial representation of the aneurysm models as well as to evaluate the smart visibility characteristics to promote the flow exploration were considered. We evaluated the techniques with respect to the participants' task performance (accuracy and time) and with respect to their personal preferences. This controlled perceptual evaluation has been published in Baer et al. [8].

6.1 MEDICAL BACKGROUND AND DATA FLOW PIPELINE

Gasteiger et al. [57] aimed at an adapted visualization of the enclosing aneurysm surface that depicts shape and spatial perception whilst simultaneously gaining maximum visibility of the embedded flow visualization. Cerebral aneurysms represent a local widening of cerebral vessels, which is a serious disease (high rupture risk). Furthermore, Gasteiger et al. [57] presented the medical requirements for cerebral aneurysm analysis in detail. Briefly, neuroradiologists analyze morphological features of the aneurysm, e.g., they search for satellites (high local bulge on the aneurysm, indicating a former bleeding) and they analyze the inflow and outflow region to assess an aneurysm's rupture risk. The data flow pipeline consists

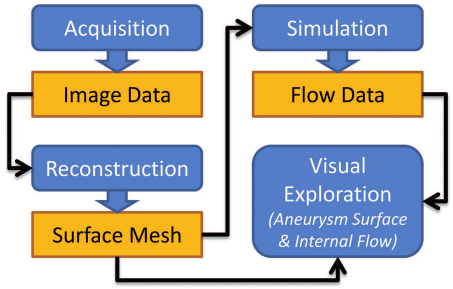


Figure 6.1: Data flow for visual exploration of blood flow in cerebral aneurysms. (Image reprinted, with permission, from Gasteiger et al. [57] © Eurographics Association 2010.)

of three processes, illustrated in Figure 6.1. In the acquisition step, clinical image data (MRA, CTA, or 3DRA) of the aneurysm is obtained. The aneurysm surface is reconstructed using a simple threshold followed by a connected component analysis to separate the aneurysm and its parent vessel from the surrounding tissue. The overall segmentation takes about 10 minutes for clinical datasets. Based on the segmented mask, the surface morphology of the aneurysm is reconstructed and optimized with respect to mesh quality [30]. The resulting mesh is used for constructing a computational grid, on which a CFD (Computational Fluid Dynamic) simulation is performed. Regarding the resulting blood flow data, Gasteiger et al. [57] only considered information about the flow direction and velocity. The surface mesh and the flow information form the input for the final visualization. Although this pipeline was developed for a specific purpose, it is similar to other pipelines, where simulations are involved [181, 44].

6.2 VISUALIZATION TECHNIQUES

The perceptual evaluations focused on three different visualization techniques to depict the surface of the vascular structure. In the exploration of blood flow data, the internal flow is the focus object, since this is the most important visual information and the vascular structure serves as context. Gasteiger et al. [57] introduced two ghosted view techniques applied to the enclosing aneurysm surface. The medical background for cerebral aneurysm diagnosis was discussed by Augsburger et al. [4]. Both ghosting techniques were developed to depict shape and spatial perception, whilst simultaneously gaining maximum visibility of embedded flow visualization and thus to improve the previously applied semitransparent visualization technique. In all techniques, the flow information was depicted with color-coded streamlines where the color represents the local velocity (see Figure 6.2). An optimized color scale according to Levkowitz [106] was used to enhance the quantitative character of the velocity data. Moreover, the same light conditions were applied: which is a white light pointing from the upper left.

6.2.1 *Semitransparency*

The first technique is a semitransparent (**S**) surface rendering as a common visualization method for enclosing surfaces. The transparency was set to 0.5 and was implemented with depth peeling to get a correct blending (see Figure 6.2, upper row). As surface color, we used a bright brown which is distinguishable to the color scale of the streamlines. Since a few recruited participants had only passing or no medical knowledge, this color facilitated the vascular structure recognition. The bias factor of this color, however, was not investigated and thus not quantified. Because of the global transparency, depth cues to convey the aneurysm surface shape are strongly reduced. Additionally, regions where several layers of semi-transparent surfaces lie in front of each other, can cause a misleading interpretation of spatial relationships. Both problems decrease the observers' ability to mentally link the aneurysm morphology with the internal flow data.

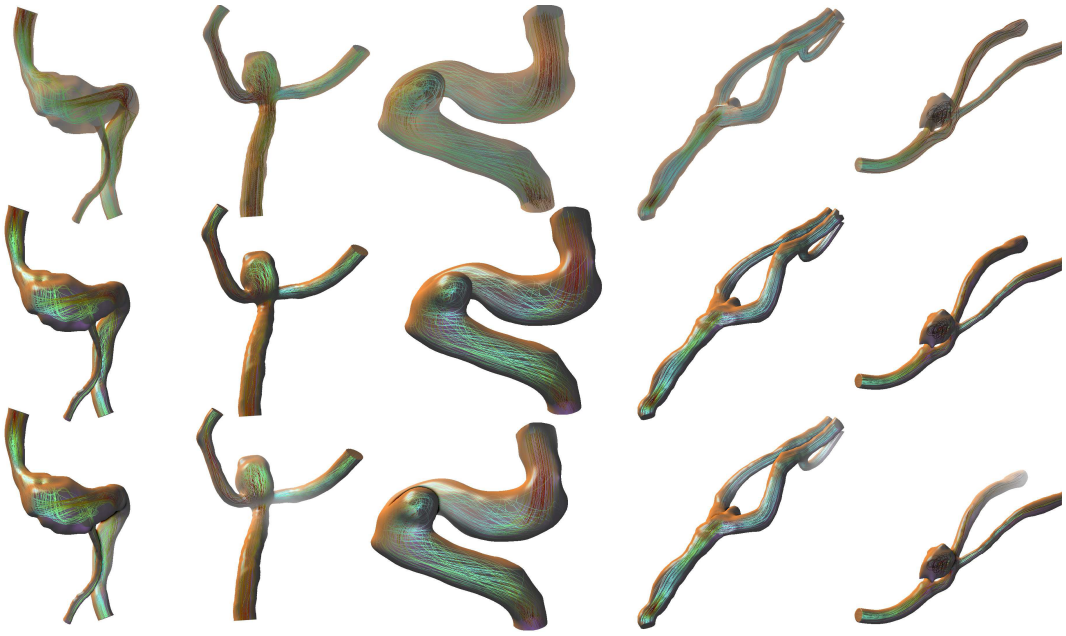


Figure 6.2: Three visualization techniques are evaluated by means of five datasets. The 3D aneurysm models of the upper row are visualized with the semitransparent **S**, in the middle row with ghosting **G** and in the bottom row with the ghosting with depth enhancement **GD** technique.

6.2.2 *Ghosting*

Inspired by smart visibility techniques [162], we investigated the ghosted view by means of a view-dependent transparency rendering. This ghosting technique (**G**) developed by Gasteiger et al. [57] was the second visualization method. An approximation of the Fresnel reflection model [144] was employed on the front faces of the aneurysm surface and replaced "reflection" with "opacity". The front face color was the same as in the **S** technique and the back face color was a cool color according to Gooch et al. [59]. A better shape enhancement (due to more opacity) at surface regions facing away from the viewer as well as a maximum visibility (due to less opacity) of the embedded streamlines facing to the viewer (see Figure 6.2, middle and bottom row) was achieved. Additionally, specular reflections on the front faces were integrated to enhance the shape perception. The opaque regions, however, occluded some of the embedded streamlines which can be disturbing during flow examination.

6.2.3 *Ghosting with Depth Enhancement*

The third visualization extends the ghosted view visualization by means of additional depth enhancement and is called ghosting with depth enhancement (**GD**). Shadow and atmospheric attenuation as two important depth cues were added to the surface and to the streamlines (see Figure 6.2, bottom row). By applying the method of Luft et al. [111], shadow casting was approximated in a non-physically correct way. A modified depth buffer of the surface front faces was employed to achieve a constant appearance of shadow casting during interaction (e.g. zooming). Atmospheric attenuation was introduced by applying fog which makes the

objects fade with increasing distance. Because flow attributes, like velocity, were color-coded, color modifications were not appropriate, since they can cause a misleading interpretation of the coded attribute.

6.3 EXPERIMENTAL DESIGN

The primary research goal of this study was to measure whether and how the techniques G and GD facilitate the assessment process of cerebral aneurysms and the internal blood flow compared to the common S visualization technique. Initially, the properties and potential advantages or differences of the ghosting technique had to be specified during the goal definition step to postulate appropriate alternative hypotheses and define suitable tasks, see Section 2.1. Then, the ghosting technique had to be analyzed and compared with S. Besides a qualitative analysis of personal technique preferences, the three techniques were quantitatively compared. The techniques were evaluated with respect to their ability to fulfill the following three requirements that are essential when assisting the exploration of flow-vessel correlations:

1. **Perceptually effective shape representation**: Accurate perception of the shape and curvature of the various surfaces is essential for estimating the aneurysm's risk of rupture and options for treatment planning. Together with the internal flow characteristics, the surface morphology plays an important role in the risk assessment process.

2. **Showing embedded structures:** This feature defines the smart visibility characteristics of the technique. It refers to the visual perception of the flow visualization, since current medical research focuses on the simulations of intravascular blood flow [4]. The additional information assists the decision-making process and disease understanding.

3. **Perceptually effective spatial representation of the aneurysm's parent vessels:** A special subset of scene perception is the understanding of the relative location in depth. This is particularly critical in smart visibility techniques. Moreover, understanding the spatial arrangement (depth ordering) of the aneurysm's blood vessel structure helps to improve the understanding of overall flow characteristics (inflow and outflow regions) and this promotes aneurysm rupture risk assessment.

The quantitative evaluation was realized by means of the participants' task performance that was defined by accuracy and task completion time. Thus, the hypotheses were split into an accuracy and a task completion time part to enable an individual analysis and prevent being bound to the other task performance parameter. We postulated the following one-tailed hypotheses:

- $H_{Shape_{Acc}}$: G and GD facilitate the shape perception of the vascular structure – as measured by *accuracy* – better than S.

- $H_{Shape_{Time}}$: G and GD accelerate the shape perception of the vascular structure – as measured by the *task completion time* – better than S.

- $H_{Smart_{Acc}}$: G and GD facilitate the assessment of embedded flow – as measured by *accuracy* – better than S.
- $H_{Smart_{Time}}$: G and GD accelerate the assessment of embedded flow – as measured by the *task completion time* – better than S.
- $H_{Spatial_{Acc}}$: G and GD facilitate the spatial perception (depth ordering) of the vascular structure – as measured by *accuracy* – better than S.
- $H_{Spatial_{Time}}$: G and GD accelerate the spatial perception (depth ordering) of the vascular structure – as measured by the *task completion time* – better than S.

Since the three criteria shape representation, smart visibility and spatial relation perception were effectively independent of each other, we decided to separate the evaluation and design three individual but customized experiments. The first experiment tackled the techniques' shape representation potential, the second experiment evaluated the smart visibility benefit and the third experiment analyzed the spatial representation. Moreover, we assume that:

- GD will be more effective in each category than **G** and **S**.

The first factor in all three experiments was a *within factor* (visualization technique) which had three *levels* (S, G and GD). Thus, the visualization technique was the *independent variable* and the measured parameters accuracy and task completion time were the *dependent variables*, compare Section 2.1.3. Accuracy was defined according to the tasks for shape, smart visibility and spatial perception and will be described in Section 6.3.2.

Furthermore, the study concerning the shape perception had a second, *between-participant factor*: Interaction. Since the ghosting technique aimed at a better shape perception by increasing the opacity of the surface at regions facing away from the viewer and providing maximum visibility (less opacity) at regions facing to the viewer, this technique benefits from rotation and interaction with the structure [56]. Thus, one group saw all stimuli as static models, while the other group could rotate the models within limits, and thus the shape evaluation followed a *between-participant* design.

6.3.1 *Participants*

We recruited a total of 86 participants from various parts of the university like psychology and engineering students as well as four medical experts and two flow simulation experts who participated in the final experiments. For the shape experiment there were two groups of 17 participants each. Six women and eleven men participated in the first group ($\bar{x} = 28.72$ years) and 10 women and seven men in the second group ($\bar{x} = 31.43$ years), all aged between 20 and 35 years. One woman and three men participated in the pilot study (aged between 25 and 32 with $\bar{x} = 27.5$ years). For the smart visibility evaluation 27 participants were recruited with twelve women and 15 men aged between 22 and 36 and an average age of $\bar{x} = 32.26$ years. Two women and two men participated in the pilot study (aged between 29 and 37 with $\bar{x} = 33.25$ years). The spatial perception evaluation

was conducted by 25 participants (15 women and 10 men) aged between 18 and 30 years with an average age of $\bar{x} = 26.52$ years. Four men participated in the spatial pilot study (aged between 26 and 35 with $\bar{x} = 30.75$ years).

6.3.2 *Task Methodology*

In order to investigate accuracy and time, the participants were asked to fulfill individual tasks in each experimental evaluation. Each task was customized to the specific evaluated shape, smart visibility or spatial perception requirements and enabled a quantitative accuracy analysis as well as a task completion time measurement.

THE SHAPE PERCEPTION EXPERIMENT was performed to evaluate the techniques' shape representation ability. It was based on the gauge methodology, which is known from visual psychophysics. As introduced in Section 3.3, this method was employed to obtain local estimates of surface orientation and thus to analyze the perceived surface shape. Participants had to orient the gauge figures placed on a 3D surface to coincide with the apparent surface normal at that specific surface point. Later on, the individual normal estimate was compared with the *gold standard* surface normal vector at this specific point and thus the shape perception accuracy was analyzed. The gold standard normal vectors (gs_{normal}) were determined by using a corresponding full-color shaded opaque stimulus with the gauge figure placed at the same surface position. Since we expected the participants to perform most accurately on shaded models, the full opaque stimuli were used to define the participants' *perceived ground truth* referred to as the *gold standard*. This perceived ground truth was used instead of the calculated normal vector at a surface point to define the most accurate perceived surface normal without any specific visualization technique but an opaque shaded surface. This evaluation was aiming at perceptual validity instead of physical validity. Thus, shape stimuli with gauge figures and corresponding full opaque stimuli were required (see Section 6.3.3).

THE SMART VISIBILITY EXPERIMENT was conducted to assesses the techniques' ability and effectiveness to show embedded structures, in this case the blood flow. The accuracy of flow perception indicated the smart visibility characteristics of the illustration techniques. To validate the techniques' smart visibility potential, color-coded streamlines were used that represent the blood flow. This experiment was designed to quantitatively evaluate the embedded flow perception and not to evaluate the flow visualization itself. Participants were asked to define the average flow color at different regions to analyze the techniques' smart visibility potential. These individually estimated average color values were compared to a gold standard color value to determine the average color error. All three techniques were evaluated by the perceptual difference between the estimated color and the perceptual gold standard color (gs_{color}). Similar to the shape perception evaluation, a perceptual color ground truth was used as the gold standard (gs_{color}) and had to be defined for each tested region, since we were aiming for perceptual validity. In detail, a second kind of stimulus was required that illustrated the correspond-

ing flow without being occluded or influenced by the surface visualization and thus enabled a determination of the maximum perceived color (reference color) (see Figure 6.4, bottom row). The participants were asked to define the average flow color for these stimuli, too. Based on that, perceptual gold standard color values were calculated. Thus, two kinds of smart visibility stimuli were required, see Section 6.3.3.

THE SPATIAL RELATION EXPERIMENT was designed to analyze the spatial perception. As described in Section 3.3, a common strategy to analyze the depth and spatial perception is to ask the participants to determine the perceived depth ordering and thus to perform a depth judgment task. Due to this depth judgment strategy, the vessel branches of the aneurysm models were used to investigate the spatial relation perception and to analyze the accuracy according to $H_{Spatial_{Time}}$. Participants had to determine which branch is closer to the viewer. This evaluation required stimuli with different depth-ordered vessel branches.

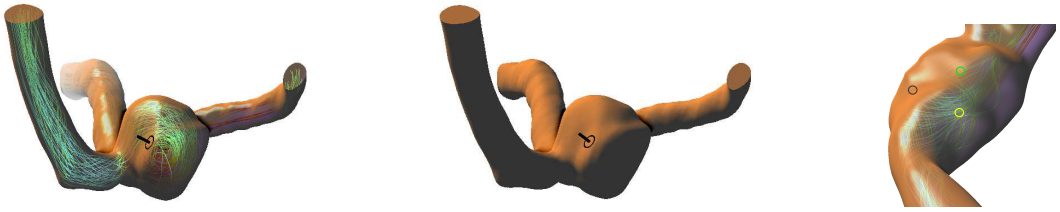
6.3.3 *Stimuli*

The stimuli were carefully chosen to be representative renderings of 3D patient-individual aneurysm models. Three of the presented models were generated by Gasteiger et al. [57] using clinical datasets according to the presented data flow pipeline. They were either derived from a magnetic resonance angiography or a computed tomography angiography and had a size of $695 \times 768 \times 149$ and $695 \times 768 \times 136$ in the x-, y- and z-axis. The resulting meshes consisted of 56 292 (first model), 46 579 (second model) and 19 234 (third model) triangles (compare Figure 6.2). Additionally, two aneurysm models were provided by Gasteiger et al. [57] as meshes with 577 744 and 271 112 triangles. All 3D models were visualized with one of the three techniques developed by Gasteiger et al. [57] (recall Figure 6.2). A Phong shading that interpolates surface normals along the edges of each triangle and subsequently computes the final pixel color was implemented. Thus, a high visual quality, which strongly reduces the visible influence of the mesh resolution was achieved. Since the surface was shaded across the triangles boundaries, the mesh resolution did not influence the visible results within in the surface. However, a reduced mesh resolution can only disturb a smooth surface perception at the surface-background boundary where the edges might become visible without influence of the employed shading technique.

SHAPE STIMULI were 3D aneurysm models visualized with S, G, GD or visualized as an opaque shaded surface that were required for the gs_{normal} determination, as introduced in Section 6.3.2. For the stimuli, viewpoints corresponding to preferred views of neuroradiologists were chosen. A single gauge was placed on the front side of the model's surface and initially pointed to the viewer. A gauge was drawn as a small ellipse representing a disc and a single line indicating the normal of the disc. Additionally, the gauge figure positions were carefully chosen based on the three different surface areas that are characteristic for ghosting techniques and illustrated in Figure 6.3 (c):

- opaque surface areas,
- semitransparent surface areas that occur between transparent and opaque regions, and
- areas primarily in the focal view of the user, where the surface is completely transparent.

To avoid cueing the participants to shape, the gauges did not penetrate or interact with the 3D model. Derived from the five datasets, five models were available and thus five stimuli for each technique. To effectively increase the number of stimuli, viewpoints were changed, which can help to reduce the ability to recognize an object as being identical to one already seen. These changed viewpoints are still similar to preferred views of neuroradiologists. Therefore, nine instead of the original five aneurysm views were available. We did not change the views for all models, since the third model visualized in Figure 6.2 was easy to recognize even with a changed viewpoint. Thus, the stimuli were generated using these nine views, each presented with one of the four techniques (S, G, GD, and opaque shaded) and with a single gauge placed on the model's surface. To enable an analysis of each technique and each characteristic region, three of these nine models were presented with gauge figures placed in opaque regions, three in semitransparent, and three models with a gauge placed in fully transparent regions. In summary, a total of 36 stimuli were generated for the shape perception experiment.



(a) (b) (c)

Figure 6.3: (a) The ghosting with depth enhancement (GD) and (b) the full shaded opaque surface visualization with a gauge figure and the line representing the estimated surface normal position. The gauge figure is a small ellipse representing a disc and a single line. (Images reprinted from Baer et al. [8] © John Wiley & Sons 2011 with kind permission from John Wiley & Sons, Inc.) (c) Gauge figures were placed in opaque (black circle), semitransparent (green circle) and fully transparent (yellow circle) surface areas.

THE SMART VISIBILITY STIMULI were static renderings generated from the five 3D aneurysm models. For this evaluation, two different viewpoints were used for model 1, 2 and 5 (compare Figure 6.2) and this resulted in a total of eight used viewpoints. Model 3 and 4 were not used twice, since model 3 was easy to recognize and for model 4 it was difficult to find a further representative viewpoint with respect to a flow visibility assessment task. Figure 6.4 illustrates the stimuli. Each stimulus was overlaid with a pink rectangle selecting a certain surface and the corresponding flow region. A rectangle was placed in semitransparent and fully transparent surface regions according to the characteristic ghosting regions mentioned above. Since, by definition, an opaque surface will fully occlude the

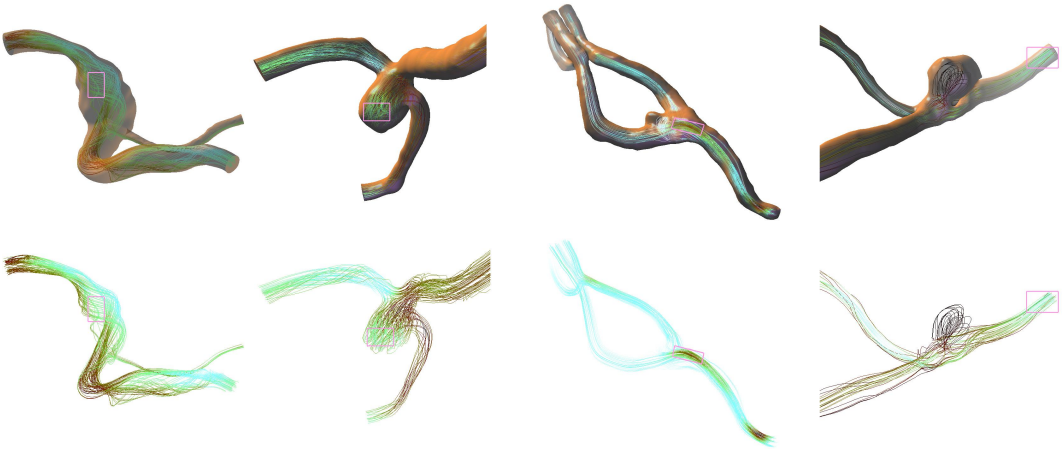
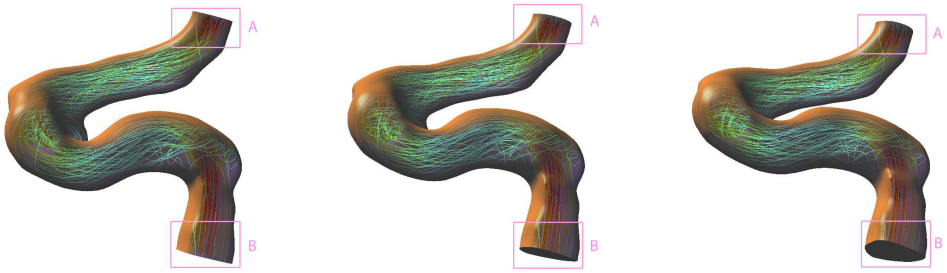


Figure 6.4: The upper row illustrates the used models and the changed viewpoints. The lower row represents flow stimuli do determine the corresponding gs_{color} for this specific region.

embedded structure, it will not be possible to determine the flow color in opaque regions. Obviously, then, S is superior to G and GD in this regard: The S technique allows you to see the flow everywhere. Since, however, the visualizations were designed with a specific region of interest in mind (the region that is transparent in the ghosting techniques), it is also logical to focus the experiments on the perception of flow color in those areas. Thus, only the semitransparent and fully transparent regions were to be investigated for this evaluation part. Eight model viewpoints were used, whereas four were overlaid with a rectangle positioned in transparent and four with a rectangle in semitransparent regions. Each model was visualized with either the S, G or GD technique, which accounts for 24 stimuli. To determine the gs_{color} for each region, eight additional stimuli images illustrating only the flow on a white background without a surface visualization were generated. This resulted in a total of 32 stimuli for this evaluation.

THE SPATIAL STIMULI were static renderings generated from four 3D aneurysm models. We left out generating stimuli of the fourth model illustrated in Figure 6.2, since the two branches were too close to separate them and thus not suitable for this depth judgment task. Stimuli illustrating the aneurysm models with two branches vertically aligned were generated to analyze the perceived depth ordering of the visualized vascular structures. To provide a depth ordering of the vessel branches and enable a depth judgment, the models were rotated around the aneurysms' x-axis clock- or counterclockwise by 0°, 10°or 20° (see Figure 6.5) The task-specific vessel branches were marked with pink rectangles labeled with *A* (upper vessel branch) and *B* (lower vessel branch) to focus the participants on these vessel regions similar to Ropinski et al. [138], who highlighted the focus vessel branches in angiography images with flashing square shaped outlines. Each participant saw stimuli of four models, each model rotated with three different angles and visualized with S, G and GD. 12 stimuli illustrated both vessel branches with a same distance, 12 with *A* being closer to the viewer and 12 with branch *B* being closer to the viewer. In summary, a total of 36 stimuli were generated for this evaluation.



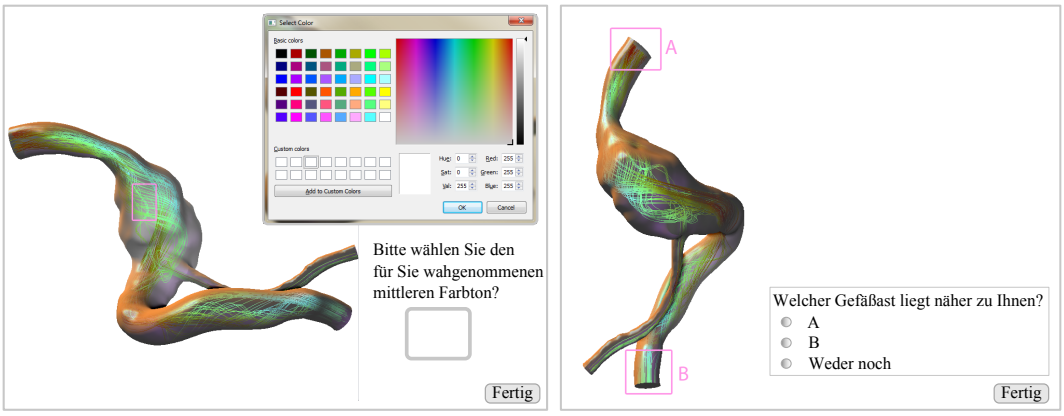
(a) 0° (b) 10° (c) 20°

Figure 6.5: Spatial stimuli for one model and one technique (G). (a) Both vessel branches have the same distance to the viewer (rotation angle of 0°). (b) The model rotated by 10° and (c) 20°, whereas the vessel branch B is closer to the viewer. (Images reprinted from Baer et al. [8] © John Wiley and Sons 2011 with kind permission from John Wiley and Sons.)

6.4 APPARATUS AND PROCEDURE

All participants were tested under the same conditions to produce meaningful results and to avoid discriminations. This means that all of them performed the experiment alone by daylight on a 26″ monitor at a full HD resolution of 1920×1200 pixels and no other processes were run on the computer during the experimental session. The participants viewed the stimuli from a distance of approximately 0.7m (each stimulus image subtended 17.8° of a visual angle on average). The evaluation framework was a stand-alone application based on VTK and Qt that included the stimuli presentation, handling of the user input, and storage of the user's response. The application was able to present the stimuli as 3D renderings, where interaction is available, or as 2D images. All participants interacted with a mouse device to solve the corresponding tasks.

Initially, small pilot studies with four participants each were conducted. The pilot studies enabled a test and refinement of the experimental design before starting full-fledged studies. Individual changes will be explained in the corresponding paragraphs below. Prior to the start of each experimental session, all observers were instructed in written form to provide the same initial requirements. One practice trial followed the instruction to familiarize each participant and to ensure that the participants understood the experimental task. Each stimulus was positioned in the center of the screen, one at a time and displayed on a white background. All evaluated illustration techniques as well as the aneurysm models were presented in random order to avoid expectation effects. The stimuli that were required to define gs_{normal} (nine stimuli) and gs_{color} (eight stimuli) were presented at the end. The participants were asked to perform different tasks according to the experimental session. Each stimulus was presented until the participants pressed a "Ready" button to indicate that they were satisfied with their answer and ready to move on to the next stimulus. Moreover, the "Ready" button was used to determine the individual task completion time for each stimulus. After finishing the experiment, the participants were asked to evaluate the techniques using a bipolar 5-point Likert rating scale, with each pole representing the performance of one technique (*technique A* *very good*, *good*, *neutral*, *good*, *very good* *technique B*).



(a) (b)

Figure 6.6: (a) For the smart visibility experiment, participants were asked to define a flow corresponding color using the CColorDialog of the Microsoft Foundation Class. (b) The task of the spatial experiment comprised a depth judgment of vessel branches. Participants had to define which branch is closer to the viewer using. Possible answers are branch *A*, branch *B* or *None*.

Thus, participants had to rate the techniques and had to compare the techniques' benefit pairwise. Based on that, a qualitative comparison of the techniques was possible, too. Finally, a short questionnaire asking for some personal details had to be filled out.

SHAPE EXPERIMENT. Participants were shown the shape stimuli and asked to orient the gauges to coincide with the apparent surface normal at that specific surface point. The orientation of the gauge was controlled by the mouse. As already mentioned, they had no control over gauge positions, each gauge initially pointed to the viewer and the gauges did not penetrate or interact with the 3D model. This experiment was divided into two groups based on the provided interaction with each stimulus. While the gauge interaction was the same for both groups, one group received static stimuli $\text{group}_{-\text{RO}}$ (group without rotation option) and one had the opportunity to rotate the stimuli $\text{group}_{+\text{RO}}$ (group with rotation option). The pilot study of $\text{group}_{+\text{RO}}$ showed that participants used the rotation option to orientate the model so that the gauge figure was on the model's silhouette. In detail, the gauge figure was at a 90° angle to the viewer and the participants, therefore, vastly simplified the task. Since the experiment's goal was to evaluate the perception for the given view, which was the preferred view for specialists, and not the perfect surface normal vector, the range of possible rotations was restricted to 15°. Thus, the shape representation of the technique benefits from the rotation and the 15° rotation option still prevents rotating the gauge to the silhouette.

SMART VISIBILITY EXPERIMENT. Each participant saw the same kind of stimuli introduced in Section 6.3.3. Their task was to define the average flow color for each stimulus-specific region. Therefore, the CColorDialog of the Microsoft Foundation Class was provided and the participants had to define a flow corresponding color. The next stimulus was presented after the participants selected a color, as illustrated in Figure 6.6a.

SPATIAL EXPERIMENT. The participants were asked to determine which branch is closer to the viewer. Possible answers were branch *A*, branch *B* or *None*, if both branches have the same distance to the viewer. Answer *None* indicates that the participants either perceived both vessel branches with the same distance or they were not able to decide which one is closer, which basically means the same for this question. A radio button dialog was provided for answering this question, see Figure 6.6b.

6.5 QUANTITATIVE ANALYSIS AND RESULTS

The accuracy, task completion times, and subjective technique preferences were analyzed. Accuracy for shape perception was determined by analyzing the perceived surface normal vectors, for smart visibility by analyzing the perceived blood flow color, and for the spatial experiment by determining the correct and false answers of the depth judgment task. Initially, all gathered data were tested for a normal distribution with the Shapiro Wilk test, since this is a major requirement for choosing an appropriate significance test and consequently achieving valid results. The test determined that the results were statistically significant different with $p \leqslant .05$ for accuracy and with $p \leqslant .001$ for task completion time compared to a normal distribution and were thus not normally distributed. To test the postulated hypotheses (recall Section 6.3), statistical significance tests were run on the results of the experiments. Non-parametric tests were applied, since the results were not normally distributed. First, the Friedmann test was applied to analyze the results and to determine if a statistically significant difference for the accuracy and the task completion time results exists. When applying the Friedman test, a statistically significant difference exists, if $\chi^2 \geqslant \chi^2_{crit}$. For $\alpha = .05$ and $df = 3-1$ the critical value is $\chi^2_{\alpha,df} = \chi^2_{.05,2} = 5.99$ and for $\alpha = .01$ it is $\chi^2_{.01,2} = 9.21$ (see Table A.1). Second, the Wilcoxon signed-rank test was used for the pairwise comparison of the technique results. Moreover, for the shape experiment we used the Wilcoxon-Mann-Whitney U test to compare $group_{-RO}$ and $group_{+RO}$, since this non-parametric test determines if there is a significant difference between two independent samples. The effect size r was calculated using the Equation 5 in Section 2.3.2. The software package IBM SPSS STATISTICS (Statistical Package for the Social Sciences) provided the statistical test and was used for the descriptive and statistical analysis.

6.5.1 *Shape Perception Results*

For the shape experiment, the x-, y- and z-components of the apparent surface normal vector were recorded for each gauge figure as well as the task completion time. The gauge figure estimates of the opaque shaded models were used as the participant's perceived ground truth (gs_{normal}) for each gauge figure and model. The average angular difference between the surface normal estimate and gs_{normal} defined the technique's accuracy and the technique's potential to support the shape perception, respectively. Thus, the scalar product of both vectors was determined and based on that the angular deviation in degrees was calculated. The average angular deviations, the standard deviations and the results of the Shapiro-Wilk test for both groups are listed in Table 6.1. The accuracy results for $group_{-RO}$

	Group -RO			Group +RO		
	$\bar{x}$	σ	Shapiro-Wilk	$\bar{x}$	σ	Shapiro-Wilk
S	30.01	6.55	$p \leq .05$	26.38	5.82	$p \leq .05$
G	28.86	3.27	$p \leq .05$	21.23	3.68	$p \leq .05$
GD	28.75	4.04	$p \leq .05$	23.91	4.12	$p \leq .05$

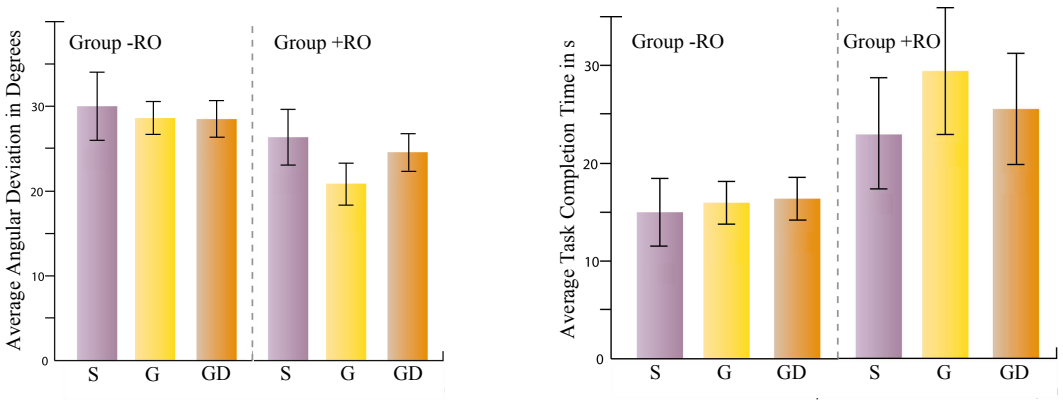
(a) Accuracy

	$\bar{x}$	σ	Shapiro-Wilk	$\bar{x}$	σ	Shapiro-Wilk
S	15.03	8.63	$p \leq .05$	22.80	10.87	$p \leq .05$
G	15.80	6.07	$p \leq .05$	28.52	14.98	$p > .05$
GD	16.23	6.49	$p \leq .05$	25.60	12.59	$p \leq .05$

(b) Time

Table 6.1: This table covers the mean values ($\bar{x}$), standard deviations (σ) and the Shapiro-Wilk test results for each illustration technique (a) for the accuracy in average angular deviation in degrees (°) and (b) for the task completion time in seconds s. Shapiro-Wilk results with $p > .05$ represent normally distributed results while $p \leqslant .05$ means that a statistically significant difference from a normal distribution exists.

were between 28.75° and 30.01° and for $group_{+RO}$ they were between 21.23° and 26.38° (compare Table 6.1a). Participants of $group_{-RO}$ were more precise with GD and of $group_{+RO}$ with G to the descriptive analysis results and the frequency distribution results, as illustrated in Figure 6.7a. The average task completion times for $group_{-RO}$ were 15.03s for S, 15.80s for G, and 16.23s for GD. Participants of $group_{+RO}$ were slower and required 22.80s when the aneurysm was semitransparently visualized (S), 28.52s when ghosting (G) was applied and 25.60s when an additional depth enhancement (GD) was used (see Table 6.1b). As illustrated in Figure 6.7b, $group_{+RO}$ required more time orienting the gauges and the results have higher standard deviations, which indicates that the time results were spread out over a wider range of values compared to $group_{-RO}$. However, $group_{+RO}$ had



(a) Accuracy (b) Time

Figure 6.7: (a) The average angular differences of the normal estimates and the corresponding gs_{normal} and thus the accuracy of S, G and GD and (b) the average task completion time results in seconds illustrated as bar charts including standard error bars for $group_{-RO}$ and $group_{+RO}$.

smaller angular errors, which means that participants oriented the gauge figures more accurately.

To test the postulated hypotheses, statistical significant tests were applied. For the techniques' accuracy results the Friedmann test confirmed a statistically significant difference for $group_{+RO}$ with $p \leqslant .05$, $\chi^2 = 7.28$. No significant difference was found between any two visualization techniques for $group_{-RO}$ with $p > .05$ and $\chi^2 = .118$ (see Table 6.2). Contrary to that, the task completion time is statistically significant different for $group_{-RO}$ with $p \leqslant .05$, $\chi^2 = 6.118$ and for $group_{+RO}$ with $p \leqslant .05$, $\chi^2 = 12.118$.

Comparison	Group -RO		Group +RO	
	Accuracy	Time	Accuracy	Time
S - G	z = - .260 p > .05; r = -.06	z = - 1.661 p ≤ .05; r = -.40	z = - 1.730 p ≤ .05; r = -.41	z = - 2.817 p ≤ .05; r = -.68
S - GD	z = - .402 p > .05; r = -.09	z = - 1.681 p ≤ .05; r = -.40	z = - 1.823 p ≤ .05; r = -.44	z = - 1.634 p >.05; r = -.39
G - GD	z = - .710 p > .05; r = - .17	z = - .282 p > .05; r = -.06	z = - 2.074 p ≤ .05; r = -.50	z = -2.864 p ≤ .05; r = -.69

Table 6.2: The p values and the calculated *z-score* values for the Wilcoxon-Mann-Whitney U test are listed in this table for each pairwise technique comparison and each group of the shape experiment. If $z \notin [-1.65, 1.65]$, the test confirms a significant difference with $p \leqslant .05$. Moreover, the Pearson's correlation coefficient r illustrating the *effect size* is included (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$ effect). Green colored results represent statistically significant differences. Pairwise technique comparisons with no significant difference are colored red.

A pairwise comparison and test for statistically significant differences was performed using the Wilcoxon-Mann-Whitney test. The results, the corresponding *z-score* values and the Pearson's correlation coefficient r to illustrate the effect sizes are listed in Table 6.2. As already mentioned, no significant differences were found for accuracy for $group_{-RO}$. In contrast, participants of $group_{+RO}$ perceived the aneurysms' surface more precisely with G ($p \leqslant .05$, $z = 1.73$, $r = -.41$) and GD ($p \leqslant .05$, $z = 1.82$, $r = -.44$) than with S. These results exhibit a medium correlation between each technique pair. Thus, $H_{Shape_{Acc}}$ can be confirmed for $group_{+RO}$. Moreover, G ($p \leqslant .05$, $z = 2.07$, $r = -.50$) enabled a significantly more precise shape perception compared to GD.

Since the gauge figures were positioned in three different surface regions (opaque, semitransparent, transparent), the participants' accuracy referring to the regions was additionally analyzed for $group_{+RO}$. G enabled a significantly more accurate shape perception within semitransparent regions ($p \leqslant 0.05$, $z = 1.71$, $r = -.41$) and GD within opaque ($p \leqslant .05$, $z = 2.02$, $r = -.48$) regions than S. Furthermore, a significant difference exists between G compared to GD for semitransparent regions ($p \leqslant .05$, $z = 1.81$, $r = -.43$), with the results for G being more accurate. Besides the accuracy results, Table 6.2 includes the required time results. Both groups oriented the gauge figures significantly faster when the models were visualized with S compared to G ($group_{-RO}$ with $p \leqslant .05$, $z = 1.661$, $r = -.40$ and $group_{+RO}$ with $p \leqslant .05$, $z = 2.817$, $r = -.68$. Furthermore, $group_{-RO}$ was signifi-

cantly faster with S compared to GD ($p \leqslant .05,\ z = 1.681,\ r = -.40$). The required time for stimuli with G compared to stimuli with GD was statistically not significantly different. In contrast, $group_{+RO}$ required statistically no significantly different task completion times for S compared to GD ($p = .054,\ z = 1.634,\ r = -.39$) but for GD compared to G ($p \leqslant .05,\ z = 2.864\ r = -.69$).

In summary, as long as the participants could rotate the models, $H_{Shape_{Acc}}$ was confirmed with the exception of transparent regions. In contrast, when the participants could not rotate the model, no significant difference was found and $H_{Shape_{Acc}}$ is highly unlikely. Even though participants performed more accurate with the ghosting techniques, they required significantly more time orienting the gauge fiqures when the models were visualized with G compared to S. Due to that, $H_{Shape_{Time}}$ is highly unlikely. The participants of both groups were faster when the aneurysm models were visualized with S, even though $group_{+RO}$ achieved better results with GD than with G. Overall, $group_{+RO}$ was significantly more accurate but slower than $group_{-RO}$.

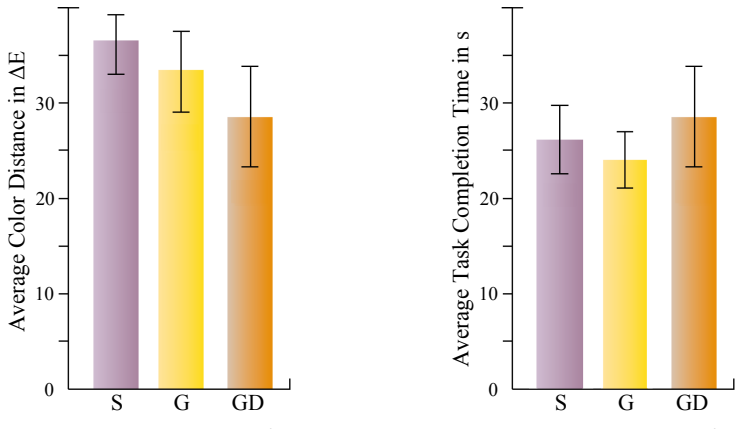
6.5.2 *Smart Visibility Results*

For the smart visibility experiment, the estimated color for each region, the corresponding gs_{color} as well as the task completion time were recorded. As mentioned in Section 6.3.2, gs_{color} was determined based on the results of the stimuli illustrating only the blood flow. gs_{color} described the average perceived color error compared to the histogram results for each region. Since we analyzed the color difference, the participants' results were stored as L*a*b* color values. This space is roughly perceptually uniform and thus, uniform changes of components in this color space roughly correspond to uniform changes in the perceived color. Thus, two colors can be treated as two 3D points and the perceptual color difference is the Euclidean distance (ΔE) between these two points. A difference of 1.0 ΔE means that the color difference between two colors is perceptually distinguishable.

	Accuracy			Task Completion Time		
	$\bar{x}$	σ	Shapiro-Wilk	$\bar{x}$	σ	Shapiro-Wilk
S	36.21	8.97	$p \leq .05$	25.83	15.37	$p \leq .05$
G	33.72	11.98	$p \leq .05$	23.77	13.25	$p \leq .05$
GD	28.83	15.01	$p \leq .05$	23.14	12.64	$p \leq .05$

Table 6.3: This table comprises the mean values ($\bar{x}$), the standard deviations (σ) and the Shapiro-Wilk test results for accuracy in ΔE and for task completion time in seconds. Results with $p \leqslant .05$ indicate a statistically significant difference from a normal distribution.

The average ΔE and time results are listed in Table 6.3. In contrast to the average color distance of 36.21 ΔE for S, the participants estimated the flow color more precise, in terms of smaller distances to their specified gs_{color} when a model was visualized with a ghosting technique (33.72 ΔE for GD and 28.83 ΔE for GD). Especially the average distance of 28.83 ΔE for GD exhibited that this technique enabled the participants to estimate the illustrated flow color more precise. The required task completion times were very close and differ at the most between S



(a) Accuracy (b) Time

Figure 6.8: (a) The average Euclidean distance (ΔE) of the color estimates and the corresponding gs_{color} and thus, the accuracy of S, G and GD and (b) the average task completion time results in seconds illustrated as bar charts including standard error bars. (Images reprinted from Baer et al. [8] © John Wiley & Sons 2011 with kind permission from John Wiley & Sons, Imc.)

with 25.83 s and GD with 23.14 s, with GD enabling a faster color estimation than S. This difference was less than two seconds. Moreover, models visualized with G (23.77 s) enabled the participants to estimate a color almost in the same time as models visualized with GD. These descriptive analysis results are illustrated as frequency distribution histograms in Figure 6.8.

The Friedmann test determined a statistically significant difference with $p \leqslant .01$, $\chi^2 = 20.75$ for accuracy and with $p \leqslant .05$, $\chi^2 = 5.99$ for the task completion time results. The technique comparison revealed that participants performed significantly more accurate and thus estimated the flow color more precisely with GD ($p \leqslant .001$, $z = -3.060$, $r = -.58$) than with S and with G ($p \leqslant .01$, $z = -2.779$, $r = -.53$). No significant difference with $p = .08$, $z = -1.37$, $r = -.26$ exists between S and G. Even though the required times were very similar, a just significant difference exists for S compared to GD ($p \leqslant .05$, $z = -1.654, r = -.31$) with GD accelerating the color estimation. No statistically significant difference

Comparison	Accuracy	Time
S - G	z = - 1.373 p > .05; r = -.26	z = - 1.373 p > .05; r = -.26
S - GD	z = - 3.060 p ≤ .001; r = -.58	z = - 1.654 p = .05; r = -.31
G - GD	z = - 2.779 p ≤ .05; r = - .53	z = - .443 p > .05; r = -.08

Table 6.4: The p values and the corresponding *z-score* values for the Wilcoxon signed-rank test are listed in this table for each pairwise technique comparison. If $z \notin [-1.65, 1.65]$, the test confirms a significant difference with $p \leqslant .05$. Moreover, the Pearson's correlation coefficient r is included (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$ effect). Green colored results represent statistically significant differences, red colored results are not significantly different.

exists between G and GD ($p > .05,\ z = -.443, r = -.08$) and between S and G ($p > .05,\ z = -1.373, r = -.26$), as listed in Table 6.4.

Due to the statistical analysis results for accuracy and task completion time, $H_{Smart_{Acc}}$ and $H_{Smart_{Time}}$ are highly unlikely. We are, however, able to confirm that GD enabled a more accurate and accelerated assessment of embedded flow compared to S.

6.5.3 *Spatial Perception Results*

The knowledge of the correct depth ordering for each stimulus and the recorded answers enabled the analysis of correct and false responses. Additionally, the techniques S, G, and GD were analyzed according to the three different rotation angles. The accuracy results listed in Table 6.5 and the histogram in Figure 6.9a document

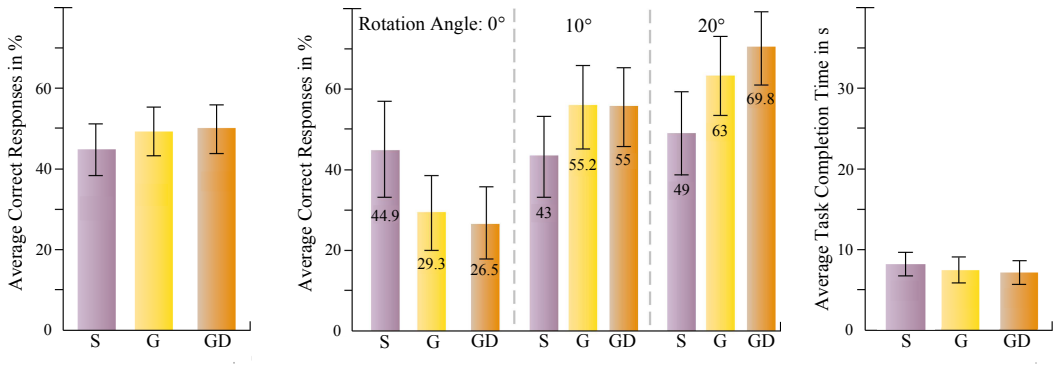
	Accuracy			Task Completion Time		
	$\bar{x}$	σ	Shapiro-Wilk	$\bar{x}$	σ	Shapiro-Wilk
S	45.03	16.39	$p \leq .05$	8.03	4.30	$p \leq .05$
G	49.75	14.53	$p \leq .05$	7.53	4.38	$p \leq .05$
GD	50.47	14.32	$p \leq .05$	7.21	4.30	$p \leq .05$

Table 6.5: This table comprises the mean values ($\bar{x}$), the standard deviations (σ) and the Shapiro-Wilk test results for accuracy in % of correct depth ordering and for task completion time in seconds. Results with $p \leqslant .05$ indicate a statistically significant difference from a normal distribution.

that participants achieved 45.03% correct responses for S, 49.75% for G and 50.47% for GD. The angle-dependent frequency distribution histogram in Figure 6.9b depicts the average correct responses for each technique and each angle. Stimuli with both vessel branches at the same distance (0°) to the participant were correctly perceived more often with S (44.9%) than with G (29.3%) or with GD (26.5%). Contrary, stimuli with branches rotated by 10° and thus with a presented branch depth order were correctly perceived more often with G (55.2%) than with GD (55.0%) or with S (43.0%). Moreover, a rotation by 20° resulted in a more correct depth judgment for stimuli with GD (69.8%) than with G (63.0%) and with S (49.0%).

In contrast to the other experiments, the participants were faster selecting the closest vessel branch than orienting gauges or defining colors. The task completion time results depict that this experiment required less time to answer the task than the other two experiments and that the required times were very close for all techniques (see Figure 6.9c). The longest times were required for S with 8.03 s and the shortest with 7.21 s for GD. However, these were the fastest time results.

The Friedmann test confirmed a statistically significant correlation between the techniques and the number of correct responses with $p \leqslant .05$ and $\chi^2 = 6.126$. No significant difference was found ($p > .05, \chi^2 = 3.920$) between any two techniques for the task completion time results. Significantly more correct answers were recorded for G ($p \leqslant .05,\ z = -1.661, r = -.33$) and for GD ($p \leqslant .05,\ z = -2.163, r = -.43$) compared to S. Both ghosting techniques exhibited statistically no significant difference when analyzing the accuracy based on the correct depth judgment answers over all angles. However, $H_{Spatial_{Acc}}$ is highly likely. The angle-



(a) Accuracy (b) Angle-Dependent Accuracy (c) Time

Figure 6.9: (a) The average number of correct responses in % for all stimuli indicates the participants' accuracy for the depth judgment task. (b) The histogram illustrates the accuracy based on the three different ration angles. (c) The required task completion times illustrated as bar charts. (Images reprinted from Baer et al. [8] © John Wiley & Sons 2011 with kind permission from John Wiley & Sons, Inc.)

dependent frequency distribution histogram in Figure 6.9b illustrates larger accuracy differences between the techniques when analyzing the correct depth judgment for each of the three angles. Thus, we analyzed the results individually for 0°, 10°, and 20°. The results for 0° were completely different than for 10° and 20°. When no depth ordering exists, the participants achieved more correct answers when the models were visualized with S compared to G ($p \leqslant .05$, $z = -1.857, r = -.37$) and compared to GD ($p \leqslant .002$, $z = -2.764, r = -.55$), see Table 6.6. If a depth ordering of the vessel branches exists, the ghosting techniques' assist a correct depth judgment changes. The ghosting techniques enabled more correct answers while the benefit of the semitransparent visualization technique decreased. G enabled statistically significantly more correct answers for 10° with $p \leqslant .05$, $z = -1.656, r = -.31$ and for 20° with $p \leqslant .01$, $z = -2.594, r = -.51$ compared to S. The participants, moreover, performed significantly more correct with GD compared to S for 10° with $p \leqslant .01$, $z = -2.546, r = -.50$ and for 20° with $p \leqslant .001$, $z = -3.346, r = -.66$. The larger the rotation angle, the higher the

Comparison	All Angles	0°	10°	20°
S - G	z = - 1.661 p ≤ .05; r = -.33	z = - 1.857 p = ≤ .05; r = -.37	z = - 1.654 p = ≤ .05; r = -.31	z = - 2.594 p = ≤ .025; r = -.51
S - GD	z = - 2.163 p ≤ .025; r = -.43	z = - 2.764 p = ≤ .025; r = -.54	z = - 2.546 p = ≤ .025; r = -.50	z = - 3.346 p = ≤ .001; r = -.66
G - GD	z = - .141 p > .05; r = -.02	z = - .665 p > .05; r = -.13	z = - .535 p > .05; r = -.10	z = - 1.414 p > .05; r = -.28

Table 6.6: The p and *z-score* values for the accuracy results of the spatial experiment are listed in this table. The results based on the correct answers are divided into results for all angles, for 0°, 10°, and 20°. If $z \notin [-1.65, 1.65]$, the test confirms a significant difference with $p \leqslant .05$. Moreover, the Pearson's correlation coefficient r is included (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$ effect). Green colored results represent statistically significant differences, red colored results are not significantly different.

significant difference and the larger the difference of correct responses between G (63%) and GD (69.8%). However, since the Wilcoxon signed-rank test was applied twice to the spatial results – all angles and angle-dependent –, a Bonferroni correction had to be considered. That means, a difference is statistically significant with $p \leqslant .025$. Thus, S compared to GD showed statistically significant differences for both analysis steps (all angles and angle-dependent). Additionally, the participants were statistically more correct with G compared to S for 20°. Even though no significant difference was confirmed between G and GD, $H_{Spatial_{Acc}}$ is highly likely for 20°. The required times to complete the task showed no significant effects for this experiment. $H_{Spatial_{Time}}$ is highly unlikely, since G and GD did not accelerate the spatial perception.

6.6 QUALITATIVE RESULTS

The questionnaire analysis measured the participants' attitude towards a technique. The results of each pairwise technique comparison for each experiment (colored bars) are illustrated in Figure 6.10. We had a total of 86 participants. As

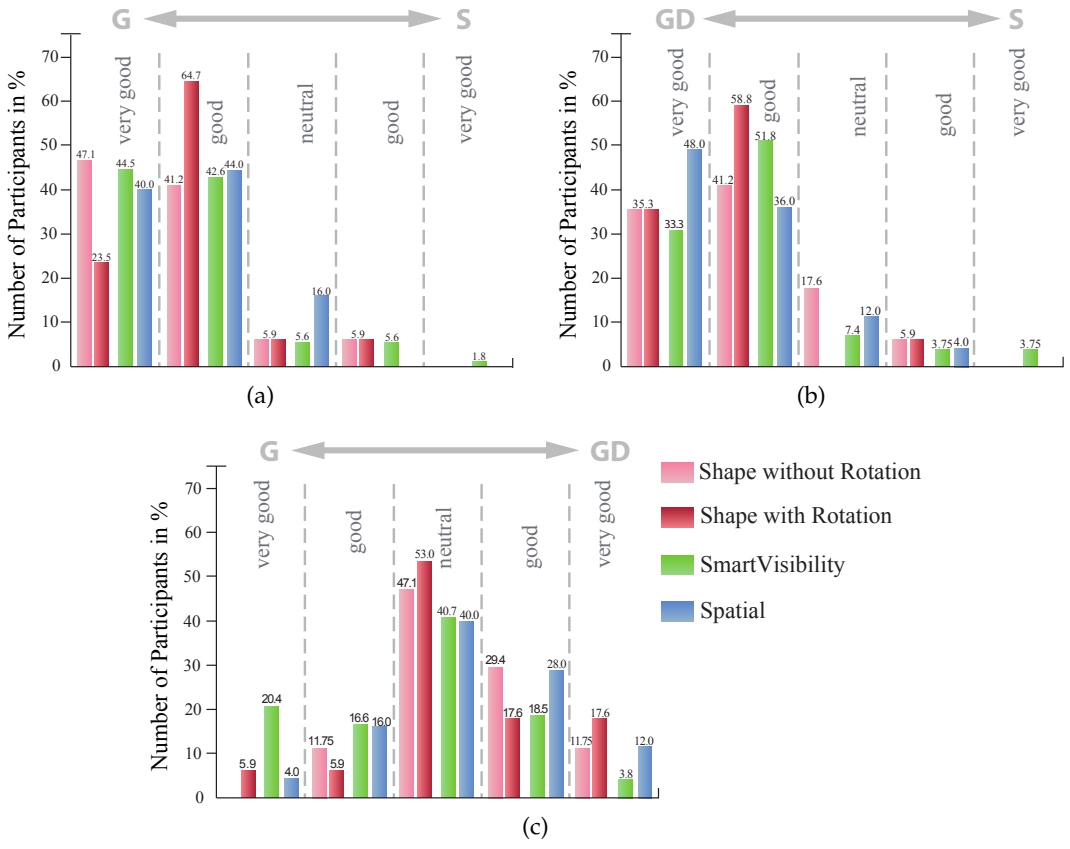


Figure 6.10: These are the results of the qualitative technique comparison for each experiment. A 5-point Likert scale was used to qualitatively evaluate the techniques' support and to compare (a) the ghosting technique G with the common semi-transparency S, (b) the ghosting with depth enhancement GD with semitransparency S, and (c) both ghosting techniques. (Images reprinted from Baer et al. [8] © John Wiley & Sons 2011 with kind permission from John Wiley & Sons, Inc.)

illustrated in Figure 6.10a, 34 participants (39.5%) rated G compared to S as *very good* and 41 (47.7%) as *good*. Thus, 75 participants (87.2%) preferred G over S. The comparison of GD and S (see Figure 6.10b) showed that 73 participants (84.9%) rated GD as *very good* or *good*, and thus, preferred this technique over S, too. Finally, Figure 6.10c includes the results of the ghosting technique comparison. 19 participants (22.1%) rated G as either *very good* or *good*. In contrast, 29 participants (33.7%) preferred GD and chose *very good* or *good*, and 38 participants (44.2%) liked both kinds of visualization techniques and chose *neutral*. Even the experiments individual results showed an overwhelming preference for both ghosting techniques. Although the three experiment tasks tackled different technique characteristics and asked for different assessment tasks the qualitative results were very similar. The majority of all participants preferred G and GD over S and a small tendency to GD was visible.

6.7 RESULT DISCUSSION

As presented in the previous section, the analysis showed that two of the three accuracy hypotheses are highly likely with medium and large effect sizes. The postulated hypotheses for the task completion times, however, are all highly unlikely.

First, for the shape experiment we were able to confirm $H_{Shape_{Acc}}$ for $group_{+RO}$ but had to reject $H_{Shape_{Time}}$ for $group_{+RO}$ and $group_{-RO}$. G and GD enabled a more accurate shape perception as long as the participants had the opportunity to rotate the models, since $group_{-RO}$ showed no statistically significant difference. The achieved results for $group_{-RO}$ were to be expected and showed the importance of an appropriate experimental design. Since G and GD were developed for interactive 3D visualizations, they had to be evaluated in 3D as well. The estimated surface normals of $group_{+RO}$ had smaller angular errors than of $group_{-RO}$, even though they required more time. Even the specific surface regions exhibited more accurate perception using G and GD for opaque and semitransparent regions. As a standard speed-accuracy trade-off, the participants required more time for both ghosting visualizations, since they perceived the surface shape better, and thus aimed at orienting the gauge figures as close as possible to the gs_{normal}. Since the participants were asked to comment on the experiment at the end, they reported on this situation and confirmed our suspicion. Moreover, they talked during the experiment and tried to be even more accurate with orienting the gauge figures if the models were visualized with G or GD. If the participants had the feeling that they were not able to orient the gauge figure according to a surface normal, since they could not perceive the real surface, they were less motivated to try and wanted to move on to the next stimulus. This happened especially for stimuli illustrating the models with S. However, this is only based on informal and additional experiment observations without being based on a real structured observation and recording. Thus, this needs to be done in further properly designed observations or within a further shape experiment using specific tasks that tackle this issue. Ghosting techniques, however, do not accelerate but enable a more precise shape perception.

The second experiment evaluated the smart visibility characteristic of G and GD and surprisingly resulted in the rejection of $H_{Smart_{Acc}}$ and $H_{Smart_{Time}}$. So far, we are able to confirm that GD enables a more accurate assessment of embedded flow than S and G, and GD accelerates the assessment of embedded flow compared to S. The majority of the participants, however, preferred both ghosting techniques over the S technique for this experiment, as illustrated in Figure 6.10. One reason for the accuracy results and statistically no significantly difference might be that the participants defined the flow color for S , which they expected to be the right one and not the color they really saw. When the participants were asked to explain their adjusted colors at the end of the experiment and some of them unintentionally commented their choices during the experiment, they confirmed this assumption. This happened especially after they saw the first stimulus with an aneurysm visualized with a ghosting technique. This might be eliminated by designing a *between-participant* experiment for the smart visibility characteristic, too. Furthermore, the color dialog and task specification turned out to be another experimental design problem that needs to be addressed. Participants saw a bundle of colored streamlines and their task was to estimate the average displayed color. On the one hand, the estimation of an average color of a colored streamline bundle is already difficult on its own. On the other hand, however, the color dialog that showed a fully-colored rectangular area displaying one color – the estimated average color – required a high cognitive load. Both the task and the answer possibility require a re-design. The first and most simplest possibility is to re-design the rectangular color area of the color dialog showing colored lines and thus being more similar to the colored streamlines. This answer possibility, however, does not simplify the task of averaging the different streamline colors that were displayed in one region and that had to be mentally averaged. A new task to evaluate the smart visibility characteristic would be more appropriate. However, a re-design and further evaluations are required.

The third experiment evaluated the techniques' effectiveness supporting the depth perception. Our results confirmed that $H_{Spatial_{Acc}}$ is highly likely with medium and large effect sizes. Even though G achieved faster responses than S and this was close to a significant difference, we had to reject $H_{Spatial_{Time}}$. All participants were, however, faster with G and GD than with S. The accuracy results required a detailed analysis. Stimuli with both vessel branches at the same distance (0°) were more often correctly assessed if they were visualized with S than with one of the ghosting techniques. This may reflect a bias towards a response of "no separation in depth" for S. In such a case, lower performance for the two larger rotations is to be expected, which is precisely what we found. Stimuli showing a rotation angle of 10° and 20° were more often correctly perceived with G and GD than with S. In summary, G and GD facilitate depth judgment compared to S. Similar to the previous experiment, a re-design is recommended for further experiments, too. Especially the vessel branch endings that were visible for 10° and 20° have to be removed, since this influences and facilitates the judgment, recall Figure 6.5b and 6.5c. Viewpoints that cut the branch shortly before the model's end will avoid this unwanted depth cue.

LIMITATIONS. A few experiment-specific drawbacks were already mentioned. A main limitation is the determination of gs_{normal} and gs_{color}. We used the fully opaque shaded model and the streamline visualization to define a perceptual gold standard. However, we did not test, if the surface was influenced by the applied surface color. This was done to facilitate the recognition of a blood vessel, since it was not possible to test only medical experts or at least to have participants with passing knowledge. Moreover, the streamline visualization used for the determination of gs_color was not evaluated, too. Furthermore, the number of participants varied between the experiments. The shape perception results were gathered from 17 participants each, while 27 and 25 participants performed the smart visibility and the spatial perception experiment. Different numbers of participants hamper an overall statement of the techniques' effectiveness for all three requirements mentioned in Section 6.3.

Overall, the results showed that G and GD enable more accurate assessments than S for each experiment. Furthermore, G and GD accelerated the individual tasks, except for the shape perception experiment. In this experiment, more accurate shape perception was maintained at the expense of task completion time. This general result was confirmed by the qualitative results. We found overwhelming preferences for the two ghosting techniques over S. There was also a small trend towards a preference of GD over the simple G.

6.8 LESSONS LEARNED

Smart visibility techniques combine several individual technique features to enable a facilitated structure or visualization exploration. Evaluations of complex illustration techniques or visualizations benefit from a detailed analysis. Each individual feature and innovation may contribute to or hamper the effectiveness. A detailed analysis and comparison gives insights into potential advantages and drawbacks and outlines required improvements. When comparing illustration techniques without such a feature-dependent categorization, potential drawbacks may be undiscovered or the technique fails in the worst case even if some features would improve a structure perception. An evaluation design that does not consider these individual features separately, can not result in a general conclusion.

However, a focus to the major features is recommended to control the extent and completion time of the evaluation. A good strategy to define the evaluation steps and categories is a detailed analysis of the comparative motivation, in advance. The typical motivation of comparative evaluations is the assumption that one technique, visualization or device is *better* than the other. The required evaluation steps can be derived from the definition of "*better*" and the following questions:

- What means one technique is better than the other?
- What is the precise meaning of better and how is it defined?

Answering these questions facilitates the evaluation design and required tasks. The presented ghosting techniques were developed to support the vessel's shape perception and the exploration of the embedded flow. Additionally, a depth enhancement was included to improve the depth perception. Based on these features

and improvements compared to the common semitransparency visualization technique, three evaluations were designed.

Interaction is integrated for the shape perception evaluation, since the illustration technique and the surface perception, respectively benefits from the viewers interaction with the 3D scene. In all other evaluations the rotation option was disabled to minimize the interaction effort, and thus to preserve the participants' concentration and focus to the main tasks. Moreover, since the visibility of embedded structures as well as the depth enhancement can be evaluated using static stimuli, there was no need to enable rotation as an interaction option. The technique enables a view-dependent exploration of the embedded blood flow. Thus, the static 3D visualizations provide an optimal view.

6.9 SUMMARY

This chapter presented three controlled task-based experiments investigating the visualization of the cerebral aneurysm anatomy with embedded flow visualization derived from five clinical datasets. Quantitative and qualitative evaluations were performed to evaluate and compare the common semitransparent visualization technique with a ghosted view and a ghosted view with depth enhancement technique developed by Gasteiger et al. [57]. In detail, three studies analyze the techniques' capability to facilitate and promote the shape and spatial representation of the aneurysm models as well as evaluating the smart visibility characteristics. The techniques were quantitatively evaluated with respect to the participants' accuracy and required time to complete the task, and qualitatively in terms of their personal preferences for each experiment. The experimental design process including the task methodologies and stimuli design was presented and explained together with the individual experimental procedure and the analysis methods. Results were presented and discussed to gain further insights into the confirmation and rejection of the postulated hypotheses and to highlight the potential for further improved experiments. Overall, there was an overwhelming preference for the two ghosted techniques over the semitransparent technique. The quantitative analysis determined the advantage of both ghosting techniques and clearly showed that both techniques support a more accurate analysis of aneurysms than the traditional S technique. Based on the time results, the generalization of a shape perception acceleration by a ghosting visualization turned out to be invalid. The shape experiment, however, outlined that this is a speed-accuracy trade-off and leads to a more correct shape assessment, which is more relevant for the clinical routine.

The above-mentioned limitations and drawbacks do not depreciate these evaluations. An experiment analysis based on the gained knowledge and observations during the experimental sessions is provided and essential. On the one hand, without such experiments the visualization techniques are not perceptually evaluated and verified. On the other hand, the experimental design, procedure and analysis itself can be evaluated to identify advantages and disadvantages and point out improvement potential. Since smart visibility techniques are rarely evaluated, this may serve as orientation for further studies and enable first insights into the

individual potential. The three techniques were evaluated in the context of the visualization of cerebral aneurysms, but the experiments and outcomes may also be applied to other vessel flow domains, such as aortic or cardiac flow. Since the shape evaluation is not limited to a vessel-shaped structure, it is also applicable to other surface renderings where additional internal information have to be visualized, e.g., the heart or organs with internal tumor visualizations such as liver or lung tumors. Contrary, the spatial perception evaluation is designed for elongated or occluding structures such as vessels or vessel trees, due to the fact that two well-defined structures are required to determine their spatial relation. This evaluation design resembles other depth judgment studies with vessel trees where individual features such as branches have to be spatially separated in 3D [137, 84] or even in 2D angiography images [138]. However, the results of our experiments should not be overgeneralized, since it is unclear how the results can be generalized to other, complex anatomic or botanic shapes. This needs to be analyzed.

7

COMPARISON OF A 2D VERSUS TWO 3D DISPLAYS FOR A MEDICAL IMPLANT PLACEMENT TASK

This chapter is based on the following publication:

"A Comparative User Study of a 2D and an Autostereoscopic 3D Display for a Tympanoplastic Surgery". Alexandra Baer, Antje Hübler, Patrick Saalfeld, Douglas Cunningham and Bernhard Preim. In *Proceedings of Eurographics Workshop on Visual Computing for Biology and Medicine (EG VCBM)*, pages 181-190, 2014

In order for surgeons to navigate through a patient's anatomy, they usually must rely on very accurate depth judgment and spatial orientation abilities. The correct localization of the surgical instruments as well as the identification of relevant anatomical and pathological structures usually requires high perceptual skills. These skills are essential for performing fine dissections, to avoid injuring risk structures, or for the correct position and alignment of implants. In particular the microscopes or endoscopes used for minimal interventions provide a very restricted field of view, which contains limited (stereoscopic) depth cues and spatial information. Thus, especially surgeons-to-be need a lot of training and trial-and-error experience to gain adequate surgical skills and experiences, e.g., depth judgment and spatial assessment of structure relationships. Virtual training improves the accuracy, the time taken to perform a task, and minimizes errors compared to no training [64]. An extensive training with different pathologies in advance improves the surgical skills. To design a virtual training scenario, an appropriate 3D training environment is required.

As mentioned in Section 3.3.3, stereoscopic cues and motion parallax are the most significant sources of depth information, beyond the depth cues of visualization techniques [166]. In- and output devices like stereoscopic displays, haptic devices or 3D navigators and stylus input devices are developed to enable intuitive 3D visualizations, navigations and interactions with 3D scenes. As discussed in Section 3.3.3, several studies document the advantages of 3D displays including shutter and polarized glasses or glass-free autostereoscopic systems. When using a 3D display, a simple rendering of structures is sufficient to explore a patient's anatomy. However, 3D displays are still not accepted in surgery, even though their technology enables binocular vision without specific visualization techniques and thus should support an intuitive depth perception. In the past, 3D displays suffered from negative side effects of stereopsis, such as the perception of double images or physical effects like nausea and headache [160, 133]. The latest

3D imaging systems provide improved image quality and resolution, similar to 2D monitors. Therefore, 3D displays represent a potential alternative for surgery training systems compared to manifold developed visualization techniques. Moreover, stereoscopic depth cues provided by the used microscopes can be integrated in a training scenario and thus used to improve the surgical skills. However, visualization techniques as well as 2D and 3D in- and output devices have to be investigated to provide an optimal 3D environment with appropriate depth cues.

Baer et al. [9] focused on the design and execution of a comparative experimental study investigating the effectiveness of 2D versus 3D displays for an otologic training scenario. We compared a 2D display with a glasses-free 3D autostereoscopic display in detail and investigated in a follow-up pilot study a 3D zSpace system[1] using a stylus as input device. The experimental study is based on a developed training scenario for a tympanoplastic surgery. Tympanoplastic is a general term used to describe a surgical procedure designed to remove a pathology and repair defects in the middle ear (ossicular chain) and the eardrum.

7.1 MEDICAL BACKGROUND

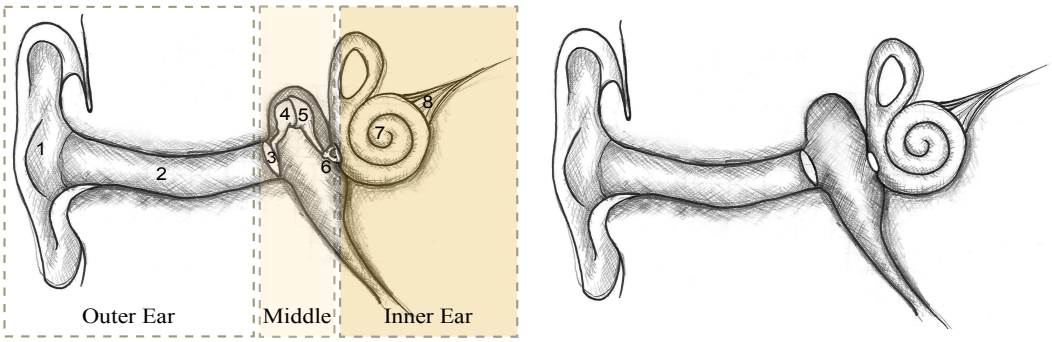
Three parts of the human ear are responsible for converting sound waves into electrical impulses: the outer, the middle, and the inner ear. Each of these has tiny, complex structures that detect vibrations, transmit mechanical energy, or convert mechanical energy into electrical nerve impulses (see Figure 7.1a). If one part or structure is malformed, damaged, lost or not fully functioning, the patient's ability to hear is either impaired or totally lost. The partial or total inability to hear is called deafness and is caused by sensorineural or conductive hearing loss.

SENSORINEURAL HEARING LOSS known as nerve-related hearing loss occurs when there is a damage to the inner ear (cochlear) or to the nerve pathways from the inner ear to the brain. With this type of hearing loss it is not always possible to tell which part is damaged and it is therefore often summarized as sensorineural hearing loss. The causes are varied but can be generally put into two categories: congenital (genetic or hereditary) and acquired (illness, trauma, noise exposure, or age-related).

CONDUCTIVE HEARING LOSS is caused by problems with the ear canal, eardrum, middle ear, or abnormalities in mobile portions of the ear. These movable parts transmit sound from the outside to the inner ear, where our nervous system takes over and transmits signals to the brain. Conductive hearing loss occurs when these movable parts are damaged, lost (see Figure 7.1b) or when their mobility is impaired, e.g., caused by genetics, physical trauma or ear malformations.

MIXED HEARING LOSS refers to a combination of conductive and sensorineural hearing loss. This occurs when there is a damage in the outer or the middle ear and in the inner ear (cochlea) or the auditory nerve.

1 www.zspace.com



(a) Anatomy (b) Conductive Hearing Loss

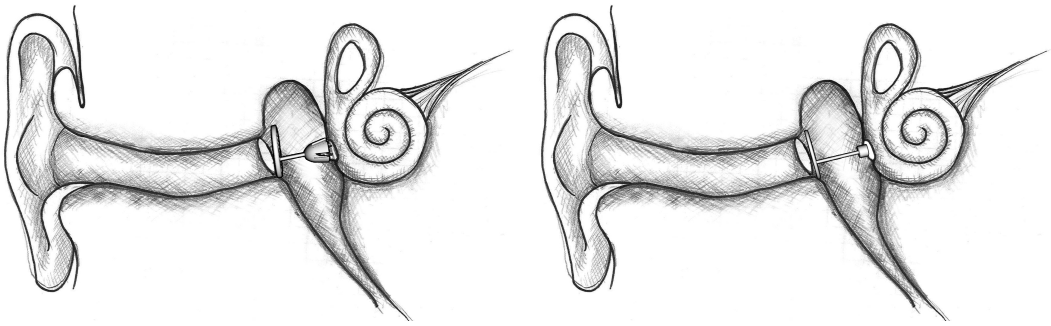
Figure 7.1: (a) The three parts of the human ear. The (1) pinna and (2) auditory canal are the outer ear. (3) Eardrum, (4) mallus, (5) incus, and (6) stapes bone are the middle ear and the (7) cochlear and (8) auditory nerve belong to the inner ear. (b) Conductive hearing loss occurs when the ossicular chain (the movable parts (3)-(6)) is damaged, lost or when its mobility is impaired. (Images reprinted, with permission, from Baer et al. [9] © Eurographics Association 2014.)

7.1.1 *Treatment of Deafness*

Nowadays, hearing can be medically or surgically corrected. A surgical procedure comprises the implantation of prostheses to restore the hearing ability. Cochlear implants are used for sensorineural and ossicular prostheses for conductive hearing loss.

A cochlear implant consists of an internal and external component. The internal component is surgically inserted under the skin behind the ear, and a narrow wire is threaded into the inner ear. The external component is connected to the internal one and sound waves are converted to electrical impulses bypassing the defective inner ear and providing patients with the ability to hear [118]. A tympanoplastic surgery re-establishes the ossicular chain and the non-functioning ossicles may be reshaped to fit properly or be replaced with a prosthetic (artificial) implant. This is a mobilization surgery to restore the ossicular chain [12]. As sketched in Figure 7.2a, gaps between the intact stapes and either the incus, malleus handle or eardrum are bridged with a *partial ossicular prosthesis* (PORP) [60]. If there is no stapes superstructure and the prosthesis connects the stapes footplate to the other ossicles or eardrum, it is called a *total ossicular prosthesis* (TORP), see Figure 7.2b.

However, successful reconstruction and implanting presents significant challenges. Besides a working knowledge of available materials, prosthesis design, reconstruction techniques and their adaptability to a variety of problems, profound surgical skills in otologic and micro-surgical techniques are essential [29]. Reconstruction results are variable, with some procedures giving complete normal hearing and some giving no improvement in hearing at all [60]. Management of the conductive hearing loss associated with ossicular damage will depend on the exact pathology and the integrity of the tympanic membrane. The results achieved from the implantation of a PORP or TORP are dependent on [150]:



(a) PORP (b) TORP

Figure 7.2: (a) Gaps between intact stapes and either the incus, malleus handle or eardrum are bridged with a partial ossicular prosthesis (PORP). (b) Total ossicular prostheses (TORP) implants are used to bridge the gap between the eardrum and the stapes footplate to restore the hearing ability. (Image reprinted, with permission, from Baer et al. [9] © Eurographics Association 2014.)

1. the pathology present in the area to be implanted and the surgical skills employed in the implantation,
2. the implant's biomechanical properties, and
3. the biocompatibility of the material implanted.

Any procedure needs to be carried out with great care, as with excessive movements structures may be damaged or energy may be transferred to the cochlear causing sensorineural hearing loss. Otologic surgeons-to-be need an extensive and long training and trial-and-error experience in hearing restoration, including the correct prosthesis length judgment and trimming during surgery as well as the correct prosthesis positioning.

7.1.2 *Tympanoplastic Surgery Workflow*

Since we aimed at a training scenario for a tympanoplastic surgery, a workflow analysis was performed to design an appropriate training scenario and experimental evaluation. Tympanoplastic surgery is a minimally invasive intervention performed through the ear canal or an incision in or behind the ear. A microsurgical technique is used for a tympanoplastic procedure to enlarge the view of the ear structures, giving a more detailed image to the ear surgeon. In clinical routine, stereoscopic microscopes are used to support the spatial orientation and to improve the navigation.
In tympanoplastic surgery, the surgeon tries to position a prosthesis implant through the ear canal using surgical instruments while looking through a microscope. The remaining eardrum is elevated away from the bony ear canal and lifted forward. A prosthetic implant is placed into the middle ear underneath the remaining eardrum to bridge the gap between the eardrum and the damaged or missing bone structures (recall Figure 7.2b). The major challenge is the depth judgment for a correct position and the length estimation of the prosthesis. During surgery, the cutting nib with known extension is used to estimate the distance between eardrum and footplate. It may require placing the prosthesis in the ear several

times to estimate the correct length. The correct length is such that the prosthesis only touches the undersurface of the eardrum without tenting it and bridging the existing gap, as illustrated in Figure 7.2.

Besides very fine motor skills (hand-eye and hand-hand coordination), the surgeon has to perceptually combine the images seen through the microscope with his actions. Additionally, a patient-specific depth judgment of the structure relationships is required, since vital anatomical landmarks may be obscured by disease or exposed to serious injury.

7.2 EXPERIMENTAL DESIGN

Initially, we analyzed the tympanoplastic surgery workflow by observations and interviews to specify the research goal and identify the essential parts of a controlled experimental evaluation, e.g., design, task and analysis methods, explained in Section 2. As described in the previous section, we identified the essential tasks that require high perceptual skills of an experienced surgeon and thus are essential for a training scenario. We focused on a TORP implant surgery with no incus, malleus or stapes bones, but footplate existing (see Figure 7.3a). Thus, participants were asked to select the correct TORP length and to position it correctly. We quantitatively compared the displays by means of the participants' task performance that was defined by accuracy, required number of interactions and task completion time:

1. **Accuracy** was measured to evaluate the depth perception. Since the prosthesis implantation is a complex task, we defined accuracy based on the correct *TORP length judgment* and TORP position defined by the *TORP placement* (xyz position) and the *TORP orientation*. A correct position was achieved when the TORP touches the eardrum without penetrating it and bridging the gap between eardrum and stapes footplate.

2. **Interaction** was measured to analyze the difference of required interactions. We assumed that almost no scene interaction is required when the training scenario is performed using a 3D display. Since the 2D display did not support the depth perception, we assumed that the participants would interact (limited rotations) with the middle ear model to enable a spatial orientation. This measured parameter defines the virtual implanting workload. We split the interaction into the number of scene and the number of required TORP interactions.

3. **Time** was measured to analyze if an acceleration exists. This parameter was defined as the participants' time to complete the implanting task.

Additionally, we asked for personal preferences to qualitatively evaluate each display. We did not expect that one display is better or worse than the other in every single investigated aspect, but we expected a difference. No hypotheses were postulated for the follow-up study, since this study investigated a combination of in- and output device and was therefore analyzed exploratory. The number of participants deviated from the main experimental evaluations and, thus, a statistical

comparison was not appropriate anyway. However, we assumed that the participants would perform more accurate with the zSpace and would prefer the zSpace system. Additionally, the TORP length estimation was measured and analyzed exploratory as well. Since we did not integrate a cutting nib used in surgical procedures to estimate the length, the length estimation task for this training scenario requires further improvements and, thus, a statistical analysis is not suitable.

We defined one- and two-tailed hypotheses. Furthermore, we divided the hypotheses with regard to the single tasks and measured variables. This enabled a more precise analysis of the displays.

ONE-TAILED HYPOTHESES. We postulated the following one-tailed hypotheses for accuracy, divided into object placement and orientation.

- $H_{accTransl}$: TORPs are placed more accurately – as measured by *smaller translation deviations* – using a 3D autostereoscopic display compared to a 2D display.
- H_{accRot}: TORPs are aligned more accurately – as measured by *smaller angular deviations* – using a 3D autostereoscopic display compared to a 2D display.

TWO-TAILED HYPOTHESES. We hypothesized that there is a difference between the 3D autostereoscopic and the 2D display in the number of required interactions and in the mean time to respond and to complete the task:

- H_{Time}: There is a difference between the 3D autostereoscopic display and the 2D display in *the mean task completion time*.
- $H_{actionScene}$: There is a difference between the 3D autostereoscopic display and the 2D display in *the number of scene interactions*.
- $H_{actionTORP}$: There is a difference between the 3D autostereoscopic display and the 2D display in *the number of TORP interactions*.

The evaluation followed a *between-participant* design. The *between factor*: display with the 2D and the 3D autostereoscopic display as the two *levels* of this factor. Since we designed stimuli with three different degrees of difficulty, which were presented to all participants, difficulty was the *within factor* with three *levels*: simple, moderate and difficult. Display and the degree of difficulty were the *independent variables*, and the measured accuracy, interaction and time were *the dependent variables*.

7.2.1 *Participants*

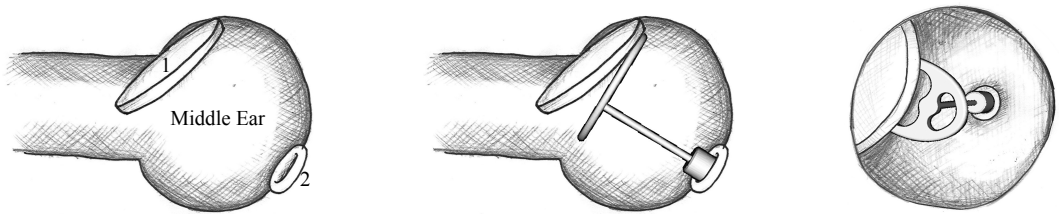
Our participants were recruited from various parts of the university, including medical experts. Although it is recommended to recruit prospective users, in our case ENT (ear, nose and throat) surgeons, participants from the general population can also provide useful insights, see Section 2.2.3. Moreover, it is likely that the measured results can be applied to prospective users concerning the perceptual effectiveness, even though ENT surgeons may achieve better accuracy because of

their clinical experience. Ten participants aged between 29 and 38 years and with $\bar{x} = 34.2$ years (seven male and three female) were recruited for the pilot study. 42 participants, 21 using the 2D and 21 using the 3D autostereoscopic display participated. The 2D group comprised 13 female and eight male participants aged between 18 and 35 years with $\bar{x} = 24.5$ years. The 3D autostereoscopic group comprised nine female and twelve male participants aged between 20 and 46 with a mean age of $\bar{x} = 27.4$ years. In summary, we took care of two similar groups with respect to the age and gender ratio for the comparison of the 3D autostereoscopic and the 2D display.

The follow-up evaluation was performed by twelve participants. Five women and seven men aged between 27 and 53 years with a mean age of $\bar{x} = 34$ years participated in the zSpace system.

7.2.2 *Stimuli*

We designed a stimulus setup for a virtual training of a tympanoplastic surgery focusing on a TORP implanting procedure when no incus, malleus or stapes bones exist (see Figure 7.3a). This surgery was chosen based on the minimum number of existing structures (middle ear, eardrum and stapes footplate) and on the structures' simplicity, since the footplate and the eardrum are easily recognizable. By focusing on one prosthesis type, we minimized bias factors, e.g., different prosthesis types, different structures and individual anatomy. Bias factors are every variation from one to another stimulus that may influence the results (stimuli similarity, recall Section 2.2.4).



(a) (b) (c)

Figure 7.3: (a) The conceptual setup of a conductive hearing loss disease. The middle ear with (1) the eardrum and with no stapes bone, but (2) the footplate. (b) TORP implants are used to bridge the gap between the eardrum and the footplate and to restore the hearing ability. (c) The viewpoint was designed similar to the surgeon's view through the microscope. (Images reprinted, with permission, from Baer et al. [9] © Eurographics Association 2014.)

Figure 7.3a illustrates the stimulus concept and design for a conductive hearing loss disease and Figure 7.3b for the hearing restoration with a TORP implant. The stimuli viewpoint was chosen similar to the surgeon's view through the microscope, as sketched in Figure 7.3, since the tympanoplastic surgery is a microsurgical procedure. The same initial point of view, restricted field of view and scope of action were used. Thus, each stimulus is a 3D model of a middle ear scene viewed through the auditory canal with the eardrum near to the middle

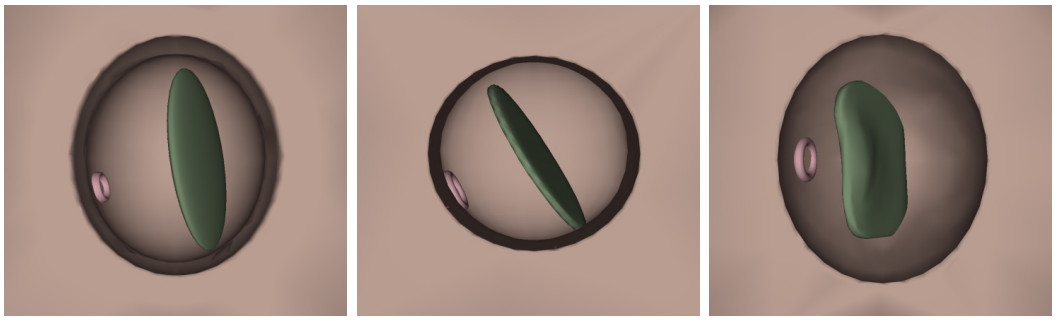
ear cave entry and the footplate at the back (see Fig. 7.3c). Compared to the real anatomical situation visualized in Figure 7.1, the stimuli were slightly modified:

- **The middle ear anatomy** was designed as a closed cave, as shown in Figure 7.3. This restricted the field of view simplified the geometric representation of the middle ear.
- **The eardrum** was represented by a disk-shaped ellipsoid. In human anatomy, it is a thin, cone-shaped membrane. We did not model an eardrum slightly folded to the side, which is the real surgical situation. Instead, a geometrical structure representing the eardrum was modeled and positioned near the middle ear entry. The eardrum is located almost in the center of the entry and slightly rotated around its longitudinal axis (compare Figure 7.4). The eardrum's geometrical shape and position were derived from surgical observations, interviews with two medical experts, and from the four segmented eardrums. However, a segmentation is difficult, since an eardrum is a circle of thin skin of about eight to nine millimeters and therefore hardly distinguishable from other structures within the clinical data.
- **The stapes footplate** was designed as a torus with an average diameter of 2 mm. In human anatomy, the footplate is the base of the stapes bone. This base is the flat portion that fits in the oval window between the middle and the inner ear.

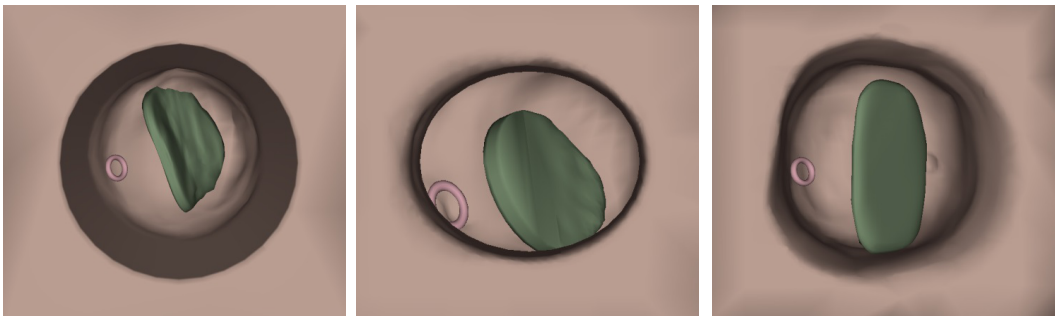
These modifications were necessary to achieve more stimuli similarity and simplicity, described as stimuli requirements in Section 2.2.4. The experimental task, therefore, was facilitated and easier to understand for the non-expert study participants. No sophisticated illustration or illumination technique was applied so that we can focus on the effect of the display technology. Only color, ambient, and diffuse illumination was used to distinguish and identify the relevant structures. We took care of smooth shadows and tried to prevent dominant shadow borders by adjusting the diffuse and ambient amount of illumination.

We had four patient-specific petrous bone CT datasets provided from our collaboration partner at the university medical center. All patients had a conductive hearing loss. Two of the four datasets contained $230 \times 230 \times 208$ voxels and a voxel length of 0.2 mm, and two contained $230 \times 230 \times 123$ voxels and a voxel length of 0.353 mm. The middle ear and the eardrum were segmented using thresholding methods and were based on the segmentation results of the surface morphology. The footplate was manually generated and integrated in the middle ear. A wall behind the footplate was integrated and closed the middle ear in the back to get a cave-like middle ear as previously described. To provide a variety of stimuli, 10 further stimuli were generated using the software tool BLENDER.[2] These generated stimuli were based on the characteristics of real datasets, surgery inspections and video recording and were generated in cooperation with our medical experts. Compared to the patient-specific stimuli, the geometry of the manually generated middle ear stimuli was smoother and more uniform.

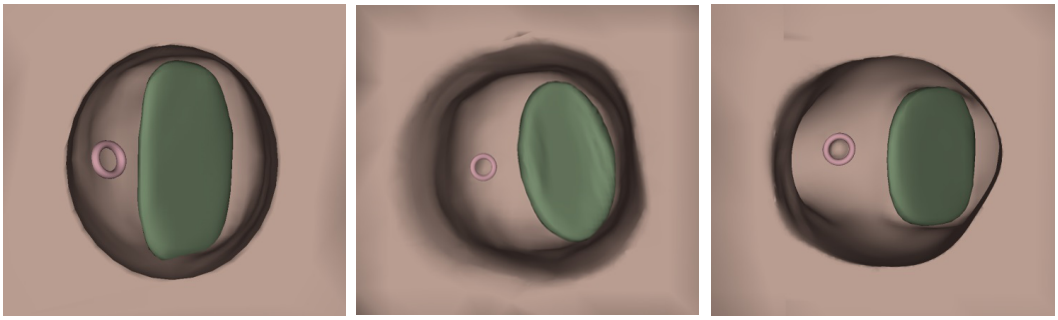
2 www.blender.org



(a) Simple



(b) Moderate



(c) Difficult

Figure 7.4: (a) Simple, (b) moderately difficult, and (c) difficult stimuli were used during the study. Increasing occlusion caused by the eardrum and non-parallel orientations of eardrum and footplate lead to a larger TORP positioning effort.

The virtual 3D prosthesis implant models were provided by KURZ GmbH Medizintechnik[3] and correspond to the titanium implants used in the clinical workflow. They manufactured the TORP implants used by our medical experts for tympanoplastic surgery. We scaled those models along their longitudinal axis to get different TORP lengths with a step size of 0.25 cm. TORPs with a length between 1.75 cm and 6.0 cm were used. The TORP translation and rotation was performed with the mouse for the 2D and autostereoscopic display. To facilitate the TORP interaction, a bounding box widget was added and visualized as thin box wire. For the zSpace follow-up study the stylus was used as input device and visualized within the stimuli as a virtual ray. Since this device provides six DOFs, the box widget was unnecessary and the TORP was directly picked by the visualized ray. A TORP interaction was realized by interacting with the stylus.

3 www.kurzmed.de

Due to the results and comments of the pilot study, four patient-specific and five manually generated stimuli were presented. Moreover, these nine stimuli were categorized into three simple, three moderately difficult and three difficult stimuli (see Figure 7.4). Thus, different degrees of difficulty similar to patient-specific anatomy were provided. These categories were defined by the eardrum's orientation combined with the resulting occlusion. If the disk-shaped eardrum and the footplate are oriented parallel (facing each other), it is easier to position the TORP. Figure 7.4a illustrates a stimuli scene defined as a simple stimulus. The ear drum and footplate are almost parallel. Increasing deviation of this orientation aggravates the positioning. Increased occlusion of the middle ear cave and the footplate leads to increasing positioning effort and requires better depth perception (see Figure 7.4c).

7.3 APPARATUS AND PROCEDURE

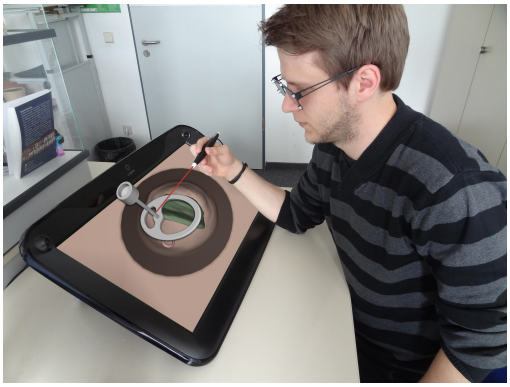
An evaluation tool was developed using MeVisLab from MeVis Medical Solutions AG[4], a development environment from MeVis Medical Solutions AG and Fraunhofer MEVIS, combined with the scripting language Python™. The evaluation tool presented one stimulus at a time and an additional dialog to select the optional yardstick in the beginning and then to chose the desired implant length. Moreover, participant-specific data was recorded and saved as a stimulus-specific xml file for each participant. The files were anonymized and included for each stimulus and participant whether a yardstick was selected or not, the chosen TORP length, TORP position, number of interactions for each TORP and scene and the task completion time. As mentioned above, one group viewed the stimuli on a 2D display and one on an autostereoscopic display, as illustrated in Figure 7.5a. The follow-up study investigated the glass-based zSpace system. The displays are full-parallax two-view (binocular) system providing head tracking and thus integrating binocular and motion parallax, see Section 3.3.3. All participants were tested alone by daylight and the stimuli were viewed from a distance of approximately 0.4 m.

THE 2D DISPLAY is a 24″ widescreen display with 1920 × 1200 pixels manufactured by Fujitsu (P24W-6IPS).

THE AUTOSTEREOSCOPIC DISPLAY is a custom-built 3D display by Fraunhofer HHI. The display is a 3D zeroCreative [5] display also with a 24″ monitor (1920 × 1200 pixel full HD). A head-tracking unit with two tracking cameras, the corresponding tracking technology and software was developed and integrated by Fraunhofer HHI [22]. The head-tracking unit tracks the participants' eyes and based on that calculates the head position. The provided software generates images for the right and left eye that were rendered as textures. The textures are merged together corresponding to the current head position (single-view mode). This image data interleaving results in one output image, which is perceived stereoscopic from the tracked viewer.

4 www.mevislab.de
5 www.zerocreative.com

(a) Autostereoscopic Display



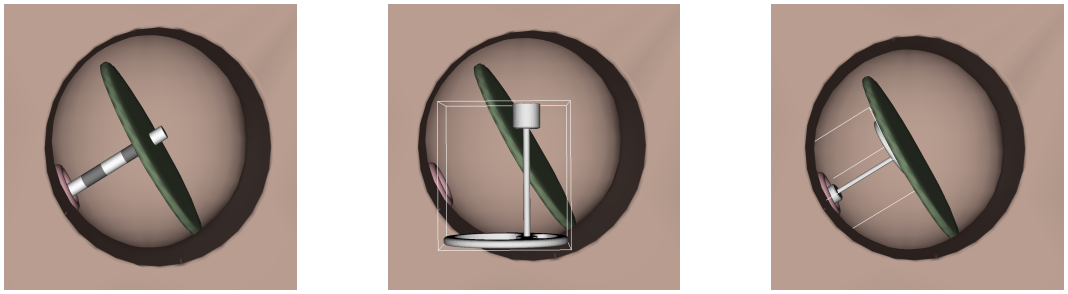
(b) zSpace

Figure 7.5: (a) The 3D autostereoscopic display is a custom-built 3D display by FRAUNHOFER HHI with a head-tracking unit to provide a view-dependent visualization. (b) The zSpace system is a glass-based system. Participants manipulated the TORP placement and orientation with the stylus input device. (Image (b) reprinted, with permission, from Preim and Dachselt [127] © Springer-Verlag Berlin Heidelberg 1999, 2010.)

THE ZSPACE SYSTEM comprises a 3D virtually imaging display with a passive circular 120 Hz stereo 3D polarization technology developed by ZSPACE INC. The left and right view required for stereoscopic vision is generated with circular polarized light. The passive 3D glass separates the two stereoscopic views by filtering the oppositely circular polarized light. Moreover, this glass-based systems includes an optical tracking giving an angle-dependent stereoscopic view of 1920×1080 pixels full HD with time-interleaved stereo frames. The position of the viewer's eyes relative to the screen is determined by tracking the infrared markers attached to the 3D glasses to generate personal perspective views and realize the binocular and the motion parallax effect (Fishtank virtual reality [171]). As illustrated in Figure 7.5b, a six DOF stylus input device is provided for intuitive 3D manipulation and navigation.

First, we performed a pilot study that followed a *within-participant* design. This study showed a stimulus recognition effect. If participants of the pilot study performed with the 2D display first, they got a mental model of the 3D stimuli scene by rotating and interacting with the individual scenes and vice versa by the depth visualization of the 3D displays. Thus, when the participants saw a stimulus the second time, their depth perception and their 3D orientation was influenced by prior stimuli knowledge. Aside from that, the participants complained about the study duration, the number of stimuli and most frequently commented on the varying stimuli difficulty. Thus, we decided to divide the participants into a 2D and a 3D group for the final evaluation and, therefore, to follow a *between-participant* design.

In advance, all observers were instructed in written form. This instruction included a short medical background information, the explanation of all shown structures as well as the task introduction and optimal TORP position definition. Two practice trials followed to familiarize each participant with the task and inter-



(a) Yardstick (b) TORP Start Position (c) End Position

Figure 7.6: (a) Initially, a depth judgment is required to chose the correct TORP length. (b) The TORP is displayed and has to be positioned between (c) the eardrum and the footplate. (Image reprinted, with permission, from Baer et al. [9] © Eurographics Association 2014.)

action technique. The more practice trials are presented, the more familiarization is achieved. Instead of one practice stimulus, we used two for this training trial to minimize variance and learning effects. These two stimuli were not used during the experiments. As soon as the participants understood the task and got familiarized with the stimuli and interaction technique, the study started. For each stimulus, participants were asked:

1. to estimate the appropriate prosthesis length and
2. to implant the TORP.

The stimuli were shown randomly to each participant. Participants were able to zoom in the stimuli until the camera reached the entry of the middle ear. A restricted scene rotation defined by the middle ear bounding box was provided, too. As long as the middle ear entry was almost fully oriented to the viewer, rotation was possible. At the end of each evaluation, a questionnaire had to be filled out. Besides demographic data like age and gender, we asked for their background and experience with 3D visualizations, medical knowledge, experience with 3D input and output devices, comments, and their personal perceived support by the display.

1. TORP LENGTH ESTIMATION. Initially, participants saw one stimulus as shown in Figure 7.4 and were asked to estimate the distance between the eardrum and the footplate. If necessary, they had the opportunity to chose a yardstick, which was integrated in the stimulus and positioned between the eardrum and the footplate, as shown in Figure 7.6a. The yardstick was a small cylinder with differently colored areas of known extension (0.5 mm). As soon as the participants felt certain about the correct TORP length, they had to choose their desired TORP. The yardstick and the implant length selection were realized within a small dialog displayed at the side. Potential TORP lengths were presented as values with a step size of 0.25 mm. Besides the correct length, three other TORP lengths were offered. After selecting the desired length, the TORP was automatically placed axis-aligned in front of the middle ear (Figure 7.6b).

2. TORP IMPLANTING. Participants were asked to navigate the TORP through the ear canal to bridge the gap between eardrum and footplate (Figure 7.6b and 7.6c). A correct position is achieved when the TORP touches the eardrum without penetrating it and bridging the gap between eardrum and stapes footplate. The TORP's orientation is defined as optimal when the TORP's "wheel" part fully touches the eardrum and the longitudinal axis is 90 degrees to the eardrum (compare Figure 7.6c). This concrete orientation definition facilitated the task understanding and the object placement. The TORP was manipulated using the mouse device for the 2D and the 3D display and with the stylus for the zSpace system.

Each stimulus was presented until the participants pressed a "Ready" button to indicate that they were satisfied with the position and ready to move on to the next stimulus.

7.4 ANALYSIS AND RESULTS

Accuracy in terms of the object placement (translation deviation) and orientation (angular deviation), the interaction effort regarding scene and TORP, and the task completion time were analyzed using inferential statistics. Additionally, we analyzed the chosen implant length and thus the depth judgment and distance assessment of eardrum and footplate descriptively. The software package IBM SPSS STATISTICS (Statistical Package for the Social Sciences) was used for the descriptive and statistical analysis.

In detail, two independent samples of 21 participants each, the 2D and the 3D autostereoscopic group as the two *between factors* and the three degrees of difficulty as *within factors* were quantitatively analyzed. The follow-up study was qualitatively analyzed, since a quantitative comparison of the 12 participant results with the 21 of the two other groups is methodically incorrect. The quantitative analysis that will be presented in this section primarily differs from the analysis introduced by Baer et al. [9] in the applied test and in the number of gathered results for the zSpace follow-up evaluation, respectively. Initially, the Shapiro-Wilk test was applied to test for normally distributed results. Based on that, the non-parametric Wilcoxon-Mann-Whitney U test for two independent samples and the parametric 2×3 factorial analysis of variance (ANOVA) combined with a t-test were used. When applying the ANOVA and the t-test, a statistically significant difference exists, if $F \geqslant F_{crit}$ and $t \geqslant t_{crit}$. As listed in Table A.2, the critical value F_{crit} for $\alpha = .05$, $df_1 = 1$ and $df_2 = 40$ is 4.08 and $t_{crit} = 2.02$. Moreover, the effect sizes r for the Wilcoxon-Mann-Whitney U test and omega squared (ω^2) was determined for ANOVA with $k = 2$ for the number of groups, see Equation 5 and 6 in Section 2.3.2. A resulting $\omega^2 = .01$ indicates a small effect, $\omega^2 = .06$ a medium and $\omega^2 = .14$ a large effect. The Pearson's correlation coefficient r defines an effect size with a range of $-1 \geqslant r \leqslant 1$, see Section 2.3.2.

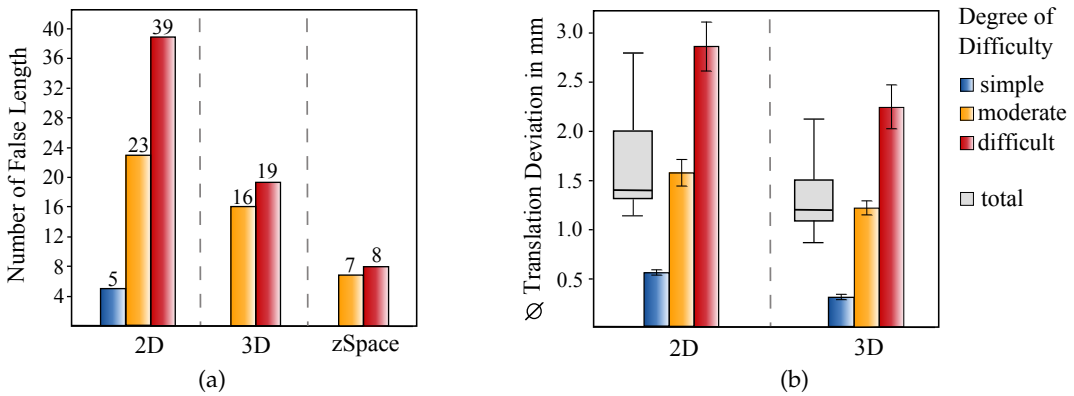


Figure 7.7: (a) The number of all falsely chosen TORP lengths separated by the degree of difficulty of the shown stimuli. (b) The average translation deviations in mm illustrated for all stimuli and for each degree of difficulty for the 2D and the autostereoscopic (3D) group. The bar charts include the standard error of the mean as standard error bars. ((b) reprinted, with permission, from Baer et al. [9] © Eurographics Association 2014.)

7.4.1 *Depth Perception*

Since the participants had to perform a depth judgment, the position, orientation, and chosen TORP length were determined and analyzed for each stimulus to evaluate the accuracy. To validate the accuracy, a *gold standard* TORP length and position is required for each individual stimulus. These *gold standard* TORPs were determined in advance by our medical experts.

LENGTH. The depth judgment of structures and their relationship was analyzed by the chosen TORP length. 189 choices were recorded for the 2D and the autostereoscopic group and 108 for the zSpace group. Figure 7.7a presents the number of all falsely chosen TORPs. Overall, there were 67 false length choices in the 2D group, 35 in the autostereoscopic group and 15 in the zSpace group. When a simple stimulus was presented, a TORP with a deviating length of .25 mm was chosen five times by the 2D group, while participants of the 3D autostereoscopic and the zSpace group chose correct lengths. The moderately difficult stimuli lead to 23 false choices in the 2D group, 16 in the autostereoscopic group and 7 in the zSpace group. In the 2D group 52% of all false choices for a moderately difficult stimulus deviated by .25 mm, 21% deviated by .5 mm, and 26% deviated by .75 mm. Contrary, all falsely chosen TORP lengths by the two 3D groups deviated from the correct length by .25 mm. While the wrong choices for the difficult stimuli of the 3D groups were almost similar to the moderately difficult stimuli, there were more falsely chosen TORPs in the 2D group (39), as shown in Figure 7.7a. The results of the autostereoscopic group included 17 TORP lengths that deviated by .25 mm, one by .5 mm and one by .75 mm from the correct TORP length. Participants of the zSpace group chose five TORP lengths that deviated by .25 mm and three with a difference of .5 mm. The 39 falsely chosen TORP lengths of the 2D group were divided into 34 choices that deviated by .25 mm, three by .5 mm and two by .75 mm. This gives only a little insight into the depth perception of structure distances, but has to be evaluated in detail.

	TORP Placement			$\bar{x}$ Degree of Difficulty		
	$\bar{x}$	σ	Shapiro-Wilk	simple	moderate	difficult
2D	1.69	.49	$p \leq .05$	.56	1.60	2.90
3D	1.29	.32	$p > .05$	.32	1.25	2.28
zSpace	.81	.29	$p > .05$	.19	.85	1.38

Table 7.1: The mean ($\bar{x}$) translation and the standard deviation (σ) in mm for the 2D, 3D and zSpace results. Additionally, the Shapiro-Wilk test results for all stimuli are listed on the left hand side. Results with $p \leqslant .05$ indicate a statistically significant difference from a normal distribution. On the right hand side are the results for each degree of difficulty.

PLACEMENT. The accuracy of the TORP placement was defined as the average translation deviation compared to the *gold standard* TORP position. For each participant and stimulus, the 4×4 transformation matrix T_{sample} of the TORP manipulation was recorded. This matrix included the translation and rotation performed by the participants to place the TORP. To calculate the translation accuracy, we defined three control points along the TORPs longitudinal axis (both ends and a mid point). These points were multiplied with T_{sample} and thus transformed according to the participants' TORP transformations. The points' distance to the *gold standard* position was then calculated. We refer to this as $\Delta trans$ for the translation difference. A little variation of $\pm$ 0.2 mm was negligible for the TORP placement, since this corresponded to the footplate's diameter. As long as the TORP's foot (thin end) was positioned within the footplate, a correct hearing restoration is possible. The zSpace group ($\bar{x}_{\Delta trans} = 0.81$ mm) and the autostereoscopic group achieved on average ($\bar{x}_{\Delta trans} = 1.29$ mm) a more accurate position estimation result than the 2D group with $\bar{x}_{\Delta trans} = 1.69$ mm. The zSpace group results definitely yield an improvement in the TORP placement accuracy.

Table 7.1 covers the Shapiro-Wilk results with the results for the 2D display being normally distributed and the autostereoscopic display being not normally distributed. Thus, the Mann-Whitney U test combined with a Bonferroni correction was used for the analysis. Since the results were analyzed twice – independent and dependent on the stimulis' degree of difficulty–, a Bonferroni correction was applied. The easiest method to use this correction is to use a critical value for p divided by the number of conducted tests [53]. In this case, a statistically significant difference exists, if $p \leqslant .05/2$. A medium to large statistically significant main effect existed with $p \leqslant 0.01, z = -2.981$ and $r = -.45$. Thus, $H_{accTransl}$ is likely to be true. TORPs were placed more accurately – as measured by smaller translation deviations – using an autostereoscopic display compared to a 2D display.

The Wilcoxon-Mann-Whitney U test confirmed a large statistically significant difference for the simple stimuli with $p \leqslant .001$, $z = -4.116, r = -.63$ and a medium effect for the difficult stimuli with ($p = .015, z = -1.850, r = -.33$), compare Table 7.2. Moderately difficult stimuli are marginally not significant with $p = .03, z = -1.850$ and $r = -.28$ (2D group: $\bar{x}_{\Delta trans} = 1.60$ mm and 3D group: $\bar{x}_{\Delta trans} = 1.25$ mm). The results of the follow-up study showed a tendency to a more accurate TORP placement for each degree of difficulty using the zSpace combined with the stylus. However, the results are based on only 12 participants. As shown in Figure 7.7b, the degree of difficulty is chosen properly. The average results confirm the three gradations. Especially stimuli with high difficulty show

Comparison	All Degrees	Simple	Moderate	Difficult
2D - 3D	z = - 2.981 p ≤ .01; r = -.45	z = - 4.116 p ≤ .001; r = -.63	z = - 1.850 p = .03; r = -.28	z = - 2.151 p = .01; r = -.33

Table 7.2: This table covers the p, the calculated *z-score* and the Pearson's correlation coefficient r (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$ effect) for the comparison of the 2D with the 3D group for all results and divided by the stimulis' degree of difficulty. If $z \notin [-1.96, 1.96]$, a significant difference exists with $p \leqslant .025$. Green represents statistically significant and red statistically not significant differences.

higher translation deviations. The autostereoscopic display improves the depth perception and facilitates the TORP placement especially for simple (2D group with $\bar{x}_{\Delta trans} = .56$ mm and 3D group with $\bar{x}_{\Delta trans} = .32$ mm) and difficult (2D group with $\bar{x}_{\Delta trans} = 2.90$ mm and 3D group with $\bar{x}_{\Delta trans} = 2.28$ mm) stimuli.

ORIENTATION. To define the orientation accuracy, each TORP was treated as a 3D unit vector $\vec{v}$. This vector $\vec{v}$ was transformed by T_{sample}, and the angular difference $\Delta angle$ to the *gold standard* orientation is calculated. Rotations around the longitudinal axis were are not considered, since the majority of our participants was medically knowledgeable and no expert. Moreover, this rotation has less impact when analyzing depth perception, it just defines the position of a TORP's "wheel" part relative to the ear canal.

	TORP Orientation			$\bar{x}$ Degree of Difficulty		
	$\bar{x}$	σ	Shapiro-Wilk	simple	moderate	difficult
2D	3.31	.77	p > .05	1.50	3.32	5.11
3D	2.27	.53	p > .05	1.01	2.54	3.25
zSpace	1.77	.32	p > .05	.88	1.86	2.37

Table 7.3: The mean ($\bar{x}$), the standard deviation (σ) in ° and the Shapiro-Wilk test results for all stimuli are listed on the left hand side. All results are normally distributed. On the right hand side are the results for each degree of difficulty.

Figure 7.8 presents the average angular deviations of both groups. The autostereoscopic group and the zSpace group achieved less angular deviations (3D group: $\bar{x}_{\Delta angle} = 2.27°$ and zSpace group: $\bar{x}_{\Delta angle} = 1.77°$) than the 2D group ($\bar{x}_{\Delta angle} = 3.31°$). However, all results were close to the *gold standard* with small standard deviations, as listed in Table 7.3. Angular deviations of $\Delta rot > 2°$ are recognizable. The first applied Shapiro-Wilk test confirmed statistically no significant difference to a normal distribution. Therefore, a parametric 2×3 factorial ANOVA with Bonferroni correction was applied. H_0 related to orientation accuracy is highly unlikely with $p \leqslant .01$ and $\omega^2 = .37$. Thus, H_{accRot} is likely to be true. The t-test with Bonferroni correction confirmed one-tailed statistically significant differences for the comparison of the 2D and the autostereoscopic display based on each degree of difficulty. As shown in Table 7.4, participants of the 3D group oriented the TORP statistically more accurate for the simple ($p < .01, t(40) = 2.851$ and $r = .41$), the moderately difficult ($p = .01, t(40) = 2.699$ and $r = .39$), and the difficult stimuli ($p \leqslant .01, t(40) = 3.206$ and $r = .45$). The re-

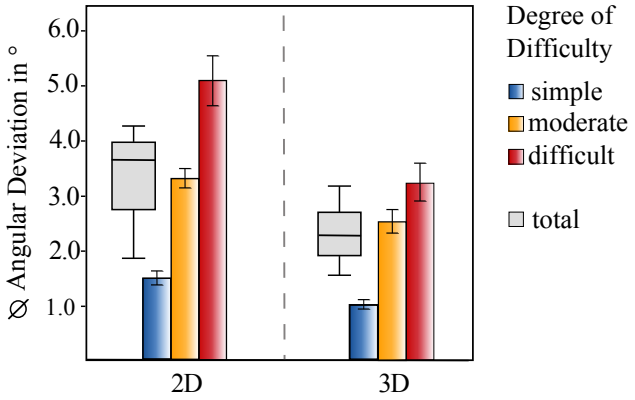


Figure 7.8: The average angular deviations in degree illustrated for all stimuli and for each degree of difficulty for the 2D and the autostereoscopic (3D) group. The bars include standard error bars representing the standard error of the mean.

sults of the zSpace group indicate a tendency to a more accurate TORP orientation for the simple, moderately difficult and difficult stimuli, compare Table 7.3.

Participants of the two 3D groups chose more correctly sized TORPs and positioned these more accurate by less translation and angular deviations. Even for the simple and moderately difficult stimuli, significant differences exist between the two displays.

Comparison	All Degrees	Simple	Moderate	Difficult
2D - 3D	$\omega^2 = .37$ $p \leq .01$	$t(40) = 2.851$ $p \leq .01$; $r = .41$	$t(40) = 2.699$ $p = .01$; $r = .39$	$t(40) = 3.206$ $p \leq .01$; $r = .45$

Table 7.4: The p value, the ω^2 effect size, the *t-statistics* ($t_{crit} = 2.02$) and the Pearson's correlation coefficient r (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$) effect for the comparison of the 2D with the 3D group based on all results and divided by the degree of difficulty are listed. All results are statistically significant.

7.4.2 *Interaction*

The interaction effort was measured by recording the number of performed interactions with the stimuli (middle ear) and the number of required TORP interactions. This was used to verify the task completion times and to analyze whether interaction with the scene and thus motion was used to perceive depth cues.

As presented in Table 7.5a, participants of the 2D group required on average 13.26 scene interactions, while participants of the 3D group required on average 9.34 and of the zSpace group .26 scene interactions until the TORP was placed. Only three participants of the zSpace group required scene interactions until the TORP was placed. All other participants did not rotate the scene and only used the head-tracking opportunity to explore the middle ear scenes. Moreover, the average scene interaction results for each degree of difficulty yield that the 2D and the 3D group required more scene interactions when a more difficult stimulus was

	Interaction			$\bar{x}$ Degree of Difficulty		
	$\bar{x}$	σ	Shapiro-Wilk	simple	moderate	difficult
2D	13.26	4.56	$p > .05$	11.01	11.29	17.47
3D	9.34	3.29	$p > .05$	6.64	8.88	12.49
zSpace	.26	.42	$p \leq .001$	0	.25	.54

(a) Scene

	$\bar{x}$	σ	Shapiro-Wilk	simple	moderate	difficult
2D	29.11	5.48	$p > .05$	15.20	32.11	40.03
3D	21.75	5.89	$p > .05$	12.50	23.90	28.84
zSpace	4.66	.92	$p > .05$	3.39	4.87	5.73

(b) TORP

Table 7.5: The mean ($\bar{x}$), the standard deviation (σ) and the Shapiro-Wilk test results for the number of (a) scene interactions and (b) TORP interactions. Red colored Shapiro-Wilk results indicate normally distributed results.

presented. However, the 2D group required almost the same number of scene interactions for simple ($\bar{x} = 11.02$) and moderately difficult ($\bar{x} = 11.29$) stimuli, while the number increases for difficult stimuli. The bar chart in Figure 7.9 illustrates the similar results for the 2D group and shows a clear distinction between simple, moderate and difficult stimuli for the 3D group. Especially the difficult stimuli required more scene (2D: $\bar{x} = 17.47$ and 3D: $\bar{x} = 12.49$) and TORP interactions (2D: $\bar{x} = 40.03$ and 3D: $\bar{x} = 28.84$). Participants of the zSpace group performed almost no scene interactions for all moderately difficult and difficult stimuli, compare Table 7.5. The TORP results in Table 7.5b and in Figure 7.9 show that participants using the 3D autostereoscopic display required less TORP interactions to position the implant ($\bar{x} = 21.75$) than participants of the 2D group ($\bar{x} = 29.11$) . Almost twice as much interactions were required for the moderately difficult and difficult stimuli compared to the simple stimuli, while the difference between the moderately and the difficult stimuli was less than eight additional interactions. Participants of the zSpace group, who positioned the TORP using the six DOF stylus device, required on average 4.66 interactions to find an adequate TORP position.

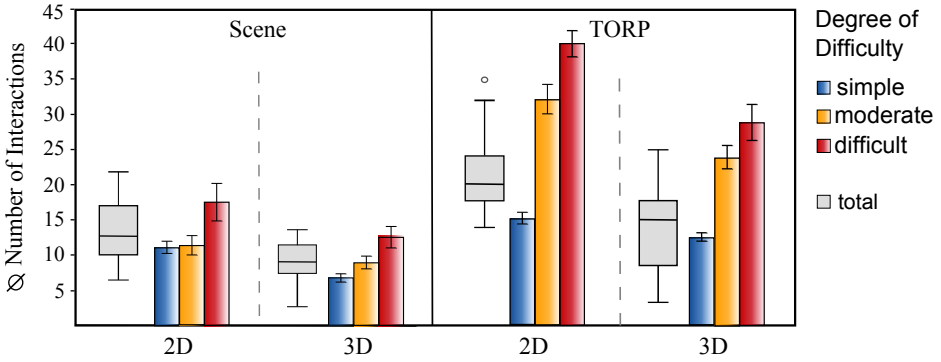


Figure 7.9: The average number of interactions for the scene and the TORP illustrated for all stimuli and for each degree of difficulty. The bars include standard error bars that represent the standard error of the mean. (Image reprinted, with permission, from Baer et al. [9] © Eurographics Association 2014.)

2D versus 3D	All Degrees	Simple	Moderate	Difficult
Scene	$\omega^2 = .18$ $p \leq .025$	$t(40) = 4.081$ $p \leq .001; r = .54$	$t(40) = 1.408$ $p > .05; r = .21$	$t(40) = 1.601$ $p > .05; r = .24$
TORP	$\omega^2 = .28$ $p \leq .01$	$t(40) = 2.470$ $p \leq .025; r = .36$	$t(40) = 2.978$ $p \leq .01; r = .42$	$t(40) = 3.450$ $p \leq .001; r = .47$

Table 7.6: The p value, the ω^2 effect size, the two-sided *t-statistics* ($t_{crit} = 2.02$) and the Pearson's correlation coefficient r (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$ effect) for comparison of the 2D with the 3D group for required number of scene and TORP interactions. The results are listed for all stimuli and divided by the degree of difficulty. Green colored results represent statistically significant differences between the 2D and 3D group.

All results were normally distributed and thus the parametric ANOVA and the t-test applied. Since the ANOVA is a non-specific test and is generally used to test whether there is a difference or not, it is appropriate for our postulated two-tailed hypotheses. Additionally, we determined the two-sided significance value for the t-test. With $F(1,40) = 10.19, p \leqslant .05, \omega^2 = .18$ the first two-tailed hypothesis $H_{actionScene}$ and with $F(1,40) = 17.55, p \leqslant .01, \omega^2 = .28$ the second hypothesis $H_{actionTORP}$ is likely to be true. There is a statistically significant difference between the 2D and the 3D group in the number of scene and TORP interactions.

With $p \leqslant .001, t(40) = 4.081, r = .41$ a large significant difference between the 2D and 3D display was confirmed for the scene, and with $p \leqslant .025, t(40) = 2.470$ and $r = .36$ a medium statistically significant difference exists for the TORP interactions when the participants saw simple stimuli. For moderately difficult and difficult stimuli, we were able to confirm medium effects for the required TORP interactions (moderate: $p \leqslant .01, t(40) = 2.978, r = .42$ and difficult: $p \leqslant .001, t(40) = 3.450, r = .47$).

7.4.3 *Task Completion Time*

The 2D group required on average 114.12 s and the autostereoscopic 3D group 99.72 s to position the TORP, see Figure 7.10. Even though the zSpace group was smaller, their task completion time results were remarkably shorter for all stimuli with $\bar{x} = 22.55$ s and for each degree of difficulty, compare Table 7.7. The standard deviation result $\sigma = 4.02$ s indicates that the measured times tended to be very

	Task Completion Time			Degree of Difficulty		
	$\bar{x}$	σ	Shapiro-Wilk	simple	moderate	difficult
2D	114.12	24.69	$p > .05$	97.16	115.22	129.98
3D	99.72	17.43	$p > .05$	82.74	105.58	110.82
zSpace	22.55	4.02	$p > .05$	17.44	22.22	27.99

Table 7.7: The mean ($\bar{x}$) and standard deviation (σ) results in seconds and the Shapiro-Wilk test results for all stimuli and for each degree of difficulty. All results are normally distributed.

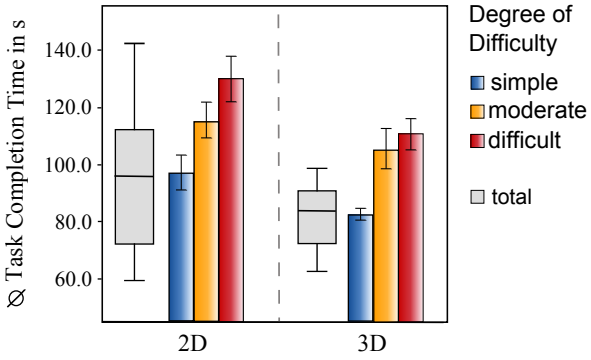


Figure 7.10: The average task completion time in seconds illustrated for all stimuli and for each degree of difficulty for the 2D and the autostereoscopic (3D) group. The bar charts include the standard error of the mean visualized as standard error bars. (Images reprinted, with permission, from Baer et al. [9] © Eurographics Association 2014.)

close to the calculated mean. Contrary to that, the standard deviation of the 2D group with $\sigma = 24.69$ s represents results that were spread out over a wider range of values. As presented in Table 7.7, the Shapiro-Wilk test confirmed normally distributed time results and therefore a parametric analysis was performed.

The ANOVA determined with $F(1, 40) = 4.77$, $p \leqslant .05$ and $\omega^2 = .08$ a medium statistically significant effect for the 2D group compared to the 3D group. Thus, we confirm that $H_{taskTime}$ is likely to be true and that there is a difference between the 2D and the 3D display. The t-test confirmed a two-tailed statistically significant difference between the 2D and 3D group ($p \leqslant .05$ and $t(40) = 2.19, r = .32$). The difference of the difficult stimuli with $\bar{x} = 129.98$ s for the 2D and $\bar{x} = 110.82$ s for the 3D group was with $p = .055$ and $t(40) = 1.97$ statistically not significant different. The same applies for the difference between the moderately difficult stimuli $p = .335$ and $t(40) = 0.98$.

Comparison	All Degrees	Simple	Moderate	Difficult
2D - 3D	$\omega^2 = .08$ $p \leq .05$	$t(40) = 2.199$ $p \leq .01$; $r = .32$	$t(40) = .981$ $p > .05$; $r = .15$	$t(40) = 1.975$ $p > .05$; $r = .29$

Table 7.8: The p value, the ω^2 effect size, the two-sided *t-statistics* ($t_{crit} = 2.02$) and the Pearson's correlation coefficient r (small: $r \geqslant |.10|$, medium: $r \geqslant |.30|$ and large: $r \geqslant |.50|$ effect) for the comparison of the 2D with the 3D group based on all results and divided by the degree of difficulty are listed. Green colored results represent statistically significant differences between the 2D and 3D group.

7.5 QUALITATIVE RESULTS

Participants of the *within-participant* pilot study were asked to compare both displays and stated their display preference. We used a 5-point Likert scale, with each pole representing one display compared to the other. Four of the ten participants rated the 2D display as *good* compared to the 3D display. Four did not prefer one display and rated *neutral* and three rated the 3D display with *good*, since they liked

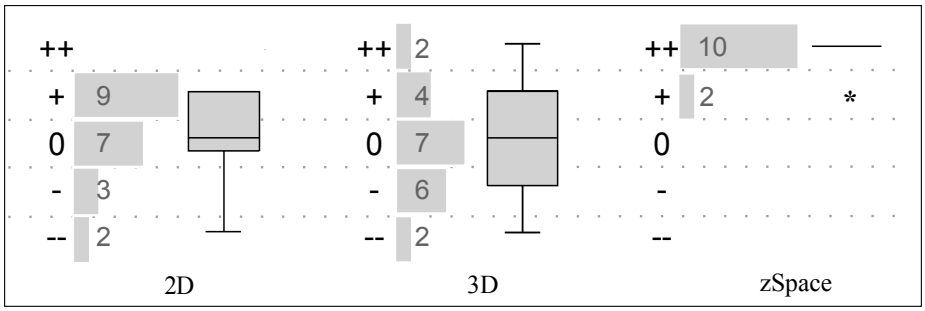


Figure 7.11: The qualitative results of all displays. While the 2D group tended to rate the display with + or 0, the 3D group results were normally distributed around 0 (neutral) and the zSpace results show a preference for ++.

the 3D depth visualization, even though there were some artifacts. No participant rated one display with *very good*.

Participants of the *between-participants* final evaluation were asked to rate one display. The personal preference regarding the display and the provided support for this task had to be assessed. Therefore, a 5-point bipolar Likert scale $(--, -, 0, +, ++)$ was used. As illustrated in Figure 7.11, the results of the 3D group were normally distributed around 0. That means the participants tended to a neutral opinion for the TORP implanting task with an autostereoscopic display. 19% rated with +, 33.3% with 0 and 28.6% with −. While two participants strongly liked the stereo visualization and rated with ++, no participant of the 2D group rated the 2D display with ++. Overall, the participants of the 2D group rated this display either with 0 or with +. Ten participants of the zSpace group rated the display and the stylus with ++ and two with +. Every single participant was enthusiastic positioning the TORP using the stylus as input device.

In summary, the 3D group tended to a neutral opinion and, thus, to no support of the 3D autostereoscopic display, while the 2D group showed a tendency to +, and therefore to a slight preference of the 2D display. This resembled the result of the pilot evaluation, where participants had the possibility to compare their experience with both displays.

7.6 RESULT DISCUSSION

Accuracy of depth judgment in terms of the TORP placement (translation deviation) and orientation (angular deviation), the interaction effort (scene and TORP), and the task completion time were quantitatively analyzed. The chosen implant length as an additional depth judgment aspect was qualitatively analyzed. In summary, all postulated hypotheses were statistically confirmed. The autostereoscopic display enabled a more accurate TORP implantation in terms of the achieved TORP position and orientation to bridge the gap between eardrum and footplate compared to a 2D display. Depth perception was improved using a 3D autostereoscopic display compared to a 2D display. Moreover, there is a statistically significant difference in the number of required scene and TORP interactions and the task completion time between an autostereoscopic display and a 2D display.

In detail, the TORP orientation results of the 2D versus 3D group exhibited remarkable differences in the average angular deviations compared to the translation results of the TORP placement. While the 3D group orientated the TORPs in all degrees of difficulty significantly more accurate than the 2D group, statistically no significant difference existed for the TORP placement when moderately difficult stimuli were presented. However, simple and difficult stimuli exhibited significant differences. Participants of the 3D group chose more correct TORPs and positioned them more accurately in terms of smaller translation and angular deviations. Comparing the number of interactions and time results, the 2D group tried to improve the depth perception by scene and TORP interactions. Especially, rotating the TORP required more interaction than the translation to find the correct depth position. Though, the task completion time results of the 3D group were still high and close to the 2D group results without the same interaction effort. Since this is considered to be a training scenario, the required time is interesting and therefore measured but not crucial.

The quantitative results of the participants' performances, however, indicated the advantages of the 3D display. Additionally, a pairwise comparison showed a statistically significant difference even for each degree of difficulty. The results illustrated the correlation between increasing difficulty and perceptual effort. Thus, the importance of different stimuli with a varying degree of difficulty was confirmed. Significant differences with $p \leqslant .01$ between the degree of difficulty for object placement, orientation, TORP interaction and task completion time were achieved. Differences with $p \leqslant .05$ were achieved for scene interaction. Thus, the stimuli classification was confirmed and simple, moderate and difficult were chosen carefully.

LIMITATIONS. Observations during the study and a few participants' comments showed that using the 3D autostereoscopic display is not as comfortable as the 2D display. Participants had to refocus during the study to receive a proper 3D stereo image. Moreover, there are still minor disturbing crosstalk effects left that hamper the 3D viewing and slow down the TORP placement. These disadvantages had an impact on the results of the qualitative analysis. Disturbing artifacts and a demanding 3D visualization caused by the display technology led to a more negative rating of the 3D autostereoscopic display compared to the new zSpace technology. Furthermore, the position and thus translation deviation was biased by the visible occlusions of TORP and footplate. Some participants used this information to find the correct position. Occlusion as a bias factor was identified by the participants' comments and observations and should be addressed and eliminated in further studies. Concluding, we did not measure the path length for each interaction. Especially, for the scene interaction, a more individual specification is useful and important to analyze the received depth from motion effect and to distinguish from interactions that were performed to improve the viewport, the middle ear access or to get an overview in the beginning.

The follow-up evaluation outlined the potential of the zSpace technology. The 3D stereo effect is very realistic and the view angle-dependent visualization technology enabled the scene exploration by moving the head. Due to that, the 12

participants required no scene interaction – rotation or zoom – to improve the depth perception or to enhance their point of view to position the implant for the simple stimuli. The moderately difficult and the difficult stimuli required almost no interaction but at least considerably less interactions than the autostereoscopic display. It is possible to look on the side of the eardrum, and thus to verify the TORP position without rotating the scene. Thus, only three participants (2.5%) required scene rotations to position the TORP. Additionally, the stylus enabled a really fast and intuitive object placement with a few stylus interactions, since six DOFs were provided. Both, the system combined with the stylus enabled a more accurate object placement and orientation and all participants were enthusiastic implanting the TORP with the zSpace system. This realistic and intuitive 3D stereo vision conduced the TORP implanting and accuracy success. Participants were impressed by the realistic 3D visualization and depth perception. However, this evaluation combined an output and input device and therefore the results cannot be clearly derived from the display. For future work, a well-defined set of tympanoplastic stimuli is required, an evaluation redesign including the improvement of the above-mentioned limitations and an extended evaluation with the zSpace system is recommended. An extended training environment is necessary, e.g., a visual or auditory feedback of the prosthesis position as well as a positioning support should be considered. Since this TORP placement task focused on the support of 3D displays as potential output devices combined with depth cues, this was neither integrated nor considered.

7.7 LESSONS LEARNED

To evaluate the potential of 3D displays for surgical procedures, preoperative planning tasks or surgical training scenarios, the design of appropriate anatomical stimuli and the applicable implementation of the surgical task is necessary. The anatomical visualizations that serve as stimuli have to be as simple and reduced as possible to focus on the important and major structures. This promotes the reduction of bias factors, supports the generation of reliable and traceable results and the stimuli understanding as well as the recruiting of a bigger sample size with a broad spectrum of knowledge to participate in the evaluation. Due to the limited range of knowledge and experience, the unfamiliar evaluation situation as well as the unfamiliar 3D displays and designed training scenario, a reduced anatomical visualization should be used even if medical students or medical knowledgeable participants were recruited. Furthermore, the presentation of a simplified and abstracted stimuli combined with a rather general surgical task - implant positioning - enables the adaptation to further anatomical domains and tasks, e.g., joint implant placement, and the gathered results represent a tendency for similar tasks. Additionally, simple and well-defined tasks prevent misunderstandings by the participants. However, the patient-individual stimuli have to be preserved to generate realistic results and provide training opportunities.

The evaluation design is difficult and dependent on the targeted results, the available number of stimuli and the participants. Preferably, the number of stimuli is sufficient. Thus, a *within-participant* design is recommended. Each participant performs the evaluation with each display. To prevent stimuli recognition

or gained depth information from the evaluation with the first display, different but equivalent stimuli are presented on the displays. Moreover, to avoid order effects, the stimuli presentation order and the display order have to be randomized and treated as a *between-participant* variable. Thus, a quantitative and qualitative comparison of displays is achieved and the statistical analysis is facilitated. All participants are able to compare the displays based on their experience. However, the evaluation completion time increases, the participants are getting unmotivated or bored, and thus the results of the last stimuli might be influenced by that. A crossover design of the display order compensates this effect but does not prevent participants from getting fatigued and bored.

If the number of stimuli is limited, which is highly likely when performing evaluations with patient-individual datasets, a *between-participant* design is recommended. Each participant performs the evaluation only on one display. There is no order of the displays, and thus no order effect. Differences between the two groups reflect differences of the displays or of the participants, since these are two different groups of people. This design relies on the assumption that the *between-participant* error is small or a sufficient number of participants is available, to ensure that any difference between the two conditions - in our case the displays - is not due to differences of the two groups of participants [40]. A well-chosen sample is therefore the major requirement to minimize potential participant differences as well as methods to determine the group for each participant such as the coin biased crossover technique. However, the same stimuli can be presented, and thus the analysis of the same stimuli on each display is possible. Furthermore, the evaluation duration time is shorter than a *within-participants* evaluation. Unfortunately, more participants are required and they are not able to directly compare the displays. Participants that viewed the stimuli on a 3D display are at least able to estimate a 3D display benefit, since they know a standard 2D display. However, a qualitative assessment of the display preference is possible.

Another possibility is the *within-participant* design with repeated measurements. In detail, all participants assess the same stimuli with all displays but delayed in time. Between the evaluation on one display and on the other display there is a specific amount of time to prevent stimuli recognition by the participants. Thus, participants perform on all displays but the time-delayed qualitative comparison comes at the cost of the participants' memory power and the possibility to recruit the same participants again.

In summary, a convenient evaluation design to qualitatively and quantitatively compare different displays is very difficult. The major evaluation aim and focus primarily defines the appropriate design. A *within-participant* design is the first choice if the subjective comparison and qualitative evaluation is desired. If quantitative differences are favored for an evaluation of displays or any other investigated devices, a *between-participant* is recommended. A *within-participant* design with time-delayed repeated measurements between the independent factors (devices) combines qualitative and quantitative goals.

7.8 SUMMARY

This chapter introduced the design and execution of a comparative experimental between-participant study with 42 participants and a follow-up study with 12 participants. Depth perception was investigated for comparing a 2D display with a glasses-free 3D autostereoscopic display in detail and with the new 3D zSpace technology including a stylus as input device. The evaluation was based on a micro-surgical tympanoplastic procedure, a treatment option of deafness disease. A tympanoplastic training scenario was designed and patient-specific data combined with virtually generated models were used as stimuli. The real-world task consisted of the prosthesis implant positioning to reconstruct the ossicular chain and thus a patient's hearing ability. A substantial benefit of the 3D autostereoscopic display compared to a 2D display regarding depth judgment, task completion time and the number of required scene and prosthesis interactions was found. A statistically more accurate depth judgment (TORP length estimation and positioning) as well as the number of required interactions and the task completion time indicated the advantages of the autostereoscopic display. However, these advantages are only present for the autostereoscopic display as long as the visualizations including color, saturation and contrast are adapted, and visible artifacts are minimized. Passive 3D display systems with the 3D image being compounded by the used glasses exhibit disturbing crosstalk effects. Especially high-contrast and saturated colors enhance the perceived ghosting images. These results are similar to the comparative display evaluation presented by Wilhelm et al. [176] for a laparoscopic task. They also evaluated an autsotereoscopic display manufactured by FRAUNHOFER HHI. Their display promoted the depth perception but failed in the preference ratings and suitability rating for laparoscopic tasks. Similar to our observations, their participants complained about visual artifacts. Thus, the 3D autostereoscopic display improves depth perception, but is not suitable for every visualization. Compared to the results of the zSpace evaluation a glasses-based system enables a very accurate stereoscopic perception and head-tracking and therefore yields in a more accurate stereoscopic 3D visualization. However, a glasses is required and an autostereoscopic displays without the need of any glasses seems to be more appropriate for medical applications, especially for the integration in the clinical routine. It is more natural, since no device is necessary to perceive the stereoscopic visualization and there would be no problems due to sterilization for the operating room. However, the head position in front of such a display is very restricted to receive the correct image for each eye and only halve of the screen resolution is achieved for the visualization, explained in Section 3.3.3. To generate an optimal stereoscopic perception and enable an accurate perception of the visualized structures, the software, the hardware and the mechanical component (e.g., the parallax barrier, the lens grating or array) have to work accurately. Thus, there is still room for improvement of 3D technology to overcome the differences to active 3D stereo systems and to promote the depth perception for individual visualizations.

The follow-up study resulted in an overwhelming depth perception improvement with the zSpace using a stylus input device. The view angle-dependent stereoscopic view enables a scene exploration without a manual scene interac-

tion. This leads to a faster task completion performance and the stylus enabled a more comfortable and precise TORP placement. Based on these overwhelming results, Saalfeld et al. [139] performed an evaluation that includes a cervical spine exploration task with head movement only to evaluate this benefit in detail (explained in Section 3.3.3). Similar to the stylus interaction technique, the participants liked the head-tracking exploration technique when they got used to it. These results combined with the enormous personal preference indicate the high potential of the zSpace system. Our results confirmed what others had found: that 3D displays are superior to comparable 2D displays. Additionally, the evaluation showed the superiority of the glasses-based zSpace technology that provides a 3D stereo visualization coupled to the head position of the observer to provide an angle-dependent view. These results combined with the work presented by Wilhelm et al. [176], Wagner et al. [167] and Storz et al. [152], who found that the improvement is between 19% and 88% in performing different expert levels when changing from 2D to 3D vision, yield to an important assistance in virtual training and surgical skill acquisition. However, virtual training with 3D visualizations and stereoscopic displays will not make a medical novice an expert surgeon, but as presented by Wilhelm et al. [176] the result of inexperienced surgeons using 3D displays were in line with the results for experts using 2D displays.

Preoperative planning tasks such as prostheses implant placement for knee, hip or shoulder joint resemble this presented TORP placement task. The optimal implant design, size and position have to be determined and simulated in advance to avoid adaptive remodeling during the surgery. Depth perception is required to position the implant correctly and spatial relations have to be assessed to chose the optimal implant related to size and shape. Since the stylus facilitates and accelerates the implant handling, this device would support the implant orientation task. Artificial joint implants have to be orientated such that the joint's mobility is restored and an optimal functionality is achieved. This is a trial-and-error task, and thus the stylus is suitable to easily rotate and interact with the implant until an optimal position is found. Moreover, other surgical tasks or treatment planning procedures could benefit from stereoscopic visualizations, as presented by Saalfeld et al. [139]. Spine surgery requires difficult path planning tasks to prevent injuries of the spine canal and at the same time to achieve the best vertebral or intervertebral discs access. There is a considerable amount of work to accomplish for simulation to be accepted as an integral part of surgical training, specifically in the area of curriculum development and the acquisition of cognitive knowledge along with hands-on skills. However, the existing evaluations outline the potential of 3D stereo display technology for surgical training or preoperative planning tasks.

Part III

SUMMARY AND CONCLUSION

8

SUMMARY

The goal of 3D medical visualizations is the representation and communication of the underlying patient-specific image data. Visualizations must allow a clear and precise depiction of complex structures and derived information to aid and facilitate the visual perception and thus the comprehension. The aim is a meaningful, expressive, abstract and simplified representation. Visualization techniques as well as stereoscopic views have a great potential to convey and to show the anatomy and pathologic structures realistically, and to reveal their spatial relations to adjacent risk structures. Thus, the development of new and the improvement of existing techniques and technology is a huge research domain. The selection of adequate techniques and visualizations, the support of an efficient exploration and the effective communication of essential information, however, is challenging. Without the guidance of perceptual-based evaluations, their potential remains underutilized in medical education, training and treatment planning systems.

An evaluation takes many forms, such as observations, interviews, cognitive walkthroughs, expert reviews, or participatory design. It can either be formative during the visualization design process to analyze goals and tasks or summative at the end of the development process. Besides widely used informal evaluations, there is a need for systematic research, since the techniques that people think will improve performance are not always those that actually do improve their performance. Visual perception and attention plays an important role, since an understanding of perception can significantly improve the quality and quantity of the illustrated information. Perceptual theories and experimental evaluations can be used to analyze and verify the suitability and potential of medical visualizations, techniques and devices, and enable a generalization for similar application domains. To generate reliable, valid and reproducible results, the designed evaluation has to measure the effect caused by the investigated technique or visualization as precise as possible. Thus, an evaluation design based on the experimental guidelines presented in Chapter 2 is important.

8.1 CONCLUSION

This thesis presented four customized evaluation designs (qualitative and quantitative), conductions and analyses. 3D isosurface visualizations, illustration techniques and stereoscopic views used for a diagnostic (Chapter 4 and 6) and an educational task (Chapter 7) were investigated. The developed evaluations comprised

a pure quantitative evaluation presented in Chapter 4, a qualitative evaluation introduced in Chapter 5 and two studies that combine both aspects in Chapter 6 and 7. Moreover, illustration techniques were evaluated in terms of their ability to attract the viewer's attention (Chapter 4), to effectively communicate essential information (Chapter 6 and 7) and to illustrate medical structures realistically and aesthetically using only feature line techniques (Chapter 5). The major contribution is the integration and adaptation of empirical criteria like objectivity, validity and reliability to medical visualizations to produce reliable and traceable results.

Each evaluation included an extensive analysis of the visualization goal and medical domain with respect to the common medical applications introduced in Chapter 3. Although the vast majority of findings and techniques from psychophysical studies exclusively use really simple stimuli, e.g., letters, graphical objects such as circle and triangle or colors, guidelines from psychophysical experiments presented in Chapter 2 were adapted to structures such as aneurysms, skull or femur bones as well as to complex scenarios such as neck, thorax anatomy and middle ear for each of the presented studies. Stimuli are generated from patient-specific image data. Tasks were derived from clinical tasks but were adapted for experimental evaluations with participants who had limited medical knowledge. The stimuli, the task abstractions, the detailed instructions and the training sessions promoted the generation of meaningful results. Since the stimuli comprised patient-specific structures and tasks included the important aspects of the clinical application, e.g., structure detection task, positioning task or shape estimation task, the presented results are adaptable to other medical domains, e.g., vessel assessment, structure detection or micro-surgical implant positioning tasks. All recorded results were analyzed with descriptive and inferential statistics methods, respectively.

As introduced in Section 1.2, this thesis aimed at answering three research questions for the evaluation of 3D medical visualizations:

What is an effective and expressive 3D medical visualization?
Since each visualization is generated for a specific purpose, the goal of medical visualizations is to transfer the patient-individual image data into simple visualizations supporting the exploration, interpretation and decision making. As analyzed in Chapter 3, a medical visualization has to accurately display the patient-specific image data, illustrate the application-specific structures and information and aim at an efficient exploration and effective communication of the essential information. To communicate relevant information effectively, the visualization must attract and then guide the viewer's attention to the relevant information. Moreover, this relevant information has to be visualized using techniques that support the visual perception and especially the shape, depth and spatial perception of structures, structure relations and further displayed information, e.g., blood flow or resection plane. This characterizes an effective, and thus, expressive 3D medical visualization.

The ability of the visualizations to fulfill these requirements can be perceptually analyzed. The effectiveness and expressiveness is defined by quantitatively

measured parameters such as task performance, e.g., task completion time, accuracy or error rate as well as subjective assessments, e.g., rating or ranking tasks as realized in the qualitative feature line evaluation presented in Chapter 5. These quantitative measurements enable a cognitive workload analysis with respect to the evaluated tasks and visualizations and an objective technique assessment. An effective visualization is characterized by a facilitated and reduced workload compared to a common visualization or defined gold standard visualization.

For example, the comparative evaluation with two ghosting and one semitransparent technique applied to aneurysm models showed the advantages of ghosting techniques for the effective and expressive visualization of vascular structures and embedded flow (Chapter 6). Moreover, the potential of the ghosting techniques was evaluated by investigating the technique's specific characteristics, e.g., shape and depth enhancement and smart visibility characteristics. This evaluation presented that ghosting techniques support more accurate analyses of aneurysms than the traditional semitransparent visualization technique and that they are more preferred. Additionally, this technique benefits from interaction with the visualized structure. The qualitative and quantitative advantages of stereoscopic views for 3D positioning tasks in micro-surgical procedures were shown in Chapter 7. Since qualitative evaluations are essential to analyze the acceptance and potential application, this aspect was investigated in a qualitative evaluation presented in Chapter 5.

How can psychophysical guidelines be applied to complex 3D isosurface visualizations of medical patient-specific image data to evaluate the effectiveness and expressiveness of the visualization?
As explained and analyzed in Section 2 and practically applied in Part II of this thesis, an experimental evaluation of medical visualizations starts with a detailed goal analysis. Based on that, the visualization can be evaluated, stimuli characteristics can be defined and tasks and measured parameters can be identified. The goal definition is essential for the evaluation concept. The more precise the goal, the more evaluation design steps and tasks are automatically included in this definition.

Generally, terms such as "better", "faster", "improved" or "facilitated" that are often used when a new visualization or technique is presented have to be specified. This specification can be combined with the task that shall be facilitated or improved by the generated visualization. If this is defined properly, there are several methods that can be used to evaluate the human perception, e.g., visual search, gauge figure technique or depth judgment experiments. Accuracy and task completion time are the common task performance measurements to analyze a 3D medical visualization in terms of guidance potential and shape and depth perception. Since these visualizations are generally used for diagnostic, treatment planning or education purposes, the accurate visual perception of the underlying data is essential and an accelerated perception contributes to a task completion acceleration. Since a medical visualization will be generated for a difficult diagnostic or treatment planning task, an accurate perception is more important than task completion time. Thus, the analysis of the technique's ability to illustrate the

structures and information effectively should be analyzed qualitatively. However, a visualization method or technology that quantitatively improves the clinical workflow is still not good enough until it is subjectively preferred and accepted as well.

Overall, it is important to design an evaluation that investigates patient-specific visualizations and real-world tasks to increase the external validity and achieve results that can be generalized to other anatomic domains and medical applications. Part II of this thesis presented experimental evaluations of different applications and anatomic domains and individual experiment designs customized to the investigated medical domain and visualization. Based on these results and gained knowledge, a few concrete restrictions and recommendations were formulated, which are the topic of the next research question.

How is an appropriate evaluation for a 3D medical visualization characterized? To produce valid and reliable results, the evaluation design must include the principles of psychophysical evaluations but still be close to the realistic data and tasks to ensure accuracy and provide realism.

Theoretically, to minimize and keep the bias factors constant, e.g., one new developed visualization technique or method had to be compared to all existing techniques in an experimental evaluation. Furthermore, to increase the internal validity, the design, conduction and analysis have to be performed from different researchers to minimize bias factors and to prevent influencing settings or participants. The evaluation design would follow a within-participant design and patient-specific visualizations of the target anatomic domain are presented as stimuli, which include all possible anatomic variations. All medical experts for this investigated anatomic domain or task were asked to perform the realistic tasks and their performance as well as their subjective preference are measured. These carefully sorted participants (e.g., age and gender) can be categorized into experts with several years of experience and medical students. Moreover, all participants are motivated and were tested at the same time of the day in the same room with the same apparatus. The study should be short and repeatable after weeks to minimize recognition and learning effects and to present all stimuli. Since these requirements do not justify the gained knowledge and results, this is only an ideal process for one evaluation.

Based on the experiments presented in this thesis, a few guidelines can be formulated. The anatomical visualizations that serve as stimuli have to be:

- **patient-specific** that are as realistic as possible and provide variations of the patient anatomy. In psychophysical methods and guidelines, these differences are unwanted and shall be prevented, since this is a bias factor itself. For the evaluation of medical visualizations it is important to include the patient-specific differences and variations, to ensure that the gained results were gathered from such data and can be applied to further individual datasets. Thus, the external validity with respect to other anatomy variations will be maximized.

- **reduced** illustrating the task-relevant structures to maximize the simplicity and focus on relevant structures and information. Each additional structure or information that is not necessary for the task understanding and performance serves as a structure that influences the result and in the worst case biases the measured effect.

- **comparable** to contribute to the internal validity. The stimuli should be as similar as possible, e.g., same number, type, view (sagittal, coronar or axial) or viewport of the illustrated structures and information. However, a limited variety between the stimuli is necessary to provide the individuality of the structure. Moreover, the similarity might be violated if the task requires that the stimuli have to be different, e.g., several steps of a diagnostic process.

Different patient-individual variations and different anatomic domains have to be integrated to promote generalizable results. If possible, real-world tasks are recommended and should be used. However, the appropriate participants to perform these tasks are difficult to recruit or limited. Therefore, tasks can be abstracted to be appropriate for both participants with less medical knowledge and medical experts. If the tasks still comprise the important aspects of the real medical tasks, such as positioning an implant to a specific location, the recorded results enable a statement of the visual perception. It is likely that the measured results can be applied to prospective users (e.g., medical doctors) concerning the perceptual effectiveness, even though medical doctors may achieve better results concerning the accuracy because of their clinical experience with the topology. Well-established tasks are recommended to quantitatively evaluate shape, depth and spatial perception as well as the guidance potential and thus to conduct an evaluation that analyzes the visual perception. Quantitative evaluations are recommended, since they enable results that are more general and objectively measured.

Besides comments and informal feedback, the subjective opinion should be quantitatively measured, too. It is recommended to use methods such as Likert scales with well-established scale items. However, the number of scale items as well as the scale type have to be chosen carefully. Therefore, it is important to define how many differentiations between the answers are required to analyze the potential of the technique. How many degrees of "like" or "dislike" are necessary to show an advantage or disadvantage of a visualization. In this thesis, 5-point Likert scales were presented. The results showed a clear distinction of rating results for the different techniques and thus, the assessment of "good" and "very good" was sufficient enough to show a technique's illustration potential.

A participant categorization according to the targeted medical application is suitable to ensure carefully selected participants and prevent a sampling bias, e.g., due to experience or knowledge. Groups of participants have to be equal to enable technique comparisons, if a between-participant design is used. In all other cases, a within-participant design is favored, except if this is not possible due to stimuli or recognition effects. An evaluation with three medical experts is not as meaningful as an evaluation with 30 participants including three medical experts. Thus, it is recommended to recruit as many medical experts or experienced participants as possible. A task abstraction enables the evaluation with more participants, who

have one similarity with the experts: they get to know and assess the new visualization technique or technology.

8.2 FUTURE WORK

Perceptual studies are an important element in the quality assessment of illustrative visualization techniques. The studies are an important tool for further analysis and refinement. Furthermore, they serve as a quality requirement for subsequent techniques. However, there are still limitations of the presented studies that have to be considered. Initially, the average age of all recruited participants was between 23.66 to 34.2 years and was very restricted to a group of young people and medical students. This might be a problem when applying the result to medical experts that are commonly older due to the period of medical education. An important aspect that influences the visual perception is the decreasing vision ability. This difference and the caused influence on the visual perception of medical visualizations have to be investigated and might be prevented by recruiting more participants with higher age.

An important aspect that needs to be evaluated is the technique's robustness for mesh resolution and quality, as this might influence the visual perception of 3D isosurface visualizations. The resolution of a surface model is required to be as high as possible to allow for a faithful depiction of fine surface details. Thus, the viewer perceives the surface as more smooth and natural and real surface features become more obvious, while a too low mesh resolution yields a stronger emphasis on the edges of the surface mesh. Depending on the triangle size and local surface curvature, these areas could misleadingly be perceived as surface features and the user might not be capable of differentiating them from real surface features. Hence, the mesh resolution should be selected depending on the local surface curvature. Areas with high curvature should be represented with a high resolution. Even though a high visual quality may be achieved by the implemented shading method that strongly reduces the visible influence of the mesh resolution, the shading may not fully compensate for low surface resolutions. A reduced mesh resolution can disturb a smooth surface perception, especially at the surface-background boundary. The edges become visible without being influenced by the employed shading technique. Besides this relation between mesh resolution and common surface rendering, illustrative techniques may have additional requirements to surface models. A high mesh resolution is essential, since computations are usually performed on a per-vertex or per-face basis. Moreover, for many computations the surface is usually required to be smooth and the surface elements are required to be of high quality and abrupt changes of size and shape of neighboring surface elements should be avoided. The latter is especially challenging since anatomical surface models are usually generated from noisy tomographic data yielding additional noise in the resulting surface models. Surface smoothing is often applied to remove these dominant surface artifacts. Illustrative techniques and visualizations have to be evaluated in terms of their robustness for the mesh resolution and quality, since this influences the visual perception and thus the technique's effectiveness and expressiveness.

The development of new illustrative visualization techniques is inspired by limitations of previous work. To identify such limitations and indicate further refinements, an analysis of different scenes or objects and further comprehensive studies comparing and analyzing illustration techniques are required. Unfortunately, only a few existing evaluations adapt psychological guidelines to design evaluations that minimize bias factors, maximize the isolation of the measured effect while still analyzing medical data and applying real-world tasks to achieve valid, reliable and reproducible results. Recently, Lindemann and Ropinski [109], Borkin et al. [17], Kersten-Oertel et al. [84] presented extensive evaluations that integrate psychophysical guidelines and findings and thus present quantitatively and qualitatively evaluated medical visualizations of direct volume renderings. Since medical application is an important domain and adequate techniques are required to support the diagnostic and treatment planning process, the development of perceptually guided technique research is important.

Besides illustrative techniques and 3D visualizations, stereoscopic displays provide the possibility to improved depth and spatial perception without a specific rendering. Despite the fact that stereopsis has been studied extensively in the literature, the effect of illustrative visualizations, techniques and stereopsis individually and in combination on the visual perception of medical images has not yet received enough attention from the medical image visualization community.

However, the presented evaluations from Chapter 4 to Chapter 7 could successfully improve the current state of the art for evaluations of 3D medical illustrative techniques, visualizations and stereoscopic views.

Part IV

APPENDIX

A

APPENDIX - CRITICAL VALUES OF DISTRIBUTIONS

A.1 CHI-SQUARE DISTRIBUTION

df	α		
	0.1	**0.05**	**0.01**
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89
40	51.81	55.76	63.69
60	74.40	79.08	88.38

Table A.1: This table covers the critical values of the χ^2 - distribution $\chi^2_{critical} = \chi^2_{\alpha,df}$. The values are calculated based on the probability level $\alpha \in \{0.1, 0.05, 0.01\}$. df is the degrees of freedom with $df = k - 1$ and k being the number of experimental conditions.

A.2 F-DISTRIBUTION

	df_1									
df_2	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88

Table A.2: This table covers critical values with $\alpha = .05$ and F_{α,df_1,df_2} for the F-distribution. df_1 is the degrees of freedom for the numerator and is defined as the degrees of freedom for the between group ($df_1 = k - 1$; k: number of groups). The degrees of freedom for the denominator is df_2 and defined as the degrees of freedom for the within group ($df_2 = N - k$; N: sample size).

A.3 T-DISTRIBUTION

	Critical values for a one-sided t-test					
α	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.004**
	Critical values for a two-sided t-test					
α / **df**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.008**
1	3.08	6.31	12.71	31.82	63.66	79.57
2	1.89	2.92	4.30	6.96	9.92	11.11
3	1.64	2.35	3.18	4.54	5.84	6.32
4	1.53	2.13	2.78	3.75	4.60	4.91
5	1.48	2.02	2.57	3.36	4.03	4.26
6	1.44	1.94	2.45	3.14	3.71	3.90
7	1.41	1.89	2.36	3.00	3.50	3.67
8	1.40	1.86	2.31	2.90	3.36	3.51
9	1.38	1.83	2.26	2.82	3.25	3.39
10	1.37	1.81	2.23	2.76	3.17	3.30
11	1.36	1.80	2.20	2.72	3.11	3.23
12	1.36	1.78	2.18	2.68	3.05	3.17
13	1.35	1.77	2.16	2.65	3.01	3.13
14	1.35	1.76	2.14	2.62	2.98	3.09
15	1.34	1.75	2.13	2.60	2.95	3.06
16	1.34	1.75	2.12	2.58	2.92	3.03
17	1.33	1.74	2.11	2.57	2.90	3.00
18	1.33	1.73	2.10	2.55	2.88	2.98
19	1.33	1.73	2.09	2.54	2.86	2.96
20	1.33	1.72	2.09	2.53	2.85	2.95
21	1.32	1.72	2.08	2.52	2.83	2.93
22	1.32	1.72	2.07	2.51	2.82	2.92
23	1.32	1.71	2.07	2.50	2.81	2.90
24	1.32	1.71	2.06	2.49	2.80	2.89
25	1.32	1.71	2.06	2.49	2.79	2.88
26	1.31	1.71	2.06	2.48	2.78	2.87
27	1.31	1.70	2.05	2.47	2.77	2.86
28	1.31	1.70	2.05	2.47	2.76	2.86
29	1.31	1.70	2.05	2.46	2.76	2.85
30	1.31	1.70	2.04	2.46	2.75	2.84
40	1.30	1.68	2.02	2.42	2.70	2.79
60	1.30	1.67	2.00	2.39	2.66	2.74

Table A.3: This table covers critical values of the T-distribution for $t_{\alpha,df}$. The degrees of freedom are defined as the sample size minus one ($df = N - 1$).

BIBLIOGRAPHY

[1] Kamyar Abhari, John S. H. Baxter, Ali R. Khan, Terry M. Peters, Sandrine De Ribaupierre, and Roy Eagleson. Visual enhancement of mr angiography images to facilitate planning of arteriovenous malformation interventions. *ACM Transaction on Applied Perception*, 12(1):4:1–4:15, 2015. (Cited on pages 46 and 47.)

[2] John R. Anderson. *Cognitive Psychology and Its Implications*. Worth Publishers, 2005. (Cited on page 43.)

[3] Dörte Apelt, Richard Rascher-Friesenhausen, Jan Klein, Hans Strasburger, Bernhard Preim, and Heinz-Otto Peitgen. Impact of luminance distribution in the visual field on foveal contrast sensitivity in the context of mammographic softcopy reading. In *Proceedings of SPIE Conference on Medical Image Computing*, 2009. (Cited on page 20.)

[4] Luca Augsburger, Philippe Reymond, Edouard Fonck, Zsolt Kulcsar, Mohamed Farhat, M. Ohta, Nikolaos Stergiopulos, and DA Rüfenacht. Methodologies to assess blood flow in cerebral aneurysms: Current state of research and perspectives. *Journal of Neuroradiology*, 36(5):270, 2009. (Cited on pages 115 and 117.)

[5] Alexandra Baer, Christian Tietjen, Ragnar Bade, and Bernhard Preim. Hardware-accelerated stippling of surfaces derived from medical volume data. In *IEEE Eurographics Symposium on Visualization (EuroVis)*, pages 235–242, 2007. (Cited on pages 51 and 76.)

[6] Alexandra Baer, Friederike Adler, Daniel Lenz, et al. Perception-based evaluation of emphasis techniques used in 3d medical visualization. In *Proceedings of Vision Modeling and Visualization (VMV)*, pages 295–304, 2009. (Cited on pages 73, 74, 75, and 78.)

[7] Alexandra Baer, Kerstin Kellermann, and Bernhard Preim. Importance-driven structure categorization for 3d surgery planning. In *Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM)*, pages 99–107, 2010. (Cited on page 42.)

[8] Alexandra Baer, Rocco Gasteiger, Douglas W. Cunningham, and Bernhard Preim. Perceptual Evaluation of Ghosted View Techniques for the Exploration of Vascular Structures and Embedded Flow. *Computer Graphics Forum*, 30(3):811–820, 2011. (Cited on pages 114, 121, 123, 129, 131, and 132.)

[9] Alexandra Baer, Antje Hübler, Patrick Saalfeld, Douglas Cunningham, and Bernhard Preim. A comparative user study of a 2d and an autostereoscopic 3d display for a tympanoplastic surgery. In *Proceedings of Eurographics Workshop on Visual Computing for Biology and Medicine (EG VCBM)*, pages 181–190,

04.-05. September 2014. (Cited on pages 140, 141, 142, 145, 150, 151, 152, 156, and 158.)

[10] Alexandra Baer, Kai Lawonn, Patrick Saalfeld, and Bernhard Preim. Statistical analysis of a qualitative evaluation on feature lines. In *Bildverarbeitung für die Medizin*, pages 71–76, 2015. (Cited on pages 103, 105, and 108.)

[11] Alethea Bair and Donald House. Grid with a view: Optimal texturing for perception of layered surfaceshape. *IEEE Transaction on Visualization and Computer Graphics*, 13(6):1656–1663, 2007. (Cited on pages 49, 50, and 56.)

[12] Juan Barroilhet. Tympanoplastic surgery. *A.M.A. Archives of Otolaryngology - Head & Neck Surgery*, 69(6):704–711, 1959. (Cited on page 141.)

[13] Dirk Bartz, David Cunningham, Jan Fischer, and Christian Wallraven. The role of perception for computer graphics. In *Eurographics State-of-the-Art Report 4*, 2008. (Cited on page 61.)

[14] Mikhail Belkin, Jian Sun, and Yusu Wang. Discrete laplace operator on meshed surfaces. In *Proceedings of Symposium on Computational Geometry*, pages 278–287. ACM, 2008. (Cited on page 103.)

[15] Steven Birr, Jeanette Mönch, Dirk Sommerfeld, Uta Preim, and Bernhard Preim. The LiverAnatomyExplorer: A webgl-based surgical teaching tool. *IEEE Computer Graphics and Applications*, 33(5):48–58, 2013. (Cited on page 36.)

[16] Alexander Bock, Norbert Lang, Gianpaolo Evangelista, Ralph Lehrke, and Timo Ropinski. Guiding Deep Brain Stimulation Interventions by Fusing Multimodal Uncertainty Regions. *Proceedings of the 2013 IEEE Pacific Visualization Symposium (PacificVis)*, pages 97–104, 2013. (Cited on page 37.)

[17] Michelle A. Borkin, Krzysztof Z. Gajos, Amanda Peters, Dimitrios Mitsouras, Simone Melchionna, Frank J. Rybicki, Charles L. Feldman, and Hanspeter Pfister. Evaluation of artery visualizations for heart disease diagnosis. *IEEE Transactions on Visualization and Computer Graphics*, 17:2479–2488, 2011. (Cited on pages 4 and 173.)

[18] Silvia Born, Michael Markl, Matthias Gutberlet, and Gerik Scheuermann. Illustrative visualization of cardiac and aortic blood flow from 4d mri data. In *PacificVis*, pages 129–136, Sydney, 2013. (Cited on page 35.)

[19] Breght R. Boschker and Jurriaan D. Mulder. Lateral head tracking in desktop virtual reality. In *Proceedings of the Tenth Eurographics Conference on Virtual Environments*, pages 45–52, 2004. (Cited on page 60.)

[20] Christian Boucheny, Georges-Pierre Bonneau, Jacques Droulez, et al. A perceptive evaluation of volume rendering techniques. *ACM Transaction on Applied Perception*, 5(4):1–24, 2007. (Cited on page 56.)

[21] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola, and Ivan Poupyrev. *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 2004. (Cited on pages 60, 61, 62, and 63.)

[22] Marius Braun, Ulrich Leiner, and Detlef Ruschin. Evaluating motion parallax and stereopsis as depth cues for autostereoscopic displays. In *Proceedings of SPIE 7863, Stereoscopic Displays and Applications XXII,*, volume 7863, 2011. (Cited on pages 60 and 148.)

[23] Stefan Bruckner and Meister Eduard Gröller. Volumeshop: An interactive system for direct volume illustration. In *Proceedings of IEEE Visualization*, pages 671–678, 2005. (Cited on page 55.)

[24] Stefan Bruckner and Meister Eduard Gröller. Exploded views for volume data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1077–1084, 2006. (Cited on pages 39, 40, and 55.)

[25] Gerhard F. Buess, Patrick van Bergen, Wolfgang Kunert, and Marc O. Schurr. Comparative study of various 2d and 3d vision system in minimally invasive surgery. *Der Chirurg*, 67(10):1041–1046, 1996. (Cited on page 64.)

[26] Oliver Burgert, Veronika Örn, Michael Gessat, Markus Joos, Gero Strauß, Christian Tietjen, Bernhard Preim, and Ilka Hertel. Evaluation of perception performance in neck dissection planning using eye tracking and attention landscapes. In *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment (Proceedings of SPIE)*, volume 6515, pages 65150–65159, 2007. (Cited on pages 59 and 73.)

[27] Franck Caniard and Roland W. Fleming. Distortion in 3d shape estimation with changes in illumination. In *4th Symposium on Applied Perception in Graphics and Visualization (APGV '07)*, pages 99–105, 2007. (Cited on page 56.)

[28] Sheelagh Carpendale. *Evaluating Information Visualizations*, volume 4950 of *Lecture Notes in Computer Science*, chapter I., pages 19–45. Springer Berlin Heidelberg, 2008. (Cited on pages 7, 8, 11, and 16.)

[29] Vincent N. Carrasco and Harold C. Pillsbury III. *Revision otologic surgery*. Thieme Medical Publishers, Inc., 1997. (Cited on page 141.)

[30] Juan R. Cebral, Marcelo Adrián Castro, Sunil Appanaboyina, Christopher M. Putman, Daniel Millian, and Ro F. Frangi. Efficient pipeline for image-based patient-specific analysis of cerebral aneurysm hemodynamics: Technique and sensitivity. *IEEE Transaction on Medical Imaging*, 24(4):457–467, 2005. (Cited on page 115.)

[31] Ming-Yuen Chan, Yingcai Wu, Wai-Ho Mak, Wei Chen, and Huamin Qu. Perception-based transparency optimization for direct volume rendering. *IEEE Transaction on Visualization and Computer Graphics*, 15(6):1283–1290, 2009. (Cited on pages 56 and 58.)

[32] Jian Chen, Haipeng Cai, Alexander P. Auchus, and David H. Laidlaw. Effects of stereo and screen size on the legibility of three-dimensional streamtube visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12): 2130–2139, 2012. (Cited on pages 66 and 67.)

[33] Alan Chu, Wing-Yin Chan, Jixiang Guo, Wai-Man Pang, and Pheng-Ann Heng. Perception-aware depth cueing for illustrative vascular visualization. In *IEEE Computer Society*, pages 341–346, 2008. (Cited on page 52.)

[34] Forrester Cole, Aleksey Golovinskiy, Alex Limpaecher, Heather Stoddart Barros, Adam Finkelstein, Thomas Funkhouser, and Szymon Rusinkiewicz. Where do people draw lines? *Proceedings of ACM SIGGRAPH*, 27(3):107–115, 2008. (Cited on pages 50 and 91.)

[35] Forrester Cole, Kevin Sanik, Doug DeCarlo, Adam Finkelstein, Thomas Funkhouser, Szymon Rusink, and Manish Singh. How well do line drawings depict shape? *Proceedings of ACM SIGGRAPH*, 28(3):1–9, 2009. (Cited on pages 49, 50, and 91.)

[36] Carlos Correa, Deborah Silver, and Min Chen. Feature aligned volume manipulation for illustration and visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1069–1076, 2006. (Cited on page 55.)

[37] John W. Creswell. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches 4th Edition*. SAGE Publications, 2014. (Cited on pages 17 and 18.)

[38] Carolina Cruz-Neira, Daniel J. Sandin, and Thomas A. DeFanti. Surround-screen projection-based virtual reality: The design and implementation of the cave. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, pages 135–142, 1993. (Cited on page 61.)

[39] Balázs Csebfalvi, Lukas Mroz, Helwig Hauser, Andreas König, and Meister Eduard Gröller. Fast visualization of object contours by non-photorealistic volume rendering. Technical Report TR-186-2-01-09, Institute of Computer Graphics and Algorithms, Vienna University of Technology, 2001. (Cited on page 51.)

[40] Douglas Cunningham and Christian Wallraven. *Experimental Design: From User Studies to Psychophysics*. A. K. Peters, Ltd., Natick, MA, USA, 1st edition, 2011. (Cited on pages 4, 7, 9, 13, 14, 15, 16, 21, 22, 26, 27, 28, 93, and 162.)

[41] Douglas W. Cunningham, Mario Kleiner, Heirich H. Bülthoff, and Christian Wallraven. The components of conversational facial expressions. In *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*, APGV '04, pages 143–150, New York, NY, USA, 2004. ACM. (Cited on page 20.)

[42] Doug DeCarlo, Adam Finkelstein, Szymon Rusinkiewicz, and Anthony Santella. Suggestive contours for conveying shape. *Proceedings of ACM SIGGRAPH*, pages 848–855, 2003. (Cited on page 92.)

[43] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. "yours is better!": Participant response bias in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1321–1330, 2012. (Cited on page 13.)

[44] Christian Dick, Joachim Georgii, Rainer Burgkart, and Rüdiger Westermann. Stress tensor field visualization for implant planning in orthopedics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1399–1406, 2009. (Cited on pages 113 and 115.)

[45] Stephan Diepenbrock, Timo Ropinski, and Klaus Hinrichs. Context-aware volume navigation. In *Pacific Visualization Symposium (PacificVis)*, pages 11–18, 2011. (Cited on page 41.)

[46] Sebastian Eichelbaum, Mario Hlawitschka, and Gerik Scheuermann. Lineao-improved three-dimensional line rendering. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):433–445, 2013. (Cited on page 54.)

[47] Niklas Elmqvist and Ji Soo Yi. Patterns for visualization evaluation. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, BELIV '12, pages 12:1–12:8, 2012. (Cited on pages 8 and 11.)

[48] Maarten H Everts, Henk Bekker, Jos BTM Roerdink, and Tobias Isenberg. Depth-dependent halos: Illustrative rendering of dense line data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1299–1306, 2009. (Cited on pages 51, 53, and 54.)

[49] Ludwig Fahrmeir, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik: Der Weg zur Datenanalyse*. Springer, 2010. (Cited on pages 10, 22, 23, 25, 26, and 28.)

[50] Gustav Theodor Fechner. *Elements of Psychophysics*. Breitkopf & Härtel, 1860. (Cited on page 7.)

[51] Chuan Feng, Jerzy W. Rozenblit, and Allan J. Hamilton. A computerized assessment to compare the impact of standard, stereoscopic, and high-definition laparoscopic monitor displays on surgical technique. *Surgical Endoscopy*, 24(11):2743–8, 2010. (Cited on page 64.)

[52] Andy Field and Graham Hole. *How to Design and Report Experiments*. SAGE Publications Ltd, 2012. (Cited on pages 12, 15, 16, 17, 19, 20, 23, 24, 26, 28, 29, and 98.)

[53] Andy P. Field. *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)*. SAGE, third edition edition, 2009. (Cited on pages 7, 11, 16, 17, 18, 23, 24, 27, 28, 29, 107, and 153.)

[54] Roland W. Fleming, Antonio Torralba, and Edward H. Adelson. Specular reflections and the perception of shape. *Journal of Vision*, 4(9):798–820, 2004. (Cited on page 56.)

[55] Robert W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5 (6):345 – 349, 1962. (Cited on page 99.)

[56] Rocco Gasteiger. *Visual Exploration of Cardiovascular Hemodynamics*. PhD thesis, Otto-von-Guericke University Magdeburg, 2014. (Cited on page 118.)

[57] Rocco Gasteiger, Mathias Neugebauer, Christoph Kubisch, and Bernhard Preim. Adapted surface visualization of cerebral aneurysms with embedded blood flow information. In *Proceedings of Eurographics Workshop on Visual Computing for Biology and Medicine (EG VCBM)*, pages 25–32, 2010. (Cited on pages 55, 60, 113, 114, 115, 116, 120, and 136.)

[58] Rocco Gasteiger, Mathias Neugebauer, Oliver Beuing, and Bernhard Preim. The flowlens: A focus-and-context visualization approach for exploration of blood flow in cerebral aneurysms. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2183–2192, 2011. (Cited on page 41.)

[59] Amy Gooch, Bruce Gooch, Peter Shirley, and Elaine Cohen. A non-photorealistic lighting model for automatic technical illustration. In *Proceedings of ACM SIGGRAPH*, pages 447–452, 1998. (Cited on page 116.)

[60] John Graham and David Baguley. *Ballantyne's deafness*. Wiley, 7th edition illustrated edition, 2009. (Cited on page 141.)

[61] Henry Gray. *Anatomy of the Human Body*. Lea & Febiger, 1918. (Cited on page 49.)

[62] David M. Green and John A. Swets. *Signal detection theory and psychophysics*. John Wiley & Sons, Inc., New York, 1966. (Cited on pages 75 and 88.)

[63] Pascal Grosset, Mathias Schott, Georges-Pierre Bonneau, and Hansen Charles. Evaluation of depth of field for depth perception in dvr. In *Proceedings of IEEE Pacific Visualization*, pages 81–88, 2013. (Cited on page 58.)

[64] Kurinchi Gurusamy, Rajesh Aggarwal, Latha Palanivelu, and Brian R. Davidson. Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *The British Journal of Surgery*, 95(9):1088–97, 2009. (Cited on page 139.)

[65] Christian Hansen, Jan Wieferich, Felix Ritter, Christian Rieder, and Heinz-Otto Peitgen. Illustrative visualization of 3d planning models for augmented reality in liver surgery. *International Journal of Computer-Assisted Radiology and Surgery*, 5(2):133–141, 2010. (Cited on pages 51 and 53.)

[66] Christopher G. Healey and James T. Enns. Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, 2012. (Cited on pages 44 and 45.)

[67] Elaine R. S. Hodges. *The Guild Handbook of Scientific Illustration*. Van Nostrand Reinhold, 1989. (Cited on page 49.)

[68] Nicolas S. Holliman, Neil A. Dodgson, Gregg E. Favalora, and Lachlan Pockett. Three-dimensional displays: A review and applications analysis. *IEEE Transaction on Broadcasting*, 57(2):362–371, 2011. (Cited on pages 60, 61, and 63.)

[69] Wei-Hsien Hsu, Zhang Yubo, and Kwan-Liu Ma. A multi-criteria approach to camera motion design for volume data animation. *IEEE Transaction on Visualization and Computer Graphics*, 19(12):2792 – 2801, 2013. (Cited on page 41.)

[70] Liqiang Huang and Harold Pashler. A boolean map theory of visual attention. *Psychological Review*, 114(3):599–631, 2007. (Cited on page 45.)

[71] Victoria Interrante, Henry Fuchs, and Stephen Pizer. Enhancing transparent skin surfaces with ridge and valley lines. In *Proceedings of IEEE Visualization*, pages 52–59, 1995. (Cited on pages 49 and 92.)

[72] Victoria Interrante, Henry Fuchs, and Stephen M. Pizer. Conveying the 3d shape of smoothly curving transparent surfaces via texture. *IEEE Transaction on Visualization and Computer Graphics*, 3(2):98–117, 1997. (Cited on pages 49, 50, 56, and 113.)

[73] Victoria Interrante, Sunghee Kim, and Haleh Hagh-Shenas. Conveying 3d shape with texture: Recent advances and experimental findings. In *Proceedings of Human Vision and Electronic Imaging VII*, volume 4662 of *SPIE Proceedings Series*, pages 197–206, 2002. (Cited on pages 49 and 91.)

[74] Victoria Interrante, Brian Ries, Jason Lindquist, Michael Kaeding, and Lee Anderson. Elucidating factors that can facilitate veridical spatial perception in immersive virtual environments. *Presence: Teleoper. Virtual Environment*, 17 (2):176–198, 2008. (Cited on page 61.)

[75] Victoria L. Interrante and Sunghee Kim. Investigating the effect of texture orientation on the perception of 3d shape. In *Human Vision and Electronic Imaging VI*, volume 4299, pages 330–339. SPIE, 2001. (Cited on page 56.)

[76] Petra Isenberg, Anthony Tang, and Sheelagh Carpendale. An exploratory study of visual information analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1217–1226, 2008. (Cited on page 12.)

[77] Tobias Isenberg, Petra Neumann, Sheelagh Carpendale, Mario Costa Sousa, and Joaquim A. Jorge. Non-photorealistic rendering in context: An observational study. In *Proceedings of Non-Photorealistic Animation and Rendering (NPAR)*, pages 115–126. ACM, 2006. (Cited on pages 49 and 91.)

[78] Cullen D. Jackson, David B. Karelitz, Sean A. Cannella, and David H. Laidlaw. The great potato search: the effects of visual context on users. In *Poster Proceedings of IEEE Visualization*, 2002. (Cited on pages 20 and 61.)

[79] Alark Joshi, Xiaoning Qian, Donald P. Dione, Ketan R. Bulsara, Christopher K. Breuer, and Albert J. Sinusas. Effective visualization of complex vascular structures using a non-parametric vessel detection method. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1603 – 1610, 2008. (Cited on page 58.)

[80] Ian C. Jourdan, Erik Dutson, Alfonso Garcia, T Vleugels, J Leroy, Didier Mutter, and Jacques Marescaux. Stereoscopic vision provides a significant advantage for precision robotic laparoscopy. *British Journal of Surgery*, 91(7): 879–885, 2004. (Cited on page 64.)

[81] Tilke Judd, Frédo Durand, and Edward Adelson. Apparent ridges for line drawing. In *Proceedings of ACM SIGGRAPH*, pages 19:1–19:7, 2007. (Cited on page 92.)

[82] Bela Julész. A theory of preattentive testure discrimination based on first-order statistics of textons. *Biological Cybernetics*, 41(2):131–138, 1981. (Cited on page 45.)

[83] Ma Jun, J Walker, Wang Chaoli, S Kuhl, and Shene Ching Kuang. An automatic guide for exploring internal flow features. *IEEE Pacific Visualization Symposium*, pages 25 – 32, 2014. (Cited on page 40.)

[84] Marta Kersten-Oertel, Sean Jy-Shyang Chen, and D. Louis Collins. An evaluation of depth enhancing perceptual cues for vascular volume visualization in neurosurgery. *IEEE Tranactions on Visualization and Computer Graphics*, 20 (3):391–403, 2014. (Cited on pages 4, 20, 58, 59, 137, and 173.)

[85] Han Suk Kim, Didem Unat, Scott B. Baden, and Jürgen P. Schulze. A new approach to interactive viewpoint selection for volume data sets. *Information Visualization*, 12(3-4):240–256, 2013. (Cited on page 40.)

[86] Sunghee Kim, Haleh Hagh-Shenas, and Victoria Interrante. Conveying shape with texture: Experimental investigations of texture's effects on shape categorization judgments. *IEEE Transaction on Visualization and Computer Graphics*, 10(4):471–483, 2004. (Cited on pages 49, 56, and 91.)

[87] Jan J. Koendrink, Andrea J. van Doorn, and Astrid M. L. Kappers. Surface perception in pictures. *Perception & Psychophysics*, 52(5):487–496, 1992. (Cited on page 48.)

[88] Michael Kolomenkin, Ilan Shimshoni, and Ayellet Tal. Demarcating curves for shape illustration. In *Proceedings of ACM SIGGRAPH Asia*, pages 157:1–157:9, 2008. (Cited on page 92.)

[89] Robert Kosara, Silve Miksch, and Helwig Hauser. Focus + context taken literally. *IEEE Computer Graphics and Applications*, 22(1):22–29, 2002. (Cited on page 73.)

[90] Robert Kosara, Silvia Miksch, Helwig Hauser, Johann Schrammel, Verena Giller, and Manfred Tscheligi. Useful properties of semantic depth of field for better f+c visualization. In *Proceedings of the Joint Eurographics (IEEE TCVG Symposium on Visualization VisSym 2002)*, pages 205–210, 2002. (Cited on pages 56, 67, and 73.)

[91] Cornelia Kranczioch, Stefan Debener, Christoph S. Herrmann, and Andreas K. Engel. Eeg gamma-band activity in rapid serial visual presentation. *Experimental Brain Research*, 169(2):246–254, November 2006. (Cited on page 78.)

[92] Arno Krüger, Christian Tietjen, Jana Hintze, Bernhard Preim, Ilka Hertel, and Gero Strauß. Interactive visualization for neck dissection planning. In *IEEE Eurographics Symposium on Visualization (EuroVis)*, pages 295–302, 2005. (Cited on page 76.)

[93] Elizabeth A. Krupinski. Visual scanning patterns of radiologists searching mammograms. *Academic Radiology*, 3(2):137–144, 1995. (Cited on page 73.)

[94] Elizabeth A. Krupinski. Medical image perception: evaluating the role of experience. In *Proceedings of SPIE, Human Vision and Electronic Imaging V 281*, pages 281–289, 2000. (Cited on page 59.)

[95] Christoph Kubisch, Christian Tietjen, and Bernhard Preim. Gpu-based smart visibility techniques for tumor surgery planning. *International Journal of Computer Assisted Radiology and Surgery*, 5(6):667–678, 2010. (Cited on page 55.)

[96] Wolfgang Kunert, Pirmin Storz, and Andreas Kirschniak. For 3d laparoscopy: a step toward advanced surgical navigation: how to get maximum benefit from 3d vision. *Surgical Endoscopy*, 27(2):696 – 699, 2013. (Cited on page 64.)

[97] Anja Kuß, Maria Gensel, Björn Meyer, Vincent J. Dercksen, and Steffen Prohaska. Effective techniques to visualize filament-surface relationships. *Computer Graphics Forum*, 29:1003–1012, 2010. (Cited on pages 59 and 60.)

[98] Bireswar Laha, Doug A. Bowman, and John J. Socha. Effects of vr system fidelity on analyzing isosurface visualization of volume datasets. *IEEE Tranactions on Visualization and Computer Graphics*, 20(04):513–522, 2014. (Cited on page 20.)

[99] David H. Laidlaw, Michael Kirby, Cullen Jackson, et al. Comparing 2D vector field visualization methods: A user study. *IEEE Tranactions on Visualization and Computer Graphics*, 11(1):59–70, 2005. (Cited on page 20.)

[100] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transaction on Visualization and Computer Graphics*, 18(9):1520–1536, 2012. (Cited on pages 8 and 13.)

[101] Kristina Lang, Sophia Zackrisson, Kenneth Holmqvist, Marcus Nystrom, Ingvar Andersson, Daniel Förnvik, Anders Tingberg, and Pontus Timberg. Optimizing viewing procedures of breast tomosynthesis image volumes using eye tracking combined with a free response human observer study. In *Proceedings of SPIE 7966, Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment*, 2011. (Cited on page 73.)

[102] Kai Lawonn. *Illustrative Visualization of Medical Data Sets*. PhD thesis, Otto-von-Guericke University Magdeburg, 2014. (Cited on page 94.)

[103] Kai Lawonn, Tobias Mönch, and Bernhard Preim. Streamlines for illustrative real-time rendering. *Computer Graphics Forum*, 32(3):321–330, 2013. (Cited on pages 51, 54, and 76.)

[104] Kai Lawonn, Alexandra Baer, Patrick Saalfeld, and Bernhard Preim. Comparative evaluation of feature line techniques for shape depiction. In *Proceedings of Vision Modeling and Visualization (VMV)*, pages 31–38, 2014. (Cited on pages 91, 92, 93, 99, 100, 101, 102, and 103.)

[105] Kai Lawonn, Patrick Saalfeld, and Bernhard Preim. Illustrative visualization of endoscopic views. In *Bildverarbeitung für die Medizin (BVM)*, pages 276–281, 2014. (Cited on pages 51, 54, and 55.)

[106] Haim Levkowitz. *Color theory and modeling for computer graphics, visualization, and multimedia applications*. Springer, 1997. (Cited on page 115.)

[107] Wilmot Li, Lincoln Ritter, Maneesh Agrawala, Brian Curless, and David Salesin. Interactive cutaway illustrations of complex 3d models. In *Proceedings of ACM SIGGRAPH*, pages 31–1–31–10, 2007. (Cited on page 60.)

[108] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932. (Cited on page 15.)

[109] Florian Lindemann and TImo Ropinski. About the influence of illumination models on image comprehension in direct volume rendering. *IEEE Tranactions on Visualization and Computer Graphics*, 17(12):1922 – 1931, 2011. (Cited on pages 4, 57, and 173.)

[110] Aidong Lu, Christopher J. Morris, David S. Ebert, Penny Rheingans, and Charles Hansen. Non-photorealistic volume rendering using stippling techniques. In *IEEE Visualization*, pages 211–218. IEEE Computer Society, 2002. (Cited on page 51.)

[111] Thomas Luft, Carsten Colditz, and Oliver Deussen. Image enhancement by unsharp masking the depth buffer. *ACM Transactions on Graphics*, 25(3): 1206–1213, 2006. (Cited on page 116.)

[112] Eric B. Lum and Kwan-Liu Ma. Hardware-accelerated parallel non-photorealistic volume rendering. In *Proceedings of the 2Nd International Symposium on Non-photorealistic Animation and Rendering*, NPAR '02, pages 67–75, 2002. (Cited on page 51.)

[113] Xun Luo, Robert V. Kenyon, Derek Kamper, Daniel J. Sandin, and Thomas A. DeFanti. The effects of scene complexity, stereovision, and motion parallax on size constancy in a virtual environment. In *Proceedings of the IEEE Virtual Reality Conference (VR '07)*, pages 59–66, 2007. (Cited on page 61.)

[114] Suzanne P. McKee and Douglas G. Taylor. The precision of binocular and monocular depth judgments in natural settings. *Journal of Vision*, 10(5):1–13, 2010. (Cited on page 60.)

[115] Gabriel Mistelbauer, Hamed Bouzari, Rüdiger Schernthaner, Ivan Baclija, Arnold Kochl, Stefan Bruckner, Milos Sramek, and Meister Eduard Gröller. Smart super views: A knowledge-assisted interface for medical visualization. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 163–172, 2012. (Cited on pages 59 and 60.)

[116] Eva Monclús, Pere-Pau Vázquez, and Isabel Navazo. Representative views and paths for volume models. In *Smart Graphics*, pages 106–117, 2008. (Cited on page 40.)

[117] Janice M. Morse. Determing sampling sizes. *Qualitative Health Research*, 10 (1):3–5, 2000. (Cited on page 20.)

[118] Albert Mudry and Mara Mills. The early history of the cochlear implant: a retrospective. *JAMA Otolaryngology– Head & Neck Surgery*, 139(5):446–453, 2013. (Cited on page 141.)

[119] Konrad Mühler. *Animationen und Explorationstechniken zur Unterstützung der chirurgischen Operationsplanung*. PhD thesis, Otto-von-Guericke University of Magdeburg, 2010. (Cited on page 40.)

[120] Konrad Mühler and Bernhard Preim. Reusable visualizations and animations for surgery planning. In *Computer Graphics Forum (EuroVis)*, pages 1103 –1112, 2010. (Cited on page 40.)

[121] Konrad Mühler, Mathias Neugebauer, Christian Tietjen, and Bernhard Preim. Viewpoint selection for intervention planning. In K. Museth, T. Möller, and A. Ynnerman, editors, *IEEE Eurographics Symposium on Visualization (EuroVis)*, pages 267–274, 2007. (Cited on page 39.)

[122] Mary Curry Narayan. Culture's effects on pain assessment and management. *The American Journal of Nursing*, 110(4):48–9, 2010. (Cited on page 13.)

[123] Devon Penney, Jian Chen, and David H. Laidlaw. Effects of illumination, texture, and motion on task performance in 3d tensor-field streamtube visualizations. In *IEEE Pacific Visualization Symposium*, pages 97–104, 2012. (Cited on pages 56 and 57.)

[124] Emil Praun, Hugues Hoppe, Matthew Webb, and Adam Finkelstein. Real-time hatching. In *Proceedings of SIGGRAPH 01*, pages 579–584, 2001. (Cited on page 54.)

[125] Bernhard Preim and Dirk Bartz. *Visualization in Medicine: Theory, Algorithms, and Applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007. (Cited on page 38.)

[126] Bernhard Preim and Charl Botha. *Visual Computing for Medicine (Theory, Algorithms, and Applications)*. Morgan Kaufmann, 2013. (Cited on pages 33, 34, 35, and 36.)

[127] Bernhard Preim and Raimund Dachselt, editors. *Interaktive Systeme - Band 2: User Interface Engineering, 3D-Interaktion, Natural User Interfaces*. Springer Vieweg Verlag, 2015. (Cited on pages 47 and 149.)

[128] Bernhard Preim, Christian Tietjen, and Christina Doerge. Npr, focussing and emphasis in medical visualizations. In *Simulation und Visualisierung*, pages 139–152, 2005. (Cited on page 76.)

[129] Rene Przkora, William McGrady, Terrie Vasilopoulos, Nikolaus Gravenstein, and Daneshvari Solanki. Evaluation of the head-mounted display for ultrasound-guided peripheral nerve blocks in simulated regional anesthesia. *Pain Medicine*, page to appear, 2015. (Cited on page 61.)

[130] Philip T. Quinlan and Glyn W. Humphreys. Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on feature and conjunction searches. *Perception & psychophysics*, 41 (5):455 – 472, 1987. (Cited on page 45.)

[131] Peter Rautek, Stefan Bruckner, and Meister Eduard Gröller. Semantic layers for illustrative volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1336–1343, 2007. (Cited on pages 41 and 55.)

[132] Peter Rautek, Stefan Bruckner, and Meister Eduard Gröller. Interaction-dependent semantics for illustrative volume rendering. *Computer Graphics Forum*, 27(3):847–854, 2008. (Cited on pages 41 and 55.)

[133] Jenny C. A. Read and Iwo Bohr. User experience while viewing stereoscopic 3d television. *Ergonomics*, 57(8):1140 – 1153, 2014. (Cited on pages 63 and 139.)

[134] Stephan Reichelt, Ralf Häussler, Gerald Fütterer, and Norbert Leister. Depth cues in human visual perception and their realization in 3d displays. In *Proceedings of SPIE, Three-Dimensional Imaging, Visualization, and Display 2010 and Display Technologies and Applications for Defense, Security, and Avionics IV*, volume 7690, pages 76900B–76900B, 2010. (Cited on pages 46 and 47.)

[135] Bernhard E. Riecke, Markus von der Heyde, and Heinrich H. Bülthoff. Visual cues can be sufficient for triggering automatic, reflex-like spatial updating. *ACM Transactions on Applied Perception*, 2(3):183–215, 2005. (Cited on page 61.)

[136] Christian Rieder, Tim Kröger, Christian Schumann, and Horst K. Hahn. Gpu-based real-time approximation of the ablation zone for radiofrequency ablation. *IEEE Transactions on Visualization and Computer Graphics*, 17(2):1812 – 1821, 2011. (Cited on pages 36 and 37.)

[137] Felix Ritter, Christian Hansen, Volker Dicken, Olaf Konrad, Bernhard Preim, and Heinz-Otto Peitgen. Real-time illustration of vascular structures. *IEEE Transaction on Visualization and Computer Graphics*, 12(5):877–884, 2006. (Cited on pages 51, 52, 53, 92, and 137.)

[138] Timo Ropinski, Frank Steinicke, and Klaus Hinrichs. Visually supporting depth perception in angiography imaging. In *Smart Graphics*, pages 93–104, 2006. (Cited on pages 58, 122, and 137.)

[139] Patrick Saalfeld, Alexandra Baer, Kai Lawonn, Uta Preim, and Bernhard Preim. Verwendung des 3d-user interfaces zspace zur exploration und inspektion von wirbeln der halswirbelsäule. In *Bildverarbeitung für die Medizin (BVM)*, pages 83–88, 2015. (Cited on pages 67, 68, and 164.)

[140] Takafumi Saito and Tokiichiro Takahashi. Comprehensible rendering of 3-d shapes. In *Proceedings of ACM SIGGRAPH 90*, volume 24(4), pages 197–206, 1990. (Cited on page 49.)

[141] Zein Salah, Douglas W. Cunningham, Wolfgang Straßer, et al. Perceptually emphasized illustrative visualization for multiple objects. Technical report, Wilhelm-Schickard-Institut, University Tübingen, 2008. (Cited on page 76.)

[142] Christof Rezk Salama, Maik Keller, and Peter Kohlmann. High-level user interfaces for transfer function design with semantics. *IEEE Transaction on Visualization and Computer Graphics*, 12(5):1021 – 1028, 2006. (Cited on page 41.)

[143] Anthony Santella and Doug DeCarlo. Visual interest and npr: An evaluation and manifesto. In *Proceedings of the 3rd International Symposium on Nonphotorealistic Animation and Rendering*, NPAR '04, pages 71–150, 2004. ISBN 1-58113-887-3. (Cited on page 49.)

[144] Christophe Schlick. A customizable reflectance model for everyday rendering. In *Fourth Eurographics Workshop on Rendering*, pages 73–83, 1993. (Cited on page 116.)

[145] Martin Schmettow. Sample size in usability studies. *Communications of the ACM*, 55(4):64–70, 2012. (Cited on page 20.)

[146] Markus Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare*, 36(2):267–303, 2011. (Cited on page 99.)

[147] Jutta Schumann, Thomas Strothotte, Andreas Raab, and Stefan Laser. Assessing the effect of non-photerealistic rendered images in cad. In *CHI Conference on Human Factors in Computing Systems*, pages 35–42, 1996. (Cited on page 49.)

[148] Susana Segura, Pablo Fernandez-Berrocal, and Ruth MJ Byrne. Temporal and causal order effects in thinking about what might have been. *The Quarterly Journal of Experimental Psychology: Section A*, 55(4):1295–1305, 2002. (Cited on page 13.)

[149] A. Elizabeth Seward, Daniel H. Ashmead, and Bobby Bodenheimer. Using virtual environments to assess time-to-contact judgments from pedestrian viewpoints. *ACM Transaction on Applied Perception*, 4(3):18/1–18/19, 2007. (Cited on page 61.)

[150] John R. Shea and John R. Emmett. Polyethylene TORPs and PORPs in otologic surgery. In *Proceedings of the First International Symposium Biomaterials in Otology*, pages 137–152, 1984. (Cited on page 141.)

[151] Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science, New Series*, 171(3972), 1971. (Cited on page 20.)

[152] Pirmin Storz, Buess, Wolfgang Kunert, and Andreas Kirschniak. 3d hd versus 2d hd: surgical task efficiency in standardised phantom tasks. *Surgical Endoscopy*, 26(5):1454–1460, 2012. (Cited on pages 64, 65, 66, and 164.)

[153] Pjotr Svetachov, Maarten H. Everts, and Tobias Isenberg. Dti in context: Illustrating brain fiber tracts in situ. *Computer Graphics Forum*, 29(3):1023–1032, 2010. (Cited on pages 51, 53, and 54.)

[154] Christian Tietjen. *Illustrative Visualisierungstechniken zur Unterstützung der präoperativen Planung von chirurgischen Eingriffen*. PhD thesis, Otto-von-Guericke University Magdeburg, 2009. (Cited on page 76.)

[155] Christian Tietjen, Tobias Isenberg, and Bernhard Preim. Combining silhouettes, surface, and volume rendering for surgery education and planning. In *IEEE/Eurographics Symposium on Visualization (EuroVis)*, pages 303–310, 2005. (Cited on pages 20, 51, 52, 76, 89, and 92.)

[156] Christian Tietjen, Roland Pfisterer, Alexandra Baer, Rocco Gasteiger, and Bernhard Preim. Hardware-acceelerated illustrative medical surface visualization with extended shading maps. In *Proceedings of Smart Graphics*, pages 166–177. Springer Verlag, 2008. (Cited on page 51.)

[157] Lames T. Todd. The visual perception of 3d shape. *Trends in Cognitive Science*, 8(3):115–121, 2004. (Cited on page 48.)

[158] Anne M. Treismann and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 10:97–136, 1980. (Cited on pages 45, 74, and 78.)

[159] John W. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977. (Cited on page 25.)

[160] Kazuhiko Ukai, , and Peter A. Howarth. Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays*, 29(2):106 – 116, 2007. (Cited on pages 63 and 139.)

[161] Rice University. Sensory pathways. Online, 2015. (Cited on page 44.)

[162] Ivan Viola and Eduard Gröller. Smart visibility in visualization. In *Proceedings of EG Workshop on Computational Aesthetics in Graphics, Visualization and Imaging*, pages 87–94, 2005. (Cited on pages 55 and 116.)

[163] Ivan Viola, Miquel Feixas, Mateu Sbert, and Meister Eduard Gr"oller. Importance-driven focus of attention. In *IEEE Transaction on Visualization and Graphics*, pages 933–940, 2006. (Cited on pages 39, 40, and 41.)

[164] Veronika Šoltészová, Daniel Patel, and Ivan Viola. Chromatic shadows for improved perception. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering*, NPAR '11, pages 105–116, New York, NY, USA, 2011. ACM. (Cited on page 59.)

[165] Veronika Šoltészová, Mark C. Price, and Ivan Viola. A perceptual-statistics shading model. *IEEE Transaction on Visualization and Computer Graphics*, 18 (12):2265–2274, Aug 2012. (Cited on pages 48 and 56.)

[166] Leonard R. Wagner, James A. Ferwerda, and Donald P. Greenberg. Perceiving spatial relationships in computer-generated images. *IEEE Transactions on Computer and Graphics and Application*, pages 44–58, 1992. (Cited on pages 60 and 139.)

[167] Oliver J. Wagner, Monika Hagen, Anita Kurmann, Simon Horgan, Daniel Candinas, and Stephan A. Vorburger. Three-dimensional vision enhances

task performance independently of the surgical method. *Surgical Endoscopy*, 26(10):2961–8, 2012. (Cited on pages 62, 66, and 164.)

[168] Manuela Waldner, Mathieu Le Muzic, Matthias Bernhard, Werner Purgathofer, and Ivan Viola. Attractive flicker: Guiding attention in dynamic narrative visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2456–2465, 2014. (Cited on pages 47 and 73.)

[169] Christian Wallraven, Martin Breidt, Douglas W. Cunningham, and Heinrich H. Bülthoff. Evaluating the perceptual realism of animated facial expressions. *ACM Transaction on Applied Perception*, 4(4):1–20, 2008. (Cited on page 60.)

[170] Jian Wang, Matthias Kreiser, Lejing Wang, Nassir Navab, and Pascal Fallavollita. Augmented depth perception visualization in 2d/3d image fusion. *Computerized Medical Imaging and Graphics*, 38(8):744–752, 2014. (Cited on page 58.)

[171] Colin Ware, Kevin Arthur, and Kellogg S. Booth. Fish tank virtual reality. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, pages 37–42, 1993. (Cited on page 149.)

[172] Collin Ware. *Information Visualization, Third Edition: Perception for Design (Interactive Technologies)*. Morgan Kaufmann, 2012. (Cited on pages 33 and 45.)

[173] Chris Weigle and David C. Banks. A comparison of the perceptual benefits of linear perspective and physically-based illumination for display of dense 3d streamtubes. *IEEE Transaction on Visualization and Computer Graphics*, 14 (6):1723–1730, 2008. (Cited on pages 56 and 57.)

[174] Irving B. Weiner, Donald K. Freedheim, and John R. Graham. *Handbook of Psychology, Assessment of Psychology*. John Wiley & Sons, Inc., 2012. (Cited on pages 13 and 18.)

[175] Michael Wendt, Frank Sauer, Ali Khamene, Benedicte Bascle, Sebastian Vogt, and Frank K Wacker. A head-mounted display system for augmented reality: initial evaluation for interventional mri. *RoFo: Fortschritte auf dem Gebiete der Röntgenstrahlen und der Nuklearmedizin*, 175(3):418–421, 2003. (Cited on page 61.)

[176] Dirk Wilhelm, Silvano Reiserand Nils Kohn, Michael Witte, Ulrich Leiner, Lothar Mühlbach, Detlef Ruschin, Wolfgang Reiner, and Hubertus Feussner. Comparative evaluation of hd 2d/3d laparoscopic monitors and benchmarking to a theoretically ideal 3d pseudodisplay: even well-experienced laparoscopists perform better with 3d. *Surgical Endoscopy*, 28(8):2387–2397, 2014. (Cited on pages 64, 65, 66, 163, and 164.)

[177] Christian F. Wittekind, Leslie H. Sobin, and Martin Klimpfinger. *TNM-Atlas*. Springer Berlin Heidelberg, 2005. (Cited on page 74.)

[178] Jeremy M. Wolfe, Kyle R. Cave, and Susan L. Franzel. Guided search: an alternative to the modified feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15:419–433, 1989. (Cited on page 45.)

[179] Xuexiang Xie, Ying He, Feng Tian, Hock-Soon Seah, Xianfeng Gu, and Hong Qin. An effective illustrative visualization framework based on photic extremum lines (pels). *IEEE Transactions on Visualization and Computer Graphics*, 13:1328–1335, 2007. (Cited on page 92.)

[180] Soichiro Yoshida, Kazunori Kihara, Hideki Takeshita, and Yasuhisa Fujii. A head-mounted display-based personal integrated-image monitoring system for transurethral resection of the prostate. *Videosurgery Miniinv*, 9:644–9, 2014. (Cited on page 61.)

[181] Stefan Zachow, Philipp Muigg, Thomas Hildebrandt, et al. Visual exploration of nasal airflow. *IEEE Transaction on Visualization and Computer Graphics*, 15(6):1407–1414, 2009. (Cited on pages 113 and 115.)

[182] Johannes Zander, Tobias Isenberg, Stefan Schlechtweg, and Thomas Strothotte. High quality hatching. *Computer Graphics Forum*, 23(3):421–430, 2004. (Cited on page 54.)

[183] Long Zhang, Ying He, Jiazhi Xia, Xuexiang Xie, and Wei Chen. Real-time shape illustration using laplacian lines. *IEEE Transactions on Visualization and Computer Graphics*, 17:993–1006, 2011. (Cited on pages 92 and 103.)

PUBLICATIONS

Journal papers

1. **A. Baer**, R. Gasteiger, D. Cunningham and B. Preim. "Perceptual Evaluation of Ghosted View Techniques for the Exploration of Vascular Structures and Embedded Flow", In *Computer Graphics Forum (EuroVis)*, 30(3): 811-820, 2011

Conference papers

1. **A. Baer**, K. Lawonn, P. Saalfeld and B. Preim. "Statistical Analysis of a Qualitative Evaluation on Feature Lines", In *Bildverarbeitung für die Medizin (BVM)*, pages 71-76, 2015
2. P. Saalfeld, **A. Baer**, K. Lawonn and B. Preim. "Verwendung des 3D-User Interfaces zSpace zur Exploration und Inspektion von Wirbeln der Halswirbelsäule", In *Bildverarbeitung für die Medizin (BVM)*, pages 83-88, 2015
3. **A. Baer**, A. Hübler, P. Saalfeld, D. Cunningham and B. Preim. "A Comparative User Study of a 2D and an Autostereoscopic 3D Display for a Tympanoplastic Surgery", In *EG Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM)*, pages 181-190, 2014
4. K. Lawonn, **A. Baer**, P. Saalfeld and B. Preim. "Comparative Evaluation of Feature Line Techniques for Shape Depiction", In *Proceedings of Vision Modeling and Visualization (VMV)*, pages 31-38, 2014
5. **A. Baer**, K. Kellermann and B. Preim. "Importance-Driven Structure Categorization for 3D Surgery Planning", In *Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM)*, pages 99-107, 2010
6. K. Kellermann, **A. Baer** and B. Preim. "Adaptive Fokus-Kontext-Kategorisierung für Visualisierungen zur Operationsplanung", In *Proceedings of Bildverarbeitung für die Medizin (BVM)*, pages 167-171, 2010
7. **A. Baer**, F. Adler, D. Lenz and B. Preim. "Perception-Based Evaluation of Emphasis Techniques Used in 3D Medical Visualization", In *Proceedings of Vision Modeling and Visualization (VMV)*, pages 295-304, 2009
8. C. Tietjen, R. Pfisterer, **A. Baer**, R. Gasteiger and B. Preim. "Hardware-Accelerated Illustrative Medical Surface Visualization with Extended Shading Maps", In *Proceedings of SmartGraphics*, pages 166-177, 2008
9. R. Gasteiger, C. Tietjen, **A. Baer** and B. Preim. "Curvature- and Model-Based Surface Hatching of Anatomical Structures Derived from Clinical Volume Datasets", In *Proceedings of SmartGraphics*, pages 255-262, 2008
10. C. Tietjen, R. Gasteiger, **A. Baer** and B. Preim. "Curvature- and Model-Based Surface Hatching of Patient-Specific Muscle Surfaces", In *Proceedings of Bildverarbeitung für die Medizin (BVM)*, pages 117-121, 2008

11. **A. Baer**, C. Tietjen, R. Bade and B. Preim. "Hardware-Accelerated Stippling of Surfaces Derived from Medical Volume Data", In *IEEE/Eurographics Symposium on Visualization (EuroVis)*, pages 235-242, 2007

12. **A. Baer**, C. Tietjen, M. Spindler and B. Preim. "Hardwaregestütztes Stippling von medizinischen Oberflächenmodellen", In *Proceedings of Bildverarbeitung für die Medizin (BVM)*, pages 266-270, 2006

13. C. Janke, C. Tietjen, **A. Baer**, C. Zwick, B. Preim, I. Hertel and Gero Strauß. "Design und Realisierung eines Softwareassistenten zur Planung von Halsoperationen", In *Proceedings of Mensch und Computer*, pages 373-378, 2006