



Software Phantoms in Medical Image Analysis

DISSERTATION

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

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

von Dipl. Inf. Jan Rexilius

geb. am 17.03.1975 in Herdecke

Gutachterinnen/Gutachter

Prof. Dr. Klaus-Dietz Tönnies

Prof. Dr. Simon K. Warfield

Prof. Dr. Regina Pohle-Fröhlich

Magdeburg, den 16.03.2015

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Analyse von Softwarephantomen für die medizinische Bildverarbeitung. Dazu werden neue Verfahren in zwei unterschiedlichen Aufgabengebieten vorgestellt: (1) Die Entwicklung von Phantomen und deren Anwendung sowie (2) die Validierung von Phantomen.

Ein wichtiger erster Schritt für die Entwicklung von Softwarephantomen ist die Betrachtung der Phantomart und der zugrundeliegenden Anwendung. Die ersten Kapitel dieser Arbeit geben daher zunächst einmal einen Überblick über Designmethoden. Darüber hinaus werden Parameter untersucht, die man häufig für die Phantomerstellung verwendet. Basierend auf den Ergebnissen dieser Betrachtung erfolgt dann die Entwicklung eines modularen Designs für Phantome. Unser Ziel ist eine Art Baukasten, der eine formalisierte Beschreibung von Parametern nutzt. Dies ermöglicht den Austausch einzelner Parameter durch neue Modelle.

Basierend auf unserer Designmethode, entwickeln wir im nächsten Schritt Phantome für unterschiedliche Anwendungen. Besonders die Entwicklung von Multiple Sklerose (MS) Läsionsphantomen sowie von Gehirntumorphantomen stehen im Vordergrund. Dabei werden verschiedene Aspekte untersucht, wie etwa das Objektvolumen. Darüber hinaus wird eine makroskopische Simulation von Wachstumsprozessen basierend auf einem physikalisch motivierten, elastischen Modell, vorgeschlagen, das bereits für die Erfassung von Formveränderungen während neurochirurgischer Eingriffe verwendet wurde. Da für die entwickelten Phantome ein hoher manueller Aufwand notwendig ist, ist die Zahl der verfügbaren Phantomdatensätze für eine Anwendung typischerweise sehr klein. In einem weiteren Schritt schlagen wir daher eine vollautomatisierte Methode vor, mit der eine große Zahl an Phantomen erstellt werden kann. Die zugehörigen Parametermodelle werden dabei aus Patientendaten abgeleitet. Mit Hilfe unseres Baukastenprinzips haben wir zusätzlich einen interaktiven Software-Assistenten für die Erstellung von Phantomen implementiert, der den Entwicklungsaufwand für ein Phantom stark reduziert. Eine Reihe von Phantomen aus dieser Arbeit wurden mit dem Assistenten entwickelt.

Unsere Phantome bieten eine gute Grundlage für die Evaluierung von Algorithmen, die häufig in der klinischen Routine oder in Studien auftreten. Zum Beispiel analysieren wir die Qualität einer rein visuellen Bewertung von Läsionen im Rahmen einer Studie mit mehr als 20 Teilnehmern. Weiterhin werden neue Ansätze für eine genaue und reproduzierbare

Volumetrie von Läsionen untersucht. Der Fokus liegt dabei vor allem auf kleinen Läsionen, wie sie beispielsweise bei Patienten mit MS auftreten. Für die Evaluierung werden mehr als 50 Datensätze mit Läsionsphantomen unterschiedlicher Volumina und Formen erzeugt und manuell von mehreren neuroradiologischen Experten ausgewertet. Darüber hinaus nutzen wir unsere Phantome für die Bewertung von Segmentierungsalgorithmen.

Der Schwerpunkt im zweiten Teil der Arbeit liegt auf der Analyse der Phantomqualität. Auf Grund der bekannten Ground Truth für alle modellierten Parameter sind Phantome heute ein weit verbreitetes Werkzeug. Trotzdem gibt es praktisch keine Methoden, die eine Bewertung dieser Phantome ermöglichen. Unser erster Schritt in diese Richtung ist daher zunächst eine Untersuchung von Methoden verwandter Gebiete im Bereich der medizinischen Bildverarbeitung. Danach wird schließlich ein neues Verfahren für die Validierung von Phantomen vorgestellt, das Verbindungen mit dem Designansatz aus dem ersten Teil dieser Arbeit aufweist. Unsere Methode ermöglicht sowohl eine standardisierte Analyse einzelner Phantome als auch den Vergleich einer Menge von Phantomen. Darüber hinaus wird mit dem sogenannten Analytischen Hierarchieprozess eine Methode verwendet, die eine Bestimmung des Eignungsgrades für alle verwendeten Parameter ermöglicht.

Bis ein Phantom eine akzeptable Konfidenz für eine bestimmte Anwendung aufweist sind viele Schritte notwendig. Diesen Aspekt bilden wir auch auf die Phantomvalidierung ab und schlagen einen iterativen Prozess vor, der aus vielen einzelnen Validierungsmethoden besteht. Den übergeordneten Prozess teilen wir in drei Schritte ein: *Methodenauswahl*, *Methodenvalidierung* und *Phantomvalidierung*. Jede betrachtete Methode vergrößert bzw. verkleinert die Konfidenz in ein Phantom. Dazu müssen zwei Merkmale bewertet werden: die Eignung und die Korrektheit einer Methode. Für die Methodenvalidierung schlagen wir eine Funktion vor, die beide Merkmale miteinander kombiniert. Eine multiplikative Fusion der Ergebnisse aller Methoden liefert schließlich die finale Phantombewertung. Die Qualität eines Phantoms wird mit diesem Ansatz also durch eine Zahl beschrieben.

Für eine konkrete Bewertung der vorgestellten Phantomvalidierung werden die MS Läsionsphantome aus dem ersten Teil dieser Arbeit verwendet. Fünf verschiedene Validierungsmethoden werden dazu untersucht und kombiniert. Anhand einer Studie mit vier Experten bewerten wir beispielsweise die Detektionsleistung von Läsionsphantomen in Patientendaten mit echten Läsionen. Eine weitere Methode nutzt den Vergleich von Segmentierungsergebnissen zwischen Phantomen und Patientendaten. Die Analyse jeder Methode besteht dabei aus einer kurzen Beschreibung der zugrunde liegenden Daten und einer Diskussion der Ergebnisse.

Summary

The validation of algorithms is an important aspect in medical image analysis. Today, a validation study serves as a major decision factor of a method's value to the user. Thereby, phantoms have become a widely accepted tool. This thesis deals with the analysis of software phantoms in medical image processing. New approaches are proposed in two areas: (1) Phantom development and applications and (2) phantom validation.

The first step towards developing a phantom is to decide what kind of phantom is required for the targeted application. Therefore, we start our investigation with a general overview of design approaches and examine parameters used in phantom development. Based on the knowledge of the underlying application domain and the phantom type, we then focus on the actual design. Our goal is a modular phantom design based on a set of components. To this end, we formalize the overall development process and propose models for major parameters used in medical image analysis such as object position, shape, or intensity values. After modeling all relevant parameters for an application, we combine them to our final phantom.

Several phantoms are developed using our new design approach. We focus on applications in the field of neurology and neurosurgery, including lesion phantoms that appear in patients with Multiple Sclerosis (MS) as well as brain tumor phantoms. Several characteristic properties are investigated such as the object volume or the selected intensity values. A macroscopic simulation of growth-induced deformations based on a linear elastic model is proposed, which was previously used to capture shape changes of the brain during neurosurgery. Instead of developing only a small amount of hand-crafted phantoms, we also present a fully automatic method to build an arbitrary number of MS phantoms. Thereby, our parameter models such as the lesion shape and the lesion position are based on actual patient data.

Based on the modular design process, we propose a new interactive software assistant for the development of software phantoms. Several phantoms developed in this work are generated with the software assistant, reducing the time to build a phantom from hours to a few minutes. The developed software phantoms provide an excellent basis for the evaluation of typical algorithms that frequently occur in clinical practice and trials. We investigate the quality of visual assessment in MS lesion volumetry based on a study with more than 20 participants. Furthermore, new methods for an accurate and reproducible volumetry are

presented. Here, the focus is on small lesions that appear in patients with MS. More than 50 software phantoms with different lesion shapes and volumes are generated, and both manual and semi-automatic approaches are analyzed. Finally, our phantoms are used for the evaluation of segmentation methods. Several algorithms ranging from manual to fully-automatic are analyzed.

In the second part of this work, we focus on the analysis of the phantom quality. Phantoms are a widely used tool, since the ground truth is known for all modeled parameters. However, algorithms for phantom validation are widely unknown. Therefore, our first step towards an analysis of the phantom quality consists of a survey of work in related fields of medical image analysis. We then propose an approach that is closely related to the phantom design process. It enables the analysis of a single phantom as well as the comparison of phantoms. A multi-criteria decision making technique is applied to evaluate the suitability of the modeled parameters.

Several steps are required to reach a reasonable confidence in a phantom. Therefore, we define phantom validation as an iterative process that comprises several validation methods. Three steps are distinguished: *method selection*, *method validation*, and *phantom validation*. Each method increases or decreases the confidence in the phantom. Thereby, an analysis of the method's suitability and correctness is performed. A validation function is introduced that fuses these two features.

A multiplicative combination of all validation methods is used to compute the final phantom validation. In other words, the phantom quality is expressed by a single value.

To evaluate our phantom validation approach, we use the MS lesion phantoms proposed in the first part of this thesis. Different validation methods are analyzed such as a user study to assess the detection performance of lesion phantoms in patient data sets with several real MS lesions. We also propose a comparison of segmentation overlap measures between phantoms and patient data using the results obtained in the first part of this thesis. For each validation method, a short description of the underlying data and a discussion of the results is given.

Acknowledgments

When I first met Prof. Klaus Tönnies, he told me that his job is to help me out when I don't know how to move forward. That was not the last time he impressed me and fortunately he took the job as my thesis supervisor. Thank you so much for your patience and your guidance.

My gratitude also to Prof. Simon K. Warfield for guiding me into the field of medical image analysis already many years ago. Thank you for many inspiring discussions and for sharing with me your unbelievable knowledge about any topic in image processing.

Special thanks also to my third reviewer, Prof. Regina Pohle-Fröhlich, for agreeing to review this thesis on such short notice.

Many people have patiently read and re-read my drafts of this work and provided valuable suggestions including Tobias Böhrer, Stephan Heigl and many others. Especially, I would like to thank Matthias König for his pragmatism, encouragement, many ideas, and for a nudge in the right direction whenever needed. Many thanks also to Richard Rascher-Friesenhausen for his input and brilliant questions in any stage of my work. Holger Bourquain for excellent help in any medical question I had, and for countless evaluation sessions.

Finally, I would like to thank those people who always supported me in my personal life. My parents for endless support and all kinds of help. My sister Wibke for countless telephone support in every life situation.

A PhD thesis is rarely complete without at least one citation. Therefore, special thanks to my niece Thea for proposing such a great one for this thesis.

I could not have done all this work without the support from Caro. Your patience, optimism, humor, love, and encouragement have made all this possible.

Superkalifragilistikexpialigetisch
Mary Poppins

Contents

1. Introduction	1
1.1. Objectives	2
1.2. Contributions and Publications	3
1.3. Thesis Structure	4
2. Related Work	7
2.1. Phantoms – Principles and Definitions	7
2.1.1. General Requirements	8
2.1.2. The 'Perfect' Phantom	9
2.2. A Categorization Scheme for Phantoms	10
2.2.1. Physical vs. Software	11
2.2.2. Static vs. Dynamic	13
2.2.3. Artificial vs. Realistic	14
2.3. Discussion	14
I. Phantom Development	17
3. Parameters in Phantom Development	19
3.1. Morphological and Topological Parameters	19
3.1.1. Parameter Description	20
3.1.2. Phantoms	21
3.2. Imaging Parameters and Artifacts	24
3.2.1. Parameter Description	25
3.2.2. Phantoms	28
3.3. Other Parameters	32
3.3.1. Parameter Description	32
3.3.2. Phantoms	33
3.4. Discussion	35
4. Design and Construction of Software Phantoms	41

4.1. Modular Phantom Design	41
4.1.1. Modules	42
4.1.2. A General Phantom Description	44
4.2. Parameter Modeling	48
4.2.1. Morphology and Topology	48
4.2.2. Imaging Parameters	54
4.2.3. Background Design	56
4.2.4. Object Incorporation	56
4.3. A Software Assistant for Interactive Parameter Modeling	58
4.3.1. Design Concepts	58
4.3.2. Processing Steps	58
4.4. Discussion	63
4.4.1. Modular Phantom Development	63
4.4.2. Our Phantom Approach	64
5. Phantom Examples	67
5.1. Interactive Design of MS Lesion Phantoms	67
5.1.1. Morphology and Topology	69
5.1.2. Imaging Parameters	70
5.1.3. Background Design and Object Incorporation	70
5.1.4. Resulting Software Phantoms	71
5.2. Automatic Design of MS Lesion Phantoms	72
5.2.1. Morphology and Topology	73
5.2.2. Imaging Parameters	76
5.2.3. Background Design and Object Incorporation	77
5.2.4. Resulting Software Phantoms	78
5.3. Brain Tumor Phantoms	79
5.3.1. Overall Phantom Design	80
5.3.2. Modeling of Edema	80
5.3.3. Modeling of Deformations Induced by Tumor Growth	82
5.3.4. Resulting Software Phantoms	84
5.3.5. Simulation of Contrast Enhancement Characteristics	85
5.4. Discussion	86
6. Applications	89
6.1. Visual Assessment in MS Lesion Volumetry	89
6.1.1. The Software	90
6.1.2. Results	92
6.1.3. Discussion	95
6.2. Accuracy in MS Lesion Volumetry	95
6.2.1. Software Phantoms for the Evaluation of MS Lesion Quantification	96

6.2.2. Setup for Manual Expert Analysis	97
6.2.3. Semi-Automatic Volumetry	97
6.2.4. Results	99
6.2.5. Discussion	101
6.3. Segmentation of MS Lesions	102
6.3.1. Manual and Semi-Automated Segmentation	103
6.3.2. Automated Multi-Spectral Segmentation	104
6.3.3. Discussion	106
6.4. Segmentation and Quantification of Brain Tumors	106
6.4.1. Results	108
6.4.2. Discussion	110
7. Conclusion Part I	111
II. Phantom Validation	115
8. Validation in Medical Image Analysis	117
8.1. Principles and Definitions	117
8.2. Phantom-Based Validation	118
8.3. Validation without Phantoms	119
8.3.1. Expert Validation	119
8.3.2. Databases	120
8.4. Gold Standards	121
8.5. General Approaches	122
8.6. Applications	124
8.6.1. Segmentation	124
8.6.2. Quantification	128
8.7. Discussion	129
9. Validation of Phantoms	131
9.1. General Approach	131
9.2. Processing Steps	133
9.3. Iterative Phantom Validation	136
9.3.1. Fusion of Parameter Validations	137
9.3.2. Number Of Parameters	140
9.3.3. Lesion Detection Performance	141
9.3.4. Segmentation Overlap Measures: Phantom vs. Patient Data	141
9.3.5. Effect of Parameter Changes	141
9.4. Discussion	142
10. Validation of MS Lesion Phantoms	145

10.1. Fusion of Parameter Validations	145
10.2. Number of Parameters	147
10.3. Lesion Detection Performance	147
10.4. Segmentation Overlap Measures: Phantom vs. Patient Data	149
10.5. Effect of Parameter Changes	150
10.6. Result of Phantom Validation	151
10.7. Discussion	152
11. Conclusion Part II	155
Bibliography	161

1. Introduction

Medical image processing has become a powerful tool for analyzing normal and pathological processes in the human body. With the growing amount of data per case acquired by medical imaging devices, the variety and complexity of post-processing algorithms are increasing. Today, image segmentation and quantification of anatomical structures are integral parts of virtually any image analysis system for basic research and for clinical applications in surgical planning and therapy monitoring. Therefore, each algorithm has to be evaluated carefully with respect to advantages and drawbacks.

Unfortunately, each quantitative analysis is subject to errors or uncertainties. There are errors related to the acquisition process such as a limited spatial resolution or motion artifacts. Other errors are related to anatomical and pathological variability. Therefore, accuracy and reproducibility of a proposed method are fundamental issues. To reach clinical acceptance, and to understand the intrinsic characteristics and behavior of a method, dedicated validation strategies are required.

Patient data can provide a realistic evaluation basis for novel methods. However, a large data pool is often required to account for major anatomical variations and pathologies. Furthermore, the lack of common reference data sets with an exactly known ground truth for the underlying application makes validation and comparison a difficult task. Today, researchers usually have their own small pool of patient data sets for the evaluation of new image analysis techniques.

Publicly available databases have been proposed in different fields of medical image analysis such as computer aided detection of calcifications in mammograms. Today, several workshops at major conferences use a set of patient data with manual expert segmentations for an analysis of new algorithms e.g., the 'Grand Challenge' workshops at the MICCAI conference. Two major requirements for a validation tool are accomplished: A number of data sets for the development, training, and evaluation of new algorithms, plus a common basis for the comparison of performance between different approaches. One of the most popular databases in this area has been introduced by Karssemeijer (1993), containing 40 digitized film-screen mammograms of different clinically relevant cases with associated ground truth. Another, much larger, public database is the Digital Database for Screening Mammography (DDSM), maintained by the University of Florida. The DDSM database contains approximately 2.500 studies (Heath et al. 1998). In the field of lung cancer de-

tection, a cooperative effort known as the Lung Image Database Consortium (LIDC) was launched in year 2000 to construct a database that contains CT scans from both diagnostic and screening studies (Armato et al. 2004). Various information is stored for each data set in addition to the actual image data including technical scan parameters, patient information, and nodule features (McNitt-Gray et al. 2007). The assessment of lesion boundaries is based on manual outlining performed by expert radiologists.

Although such image databases provide a common source of clinically relevant images for researchers around the world, a quantitative analysis is rarely possible due to the lack of a ground truth. The exact object parameters such as boundaries and tissue quantities within a voxel are simply unknown, and a manual object labeling suffers from intra- and inter-observer variability. Therefore, phantom data are a useful tool for developing, training, and evaluating new segmentation and quantification methods.

Early approaches to simulate human anatomy in medicine were developed in the 1960's for dosimetry calculations in radiography and radiotherapy (Fisher and Snyder 1966). Today, new algorithms are often accompanied by a phantom-based evaluation study. One of the most popular digital phantoms in medical image analysis is the freely available brain phantom from the BrainWeb project (Collins et al. 1998). More than 100 groups have already used the associated data sets for various image analysis and evaluation tasks (Aubert-Broche et al. 2006). Unfortunately, there are only few other publicly available software phantoms, and a common repository or database has not yet been established. One reason might be the large manual effort required during phantom development. Fully automatic approaches could prove helpful here. Furthermore, it is important to understand and to evaluate intrinsic characteristics and behavior of the data sets, used as a surrogate ground truth. In other words, we need a structured decision process that can be documented and replicated. Today, most phantoms are accompanied by at least some kind of evaluation. However, a proper approach beyond an expert-based rating is still missing. Formalizing the overall development process is a first step towards this goal.

1.1. Objectives

A frequent argument against the use of phantoms is their lack of complexity and structures encountered under clinical conditions. A phantom typically has only a reduced set of modeled parameters, putting in question the reliability of drawn conclusions as well as the overall data quality. Two main aspects are analyzed in this work: Phantom development and phantom validation. We address the following questions:

Phantom Development

- *How to develop phantoms for medical image analysis?*
- *What are the benefits of our phantoms?*

Phantom Validation

- *Why is an analysis of the phantom quality required?*
- *What are the benefits of our phantom validation?*

1.2. Contributions and Publications

This work contains several novel contributions. Our overall aims are the development of phantoms, applications associated with these phantoms, and an assessment of the phantom quality. We focus on two clinical applications in the field of neuroimaging: Lesions that appear in patients with Multiple Sclerosis (MS) and the analysis of brain tumors. Magnetic resonance imaging (MRI) is used as underlying imaging modality. In detail, the following novelties have been presented:

- A modular approach for the development of software phantoms, which provides a formalization of the phantom design process. A general overview has been published in (Rexilius and Tönnies 2014a).
- Software phantoms for Multiple Sclerosis based on an overall manual design have been developed in (Rexilius et al. 2003; Rexilius et al. 2005). An extension using a fully automatic method to build an arbitrary number of MS data sets was later published in (Rexilius and Tönnies 2014b).
- A software phantom for brain tumors has been published in (Rexilius et al. 2004). Tumor growth is simulated based on a linear elastic model. Initially, we developed this model to capture shape changes of the brain during neurosurgery (Rexilius et al. 2001; Rexilius et al. 2002). An overview of our approach together with other methods can be found in (Warfield et al. 2002). Furthermore, our non-linear registration method has been used for a multi-modal analysis (Verhey et al. 2002; Verhey et al. 2005).
- To provide an easy-to-use tool for manual phantom design, we developed an interactive software assistant. The results have been published in (Rexilius et al. 2008).
- The implemented MS lesion software phantoms have been used for several applications. In (Rexilius et al. 2003; Rexilius et al. 2005), we used them to evaluate lesion volumetry algorithms including manual and semi-automatic methods. A phantom-based evaluation of a well-known segmentation method has been published in (Rexilius and Tönnies 2014b).
- Similar to MS lesion phantoms, we also used our brain tumor phantoms for the evaluation of a segmentation algorithm. Our current brain tumor phantom consists of

only a single sequence. A multispectral extension will enable the analysis of our own segmentation algorithm (Rexilius et al. 2007), which is part of future work.

- Today, only few approaches are used for phantom validation in medical imaging. We propose an iterative approach based on an evaluation of different validation methods. Our long term goal is a standardization of phantom validation. Therefore, the formalization of the validation process proposed in this work is an important step. Initial results of our work have been published in (Rexilius and Tönnies 2014b). The iterative validation method has been presented in (Rexilius and Tönnies 2014a).

1.3. Thesis Structure

The rest of this thesis is organized as follows:

Chapter 2. Chapter 2 provides a general overview of phantom design approaches and proposes a new classification scheme. Furthermore, we introduce design requirements, which each phantom should reflect to a certain extent. Since our work focuses on software phantoms, we propose an additional categorization for this phantom type into stylized phantoms, voxel phantoms, and hybrid phantoms.

Chapter 3. In this chapter, we perform a detailed examination of parameters used in phantom development. Furthermore, we propose a classification into different groups, each modeling a certain aspect of a phantom.

Chapter 4. In this chapter, we introduce a modular design approach that is suited to describe any phantom. Based on a formalized description for each parameter, we propose several methods for parameter modeling. Furthermore, an easy-to-use software assistant is presented that enables an interactive design of software phantoms.

Chapter 5. After discussing how to design a phantom and the underlying parameters, we introduce three example phantoms. For each of them, the template sheet introduced previously is used to provide a brief overview. Two phantoms are based on a manual object design, targeting Multiple Sclerosis lesions as well as brain tumors. Object deformations affecting the background data are modeled to describe tumor growth based on a linear elastic model. The third example extends the above methods to a fully automatic approach. A statistical map of object positions and an automatic selection of other parameters such as shape or volume enable the generation of an arbitrary number of phantom data sets.

Chapter 6. In Chapter 6, the phantoms developed so far are used for the evaluation of algorithms that frequently occur in clinical practice and trials. First, we analyze the quality of visual assessment in MS lesion volumetry. Further applications are the evaluation of current algorithms for lesion volumetry as well as an analysis of segmentation methods.

- Chapter 7. This chapter concludes the first part of our work. We discuss the main objectives related to phantom development and recapitulate our contributions. This establishes the basis for an in-depth analysis of the phantom quality in Part II of this work.
- Chapter 8. Chapter 8 gives an overview of validation approaches used in medical image analysis and summarizes important characteristics. Furthermore, we discuss current validation strategies, which can be applied to phantoms.
- Chapter 9. This chapter is dedicated to the validation of phantoms. In the first part, we recapitulate major challenges of phantom validation and identify a link to the overall requirements for phantom development already addressed in Chapter 2. To assess the phantom quality, we then propose a novel iterative validation approach. Each iteration consists of a separate validation approach, that increases or decreases the confidence in the phantom.
- Chapter 10. In Chapter 10, we apply the proposed phantom validation approach to validate the MS lesion phantoms introduced in Chapter 5.
- Chapter 11. Finally, Chapter 11 concludes this thesis. We reflect the main challenges of phantom validation and our contributions in this field. The thesis is completed with a discussion of future work.

2. Related Work

What is a phantom? In this chapter, we provide a general overview of today's phantom design approaches and outline early and more recent work. A principal definition of a phantom and general requirements are given. Furthermore, we propose a categorization into different layers that abstract the core functionality from the actual application: the phantom type, the simulation approach, and the overall phantom appearance. Since our own work largely focuses on software phantoms, we present a comprehensive survey on software phantom design. A common classification scheme into stylized and voxel phantoms as well as the combination of both, so-called hybrid software phantoms, is considered in this work.

2.1. Phantoms – Principles and Definitions

In manufacturing and engineering, quality control (QC) or quality assurance (QA) have become an important requirement to meet customer requirements. When a new MR scanner is installed or maintained within a clinical environment, extensive calibration and testing are necessary to ensure that the system is operating within the demanded specifications (Chen et al. 2004). Thereby, QA is largely concerned with image quality using specifically designed phantoms, i.e., test objects with a known ground truth with respect to different imaging parameters (McRobbie et al. 2003). Such phantoms gain even more importance for testing and calibration when data is collected and analyzed by different scanners and at different institutes, e.g., in multicenter trials (Fu et al. 2006).

Definition: Phantom

An artificial object with known imaging parameters and properties used to test certain aspects of an imaging device or analysis method.

Besides QA of scanner parameters, the usage of phantoms is also an integral tool during the design, implementation, and utilization of new image analysis methods. Simulations of different image acquisition parameters, such as the slice thickness, provide a standardized way to generate data with known ground truth. Realistic simulations of anatomical structures further allow a dedicated performance characterization and evaluation of segmentation

and registration methods. Even education and training of medical students and professionals in medical disciplines involving anatomy and radiology are now common applications for simulation environments (Hoehne et al. 2003). Furthermore, surgical training with different normal and pathological cases can benefit from phantoms based on anatomical models (Petersik et al. 2003).

Unfortunately, not all effects of an MR system or of an examined pathology can be simulated with a phantom in a sufficient way; simplifications of the geometry or of the acquisition parameters are inevitable. Human control subjects, on the other hand, yield completely realistic images and could thus provide a basis for parameter evaluation. However, a known ground truth is usually not available and reference measurements are computed by software or by human observer, e.g., manual segmentations by a radiologist. Furthermore, large multispectral image acquisition protocols can be tedious for a subject, and there may be ethical issues concerning (repeated) injection of contrast agent. Healthy and pathological structures and organs also show high variability with respect to appearance, size, or shape, complicating the generation of established and representative data sets. Therefore, a method to generate different phantoms with a known ground truth constitute a general tool for an initial analysis and a quantitative validation of both new image analysis methods and scanner parameters.

2.1.1. General Requirements

Besides the different tasks a phantom should be able to perform, the actual design process raises several questions. How can we build a phantom? How can we measure the quality of a phantom? And what is a good phantom anyway? General requirements for phantoms are suitability, flexibility, and correctness (cf. Fig. 2.1).

Suitability. In our definition we already highlighted a known ground truth for a phantom as a central aspect. Thus, the first requirement is a preferably complete modeling of *all* features of real image data. Unfortunately, it is not possible to model all features of an image, because some of them might not even be known or very complex to model. A reasonable compromise could be to model only aspects, relevant for the application in mind. For example, analyzing a lesion segmentation based on global thresholding will not require a phantom with exact positions for each lesion.

An alternative strategy could be the utilization of post-mortem images of histological sections such as cryosection images of representative male and female cadavers (Spitzer and Whitlock 1998), applied to the confirmation of in-vivo data. However, various nonlinear deformation artifacts such as shrinkage, tearing, or partial loss of tissue are introduced within the tissue, and suitable correction methods must be applied. Moreover, physical changes within the observed tissue may result in modifications of magnetic characteristics and thus in different tissue intensity contrast.

Correctness. After deciding on the required parameters of a phantom, each parameter should be generated from a correct model and in an appropriate parameter distribution.

Property	Description
Suitability	All important parameters should be modeled, i.e., with respect to clinical data.
Correctness	Parameters should be modeled as good as possible.
Flexibility	Ability to generate phantoms of arbitrary shapes, sizes, and structures.

Figure 2.1. General design properties of a phantom.

Anatomically detailed properties allow for a plausible design with respect to examined structures of the human body. Especially parameters such as a known size and volume are an important step towards representative data sets for a quantitative analysis.

However, characterizing the quality of a parameter or even of the whole phantom is difficult and heavily depends on the considered task. While phantoms used for QA usually do not depend on a mixture of components or complicated structures, a sophisticated design with complex shape assumptions and several different tissue classes is often required for phantoms. Common measurement tools comprise visual assessment of the acquired images by medical experts. Other approaches try to ensure anatomical and functional correctness by adapting the phantom to a given patient data set (Prastawa et al. 2009). For example, a heart phantom based on intensity-averaged MR series of a single healthy volunteer is proposed by Moore et al. (2003).

Flexibility. The third requirement for a phantom is a flexible design approach. Each parameter should be easily changeable, resulting in the ability to generate phantoms of arbitrary shape, appearance, structure, or contrast behavior. Furthermore, simulation of dynamic processes including contrast enhancement or deformations due to tissue growth and shrinkage respectively would allow for a broad range of different applications.

2.1.2. The 'Perfect' Phantom

After discussing general design requirements in the previous section, what would be a perfect phantom? Generally, such a phantom should reflect all requirements (suitability, correctness, flexibility) to a certain extent. However, each application will have its own definition of what is needed for evaluation. A scanner calibration procedure does not need a phantom shaped like a real organ. Therefore, an object as shown in Figure 2.2 (a) could be described as a *perfect phantom* to control certain imaging parameters of an MR scanner.

In the field of medical image analysis on the other hand, a different approach is required. Here, a simple shape is often not enough to model anatomical variability of a real patient data set (cf. Fig. 2.2 (b)). For example, validating a new liver segmentation method requires a phantom with an appropriate shape model of the liver. A following quantification approach

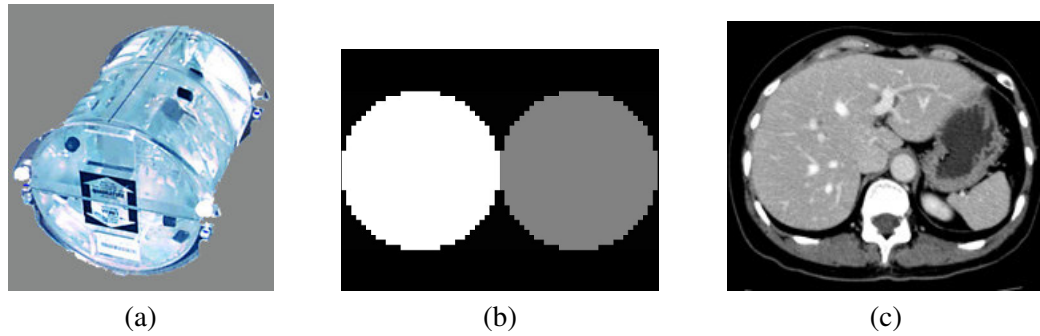


Figure 2.2. From simple phantoms to actual patient data sets — the perfect phantom for an application. (a) Physical phantom (from General Electrics) used for scanner calibration, (b) simple software phantom consisting of two circles for initial algorithmic evaluations, (c) CT patient data set of the liver without known ground truth.

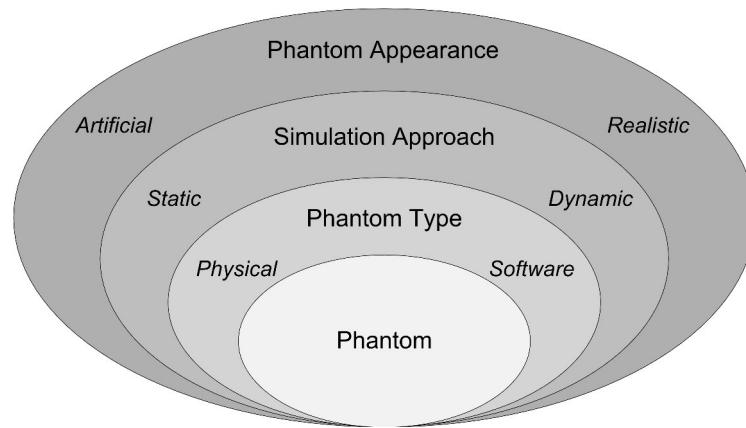


Figure 2.3. Categorization of the phantom design process.

would also require a known object volume of the phantom. Here, a *perfect phantom* consists of a data set that is not distinguishable from a patient data set. Additionally, the ground truth should be known for each parameter and should be easily changeable.

2.2. A Categorization Scheme for Phantoms

In addition to basic requirements for a phantom, a taxonomy should be established categorizing the design process. In this work, we propose a partition into different layers that are separated from the actual application. Figure 2.3 gives an overview of the proposed categories. Three main layers are identified, each employing certain constraints to determine the properties of a phantom.

2.2.1. Physical vs. Software

The first design decision in the phantom development process includes the choice of the required QA procedure for the system in mind, and thus the general phantom type. Testing of medical imaging devices requires real objects to be placed in the scanner, which are thus denoted *physical phantoms*. Physical phantoms were introduced into medicine for radiological use in the 1910's (Lee and Lee 2006). At that time, water tanks and wax blocks were usually applied for x-ray experiments. Although some are still in use in certain applications, technology has evolved along with radiological imaging equipment since. Today, physical phantoms are manufactured using a variety of available materials and processes. Tanks filled with water, oil, or other liquids are commonly used for acceptance testing of MR systems. Incorporated geometric objects based on acrylic compounds such as plastic allow for an examination of a broad range of imaging parameters.

Physical phantoms can also be generated from actual tissue. For example, Sekhar et al. (2014) propose a bovine–porcine tissue phantom for liver biopsy simulations. A model that allows to simultaneously capture internal images of a working heart and to record physiological parameters of both mammalian and human beating hearts has been proposed as well (Visible Heart Project). Another interesting approach is proposed by Klink et al. (2014). In their work, the authors develop a brain phantom from segmented MR and CT data of a patient. Different rapid prototyping methods such as a 3D printer are used to generate the modeled tissue classes.

If the development and testing of a new system is very complex or even too expensive to be implemented at every design stage, computerized models and simulations are an advantageous alternative in many areas including engineering, physics, and biology. Such *software phantoms* provide a flexible, reproducible, and cost-effective way for the evaluation of new methods and systems. Medicine as well as medical image analysis have also become an established field for the utilization of software phantoms. Early approaches for the simulation of human anatomy in medicine were used for planning of radiotherapy, developed in the 1960's (Caon 2004). Evaluations of modern segmentation and quantification methods are commonly based on both simple artificial objects such as a square or a sphere (Noe and Gee 2001; Warfield et al. 2004), as well as on models of anatomical structures (Collins et al. 1998; Kazemi et al. 2011).

The different approaches to represent human anatomy have led to a further specialization of software phantoms into three classes: stylized phantoms, voxel phantoms, and a combination of both, so-called hybrid phantoms (cf. Fig. 2.4). Each of these, also called computational anthropomorphic phantoms, differ in their basic design properties (cf. Fig. 2.1). An extensive review of the three software phantom classes can be found in (Caon 2004; Lee and Lee 2006). An example for each software phantom category is shown in Figure 2.5. A comparable categorization could also be established for physical phantoms.

A more detailed overview of parameters used in both physical and software phantom design along with their targeted applications is given in the next chapter.

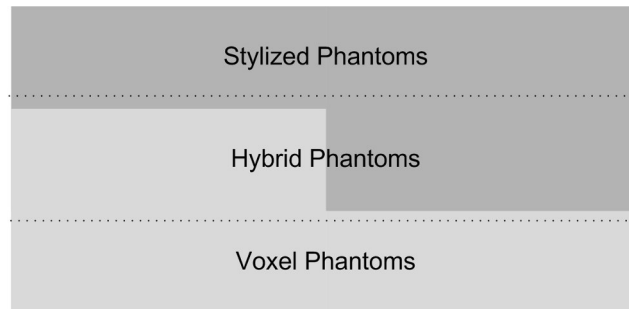


Figure 2.4. Categorization of software phantoms.

Stylized Phantoms

Stylized or mathematical phantoms describe the shape of a considered structure using simple mathematical representations such as planes, cylinders, spheres, or combinations thereof, i.e., the underlying ground truth is defined by mathematical equations. One of the first models was developed for dosimetry calculations in radiography and radiotherapy at the Oak Ridge National Laboratory (ORNL) (Fisher and Snyder 1966; Cristy 1980). These phantoms, also known as 'MIRD' phantoms (after the Medical Internal Radiation Dose Committee), assemble the major body sections and principal organs from simple equations. For example, the brain is represented by an ellipsoid, and the liver by an elliptical cylinder cut by a plane. In medical image analysis, stylized phantoms are an important step in validation studies of new segmentation and registration algorithms, e.g., (Drexler et al. 2004). Especially during early development stages, simple shapes with known geometries provide an excellent testing environment.

Voxel Phantoms

With the advances of modern tomographic imaging technology such as MR and CT, high-resolution 3-dimensional digital images of anatomical structures can be acquired, facilitating a better representation of the human body. Concomitantly, software phantoms constructed from such data sets have gained importance for different applications. These voxel or tomographic phantoms typically consist of a patient data set with manual or semi-automatic segmentations of different tissues or organs. Voxel phantoms may even comprise images from more than one individual.

The Zubal phantom offers a precisely labeled segmentation of different organs and other internal structures from a CT data set of a single subject. Furthermore, T2-weighted MR images of the brain were outlined, resulting in 62 designated neurological and taxonomical structures (Zubal et al. 1994). One of the most popular digital phantoms in medical image analysis is the brain phantom proposed by Collins et al. (1998). Due to its quality and its free availability from the Montreal Neurological Institute (MNI), it is often used in validation

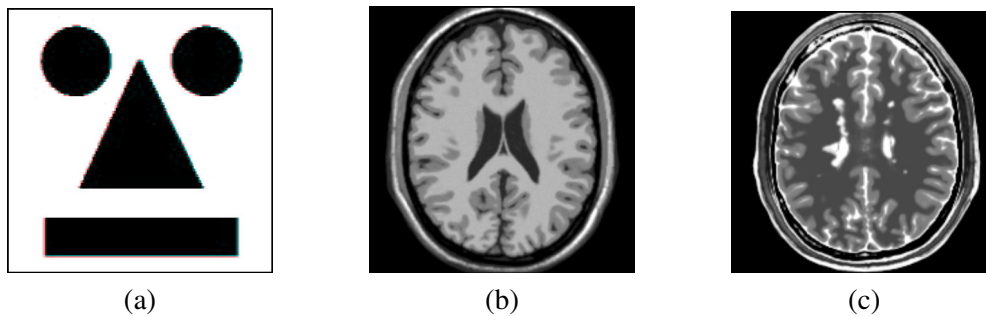


Figure 2.5. Examples of stylized, voxel, and hybrid software phantoms. (a) Stylized phantom: combination of geometric primitives, (b) voxel phantom: slice of the BrainWeb phantom (T1-weighted data set), (c) hybrid phantom: slice of the BrainWeb phantom with MS lesions (T2-weighted data set).

studies (Hahn et al. 2004; Cuadra et al. 2005), or as template in atlas-based segmentation and quantification methods (van Leemput et al. 1999).

Hybrid Phantoms

An approach that takes advantage of the flexibility of stylized phantoms and the anatomical correctness of patient-based voxel phantoms are hybrid phantoms. An interesting approach has been implemented by Segars et al. (1999). In their work, smooth surfaces are constructed from segmentations of a cardiac MR data set based on non-uniform rational B-splines (NURBS) technology. This enables the development of a model of the human heart. Similar approaches for modeling respiratory mechanics (Segars et al. 2001) as well as simulation of dynamic respiratory models for individual lung lobes and the airway tree are proposed by the same group (Garrity et al. 2003).

Burgess et al. (2003) propose a software phantom for mass discrimination in mammograms. In their work, masses extracted from digitized surgical specimen radiographs are added to regions of digitized normal mammograms. A similar technique has been applied to study the effect of wavelet-based data compression on lesion detection in digital mammograms (Suryanarayanan et al. 2005). A software phantom consisting of two concentric ellipsoidal surfaces with varying sizes and levels of segmentation error, and missing surface information for the segmentation of radiofrequency ablation induced thermal lesions is proposed by Lazebnik et al. (2002). Furthermore, a physical phantom using in vivo lesion images obtained from an animal model is used in their work.

2.2.2. Static vs. Dynamic

Once we have decided on the required phantom type, the next step is to determine what to simulate. Here we consider structural or static approaches and dynamic approaches that

simulate processes. The typical static method includes a certain object used for testing purposes, and the examined QA parameters largely denote morphological measurements such as the size and the volume of a lesion or an organ. A more complex task is to simulate a dynamic process within a phantom. Recent approaches include biomechanical modeling of deformations due to tissue growth and infiltration of brain tumors, as well as simulating treatment strategies for therapy planning (Swanson et al. 2003). A simulation of contrast enhancement characteristics is proposed in (Brix et al. 1999). A phantom that combines different organ models of the torso with cardiac and respiratory motion simulations has been developed in (Segars et al. 1999; Garrity et al. 2003).

2.2.3. Artificial vs. Realistic

A further design decision is the actual appearance of the phantom and its complexity. Artificial objects based on simple geometric primitives such as a cube or a sphere provide an easy-to-handle test framework with a known ground truth based on simple mathematical expressions (Cinti et al. 2004; Warfield et al. 2004). Moreover, they present a cost-effective design strategy for physical phantoms. For example, an inexpensive tissue-equivalent breast phantom consisting of lard (a solid cooking fat) surrounding a commercial jelly product is proposed by Liney et al. (1999). Nevertheless, anatomical correctness is favored in many cases. Especially image analysis methods that make use of anatomical prior information require realistic phantoms tailored to the considered application. Common design schemes consist of complex representations of anatomical structures using plastic cutouts (Hoffman et al. 1990), or use data sets from patients and/or healthy volunteers (Zubal et al. 1994). To take advantage of features from both, standardized mathematical equations and anatomical realism, hybrid phantoms have also been proposed by some investigators (Garrity et al. 2003).

2.3. Discussion

This chapter introduced two aspects we focus on in this work. The first part was dedicated to the question: *What is a phantom?* The second part provided an overview over common approaches in medical image analysis that will be further discussed in the context of phantom design in the upcoming chapters.

So, what is a phantom? Creating a model before creating a product has become a common technique in many engineering disciplines. Automobile designers build scale models of a new car and create 3D computer visualizations in order to provide a visible, early solution of their design. Such prototypes allow designer and customer to validate the requirements and design specifications. Phantoms in medical imaging are designed to achieve similar goals: The development of a phantom is concerned with the evaluation of new analysis approaches or the calibration and testing of imaging devices. Thereby, phantoms provide a cost-effective and time-saving approach in comparison to an analysis based on actual patient

Phantom Type \ Design Property	Suitability	Correctness	Flexibility
Stylized Phantoms	–	–	+
Voxel Phantoms	+	+	–
Hybrid Phantoms	+	○	+

Figure 2.6. Degree of agreement of the three software phantom types with the general design requirements introduced in Section 2.1.1. Good (+), average (○), low (–).

data. Furthermore, they allow focusing on the really important features of a testing procedure. For example, analyzing the field inhomogeneity of an MR scanner does not require a shape of a human organ. However, the quality and importance of the modeled parameters are essential issues for the applicability of the whole phantom, and the following chapters will discuss this aspect in detail.

Software Phantoms

Besides a general overview of design requirements we also proposed a categorization scheme that partitions a phantom into three categories, namely the phantom type, the simulation approach, and the phantom appearance (cf. Fig. 2.3). Especially software phantoms were surveyed, because they present the main category in this work. They provide a flexible and more reproducible design approach and an advantageous alternative to physical phantoms for many applications, e.g., applications that require a detailed object shape.

We divided software phantoms into three groups (cf. Fig. 2.4): stylized phantoms, voxel phantoms, and hybrid phantoms, each having their own advantages and drawbacks. Figure 2.6 illustrates the degree of agreement with the proposed design properties for each group. A similar categorization could also be used for physical phantoms.

Stylized Phantoms. The simple mathematical description of each object in stylized phantoms allows for a flexible and simple adaptation of imaging parameters. Furthermore, changing the volume or position of an object can be easily achieved. However, geometrical models based on mathematical equations alone inherently limit exact modeling of complex shapes, resulting in a low degree of agreement with respect to suitability and correctness of the considered parameters.

Voxel Phantoms. Although voxel phantoms offer the ability to create appropriate models of the human anatomy that surpass stylized phantoms (suitability & correctness), several limitations still exist. An important drawback is the missing ground truth of the scanned anatomical structures. For example, the amount of white matter in the brain or the volume of a tumor is not known and can only be approximated, which makes an analysis of new quan-

tification techniques difficult. Furthermore, a limited data pool and a limited resolution can not account for all anatomical variations and pathologies. The phantom construction can also be time consuming due to the required segmentation task for each structure. Moreover, small structures in the order of the voxel dimension may not be accurately segmented, and extensive acquisition protocols with long scanning times or repeated injection of contrast agents often prevent from scanning patients or probands in the preferred way. Therefore, this approach results in a limited flexibility.

Hybrid Phantoms. Hybrid phantoms are most suitable for the analysis of new methods and systems, since they combine the advantages of the two other categories. The resulting data sets provide a suitable model for several parameters plus a flexible handling of them. A main feature is the separation between object and background. An example of a hybrid phantom could be as follows: An object with a suitable shape incorporated into a patient data set. In other words, we focus on the development of objects with relevant parameter models, and assume the background to be already available. This approach provides great flexibility and suitability during object modeling. However, this approach results into a phantom design with only moderate correctness, since some parameters may not result from actual patient or volunteer data and have a rather simple layout. To this end, a careful selection of the important features has to be taken by the developer. The following chapters will present a survey on how different parameters can be modeled, and present our approach to phantom development. Furthermore, an analysis of possible applications and limitations is provided.

Part I.

Phantom Development

3. Parameters in Phantom Development

The development of phantoms in medical imaging is a challenging task. Important aspects are the complexity of the underlying problem domain and the difficulty of modeling all required parameters in a sophisticated way. The previous chapter has given a general overview of design approaches and proposed a categorization into different layers. Now, we present a more detailed examination of parameters used in phantom development. We propose a categorization into different groups: The first group contains morphological and topological parameters associated with the appearance of an object such as shape or volume (cf. Sec. 3.1). The second group includes parameters related to the acquisition process, such as noise, uniformity, or spatial resolution (cf. Sec. 3.2). The last two groups contain a general scanner parameter and the modeling of processes such as tumor growth (cf. Sec. 3.3). See Figure 3.1 for an overview.

For each group we present a general description of examined parameters. Furthermore, we give an overview of typical phantom examples from literature, that model the considered parameters as well as associated applications. Thereby, the description of phantoms is divided into physical and software phantoms. Each section concludes with a compact summary of the presented modeling approaches, again divided into physical and software phantoms.

3.1. Morphological and Topological Parameters

The initial design decision in phantom development is typically related to the appearance where important aspects are the morphology and the topology of the generated objects. This includes parameters such as shape, orientation, or volume. For example, a muscle has a different imaging characteristic than the liver. Normal tissue has a different appearance than pathological tissue.

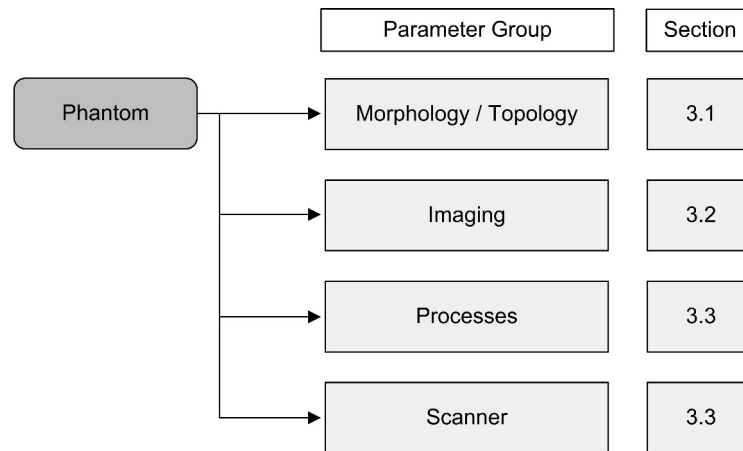


Figure 3.1. Overview of parameter groups presented in this chapter.

3.1.1. Parameter Description

Shape

The shape of a structure is an integral part of many clinical applications forming a relationship between visual appearance and function. Each organ has a characteristic shape, and pathologies are often related to an abnormal shape of the analyzed structure. Using shape related parameters has also become a common approach in medical image analysis. Many segmentation methods incorporate specific knowledge about the morphology of an organ to constrain the approximation of the object contour, e.g., by using a statistical shape model which provides a parametric description of object variations based on a set of aligned training data.

For example, Liney et al. (2006) propose a number of shape features such as convexity or circularity to differentiate benign and malignant breast lesions. Thereby, a common challenge is the typically large variability of anatomy within a shape population.

Structure

Another parameter to describe the appearance of an object is related to its underlying structure, i.e., the different components or tissue classes. Such a classification plays an important role in diagnosis and therapy planning, and multispectral acquisition protocols are often required. For example, a brain tumor can consist edema, active tumor, or necrosis. Here, a differentiation is essential of these tissue classes for treatment planning such as resection or radiation therapy.

Volume

The volume of an object is another important descriptor. Several studies have reported the clinical use of volumetric measurements as a surrogate marker for different applications and body parts. For example, brain atrophy as well as atrophy of certain brain structures such as the hippocampus have emerged as sensitive predictive markers for disease progression in Multiple Sclerosis and in Alzheimer disease (Miller et al. 1998; Henneman et al. 2009). A volumetric growth assessment of lesions is also an essential task in many applications including oncological therapy monitoring (Ko et al. 2003), estimation of total lesion load in Multiple Sclerosis (Molyneux et al. 1998), or the computation of residual tumor volume as a predictive survival rate after brain surgery (Wood et al. 1988).

Topology

Besides the general appearance of an examined object, its topology has to be considered for an in-depth analysis. We relate the topology to the location or position of an object within the background, as well as the position in relation to other objects. Although anatomical structures have their pre-defined positions or at least a reference range of typical locations, deviations can give insight into potential pathologies. Furthermore, the location of a pathological structure such as a lesion is an important feature for diagnosis and treatment planning: A tumor deep inside the brain is more difficult to access for surgical removal than one at the surface. In liver surgery, a tumor close to a large vessel can have major impact on the remaining liver volume, since a surgeon removes the tumor with a lap of surrounding healthy tissue.

3.1.2. Phantoms

After discussing common parameters related to morphology and topology, we now present a survey of phantoms modeling these features. The same categorization as above is used to group the referenced publications. Of course, the discussed phantoms in this section only present a subset of possible applications. We focus on common examples including brain, breast, heart, or liver.

Shape

Shape is an important spatial property of a phantom. Unfortunately, complex anatomical and pathological shapes are difficult to model, especially for physical phantoms. Furthermore, these phantoms have to cope with the problem of long-term physical stability. Nevertheless, several physical phantoms with a realistic anatomical appearance have been proposed, and some are even commercially available, e.g., via CIRS, Norfolk, USA. The Hoffman brain phantom (Hoffman et al. 1990) comprises a representation of the outer edge of the human brain, of the interface of white and gray matter, as well as of the ventricle region.

Nineteen independent plastic plates that can be stacked in a plastic cylinder, facilitate complex simulations of radioisotope distributions found in the brain. User-defined defects can be added to simulate clinical abnormalities. A related approach is presented in (Pupi et al. 1990).

Other groups use rather simple shapes such as spheres or ellipsoids to model their phantoms. Ko et al. (2003) use a realistically shaped chest phantom composed of different materials simulating bone, lung, muscle, and fat to study quantification methods for small pulmonary nodules. The actual nodules are acrylic spheres placed into small wells drilled into the material of each lung. A physical phantom of the upper abdomen consisting of 17 synthetic lymph nodes of ten different sizes is proposed in (Keil et al. 2009). A related, however much simpler phantom consisting of wells within a Plexiglas cylinder is presented by Drexl et al. (2004) for the analysis of vessel segmentation approaches. Physical phantoms for many other anatomical structures and applications have been proposed as well (Timinger et al. 2006; Mattila et al. 2007; Tofts et al. 1997; Cinti et al. 2004).

Similar to their physical counterpart, software phantoms have been proposed with a wide range of different shapes, each shape again being on a scale of simple geometric objects to realistic anatomical appearance. One approach to model anatomically realistic shapes within a software phantom is to use digitized real lesions, e.g., obtained after biopsy. A software phantom for mammograms is proposed by Burgess et al. (2003). Here, digitized masses from patient data sets are incorporated into normal mammograms. However, digitized real lesions are often not available, and other representations of these structures are required. A common method is to extract the shape from segmented volunteer or patient data sets. A popular example is the BrainWeb phantom by Collins et al. (1998). In a related approach, Kauffmann et al. (2003) use segmentations of the tibia and the femur geometries, as well as of cartilage for the quantification of cartilage thickness and volume changes.

Several simple geometric software phantoms have been proposed as well. Stylized phantoms are a common approach for dosimetry calculations in radiotherapy (Fisher and Snyder 1966; Lee and Lee 2006). Schlüter et al. (2005) propose a phantom for the analysis of fiber tracking algorithms on diffusion tensor data. Synthetic fiber bundles based on mathematical equations as well as clinical data serve as basis for evaluation. A spherical lesion phantom is added to assess the robustness of fiber tracking methods to fiber disturbance. A software phantom consisting of two concentric ellipsoidal surfaces with varying sizes and levels of segmentation error is proposed by Lazebnik et al. (2002) for the segmentation of radiofrequency ablation induced thermal lesions.

Structure

Developing phantoms that consist of several components is a challenging task, especially for physical phantoms. Each structure requires its own anatomical shape and imaging parameters such as contrast or noise. Therefore, common physical phantoms consist of one or more geometric structures and of a single material. Depending on the targeted applica-

tion, this can be an acceptable solution, e.g., for quality assurance of MR systems (Price et al. 1990). However, analyzing anatomical or functional patterns within the human body requires phantoms with a more detailed underlying structure. To give an example, a suitable phantom to examine brain tissue deformations during neurosurgery should include a model of all important structures such as different tissue types as well as blood vessels. A first step in this direction is presented in (Reinertsen and Collins 2006). Several physical phantoms of more than one component for other structures have been proposed as well (Hoffman et al. 1990; Timinger et al. 2006).

Software phantoms have the ability to be easily composed of different tissue classes, and phantoms have been proposed for various anatomical structures up to whole body phantoms (Lee and Lee 2006). Again, the BrainWeb phantom is a prominent example. It consists of ten tissue classes including gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Furthermore, an extension with additional structures (blood vessels, the dura matter, marrow, etc.) is proposed by Aubert-Broche et al. (2006). A brain phantom including tumor and edema is presented in (Prastawa et al. 2005). A diffusion-reaction model is used to guide edema growth and tumor infiltration. Zhang et al. (2008) propose a breast phantom including the simulation of skin, adipose tissue, and fibro-glandular tissue.

Volume

Phantoms used to analyze the volume of an anatomical or pathological structure often have a rather simple geometric shape. Reasons are the associated time and costs for developing complex objects with known volume plus correct imaging parameters and morphology. Another aspect is anatomical feasibility. Especially, methods that quantitatively assess the volume of lesions use roundish or ellipsoidal objects (Luft et al. 1996). Multiple Sclerosis lesions (Tofts et al. 1997) and pulmonary nodules (Ko et al. 2003; Winer-Muram et al. 2003) are typical examples. Kuhnigk et al. (2006) use a physical lung phantom consisting of 39 spherical and non-spherical objects of different diameters for a volumetric assessment of lung nodules.

Nevertheless, a detailed volumetric analysis often requires phantoms with a more complex shape. The BrainWeb software phantom is one of the most common data sets for the evaluation of methods for brain volumetry, see for example (Shattuck et al. 2001; Hahn et al. 2004; Cuadra et al. 2005).

Topology

The position of an object within the background structure is seldom explicitly modeled in physical or software phantoms. Instead, objects incorporated in phantoms are spread randomly or in some pre-defined pattern throughout the underlying background material (Pikus et al. 2006). Furthermore, a specific object position presumes a surrounding background related to an anatomical structure, i.e., there is no need for careful positioning if the back-

ground is made of a single material without any specific anatomical shape (Winer-Muram et al. 2003).

An approach towards a realistic topology could be to position objects at similar locations found in patient data sets. Reynolds et al. (2007) propose a direct mapping of 2D coordinates of melanoma positions onto a 3D anatomical model created from the Visible Human data set. Ko et al. (2003) place several wells at typical positions of lung nodules in the periphery of each lung at a pre-defined distance to the pleura and in the center of the lungs. Phantoms in the field of radiation dose calculation use the shape of a human body with incorporated objects (Lee and Lee 2006). Each object is shaped as an internal organ and placed at an approximately correct position.

Summary

This section provides a brief overview of the modeling approaches discussed above. Thereby, physical phantoms and software phantoms are given for each parameter.

Parameter	Physical Phantoms	Software Phantoms
Shape	Create shape	Digitized biopsy data, e.g., for lesions
		Mathematical representation
		Segmented volunteer or patient data
Structure	Combine different components of required shape	Segmented volunteer or patient data
		Model of tissue classes, e.g., growth models
Volume	Scanned object with known volume	Modeled object with known volume
Topology	Randomly spread objects	Randomly spread objects
	Objects placed at pre-defined patterns	Objects placed at pre-defined patterns
	Typical positions in the human body	Typical positions in the human body

3.2. Imaging Parameters and Artifacts

Besides morphological and topological parameters, imaging parameters and artifacts are further important aspects in phantom development. Acquiring a data set from an MR or CT

scanner comprises many features related to the acquisition process and associated artifacts. Image artifacts are intensity values in an image that do not have a corresponding anatomical basis, i.e., values that are not present in the imaged object. Thus, they influence the image appearance such as a hyperintense signal in the background or a drop-out of signal where there should be something. Each modality causes own artifacts, and considering all will go beyond the scope here. Therefore, we focus on MRI image data in this work, which have numerous potential sources of image artifacts and each pulse sequence has its own problems.

Here we only give a brief overview of common imaging parameters and artifacts. For an in-depth discussion we refer to standard literature and textbooks such as (McRobbie et al. 2003).

3.2.1. Parameter Description

Contrast/Intensity

To examine normal anatomical or pathological structures within an image, a contrast between adjacent tissue classes is needed. In MR imaging, an intensity value depends on several effects influenced by the used pulse sequence. Thereby, altering the local magnetic field with respect to a voxel position varies the signal intensity in the examined tissue. Furthermore, several new and evolving imaging techniques such as diffusion tensor imaging (DTI), magnetization transfer imaging (MTI), or MR Spectroscopy (MRS) have opened up new opportunities in MR imaging. Another possibility to change the contrast between certain tissue types is to administer a contrast agent to a patient. For example, a contrast agent based on the paramagnetic gadolinium is used to visualize areas of enhancing brain tumors with a disrupted blood-brain-barrier.

Noise

Unfortunately, imperfect imaging conditions result in fluctuations in the signal measured with an MR scanner and thus of the voxel values in an image. The term noise or random signal is used to describe this component. Different sources of noise arise during an acquisition such as thermal noise from the imaged subject or the coil temperature, or quantization artifacts from the analog to digital conversion. However, anatomical variations in a measured tissue may also result in deviations from the expected signal intensity. See (Sijbers 1998) for a more detailed overview. A noise measure is the Signal to Noise Ratio (SNR) that quantifies the amount of signal with respect to the amount of noise. A common definition is given by McRobbie et al. (2003) as

$$SNR = \frac{\mu}{\sigma} \quad (3.1)$$

SNR	Signal to Noise Ratio
μ	mean value of signal
σ	standard deviation of noise.

The SNR will be large for areas with a large mean signal value or a low noise level. The above mentioned spread of possible intensity values is commonly described by a Probability Density Function (PDF). The complex MR signal noise can be described by a Gaussian PDF as

$$p(i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \quad (3.2)$$

$p(i)$	Gaussian PDF with model parameters $\theta = (\mu, \sigma)$
μ	mean value
σ	standard deviation.

For magnitude MR images, this PDF transforms into a Rician distribution (Sijbers 1998). In the special case of low SNR, e.g., in the background of an MR image, this PDF is reduced to a Rayleigh distribution. For high SNR, the Rician PDF is well approximated by a Gaussian distribution. This result will be applied in Part II of this work on image segmentation and classification, where we use a Gaussian PDF to characterize the distribution of brain tissue classes. Other imaging modalities such as CT or PET have their own noise models.

Resolution

The spatial resolution of an image describes the level of detail, an MR scanner is able to achieve. Or simply stated, the image resolution is determined by the size of a voxel — the larger the voxel size, the lower the resolution, and the less accurate an object is imaged. Thereby, we have to consider the in-plane resolution and the slice thickness, which are often different. In this work, we typically work with data sets having an in-plane resolution of $1 \times 1\text{mm}^2$ and a slice thickness of 1 – 3mm. Unfortunately, MR resolution is limited with respect to the amount of SNR. The available scan time is another limiting factor.

Field Inhomogeneities / Nonuniformity

MR scanners seldom have a uniform magnetic field. Therefore, a measured homogeneous tissue region will often not result in a uniform MR signal intensity, but in variations across the image. These distortions are typically caused by hardware imperfections, but can also result from susceptibility of the imaged object.

Motion Artifacts

Another common artifact is caused by motion of the imaged object during the acquisition, resulting in a blurring of an entire image or parts of it. Two types of artifacts due to motion

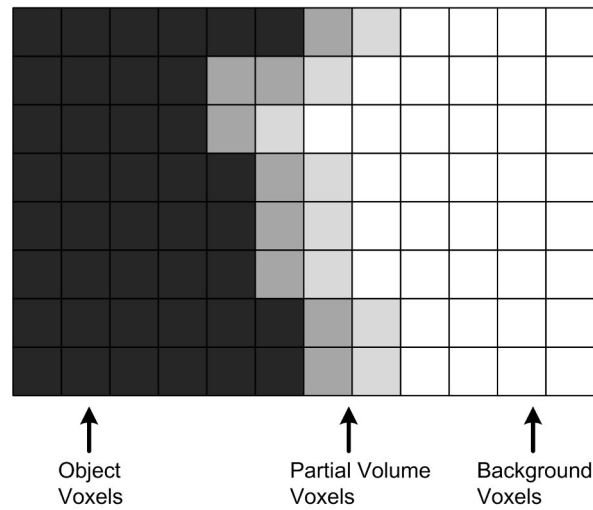


Figure 3.2. Illustration of the partial volume effect. A black object is sampled in a white background, resulting into voxels with different gray values between black and white at the object boundary.

can be observed: involuntary movements, such as respiration and cardiac motion, and random patient movements. A reduction of the first type can be used to gate the acquisition to the breathing or the cardiac cycle of a patient. Patient movement during the acquisition can be reduced by immobilizing the imaged body part.

Partial Volume Artifacts

Digitizing a continuous signal into a finite number of data points results into voxels that contain a mixture of tissue types, which is referred to as Partial Volume (PV) effect. The intensity values of these PV voxels can be calculated as proportional sum over the individual tissues. For example, a black object in a white background will result into PV voxels with mid-range gray values, i.e., the PV effect spreads out the object and it will appear larger than it actually is. This could for example have influence on treatment planning if the PV effect is not taken into consideration during measuring the object size. See Figure 3.2 for an illustration of the PV effect in such an image.

Several parameters have an effect on the PV effect, only some of which can be controlled. PV effects strongly depend on the size of an examined object. For example, the volume of the liver might be analyzed with a larger voxel size than a lesion with a diameter of only a few millimeters. A way out of this dilemma could be to increase the resolution and acquire data with a voxel size that is small enough for the examined object. Unfortunately, scanner and acquisition protocol, as well as patient acceptance of long scan times limit the feasible data resolution. Another parameter influencing the amount of PV is the shape of an object. A large surface area for a given volume will result in a larger amount of PV. Thus,

an irregular object is more affected than a spherical or cuboidal one, because a larger part of the object is close to the boundary and thus prone to PV effects.

Further Imaging Artifacts

Besides the above mentioned artifacts, several other can occur as well. Some are mentioned in this section.

Susceptibility and (metal) artifacts. Magnetic susceptibility describes the degree of magnetization of a tissue when placed in an external magnetic field. Different tissue types become magnetized to different extents, causing local distortions that result in a loss of signal. Similar artifacts occur in the presence of metals, which produce large magnetic field inhomogeneities because of their different susceptibility than body tissue.

Gibbs ringing. Truncation or Gibbs ringing artifacts are parallel bright or dark lines that can occur parallel or adjacent to high-contrast boundaries. The effect can be reduced by collecting more high-frequency data, i.e., by increasing the resolution. Another approach is to apply a smoothing filter.

Chemical shift. Chemical shift artifacts refer to signal alternations that result from differences in the resonance frequency (Lamor frequency) of nuclear spins between different body tissues. An important example is the shift between fat and water. Because fat has a lower Lamor frequency than water, signals from water and fat protons at the same location will be assigned different locations when converting from the frequency to the spatial domain, causing a misregistration.

3.2.2. Phantoms

Other than phantoms related to morphological and topological parameters presented in Section 3.1.2, phantoms modeling imaging parameters often do not require a complex appearance or a specific clinical application. Since few publications only model a single parameter, some phantoms are addressed more than once. An approach to develop phantoms dedicated to a single imaging parameter is presented by Price et al. (1990). The authors summarize different methods for quality assurance of an MR system and their associated physical phantoms. Required design criteria including appearance and material are examined, and applicable sequences and scan conditions are specified. To the best of our knowledge, a similar approach for software phantoms has not been proposed yet.

Contrast/Intensity

A physical phantom consisting of obliquely positioned cylinders within a plastic container for the analysis of brain lesion quantification methods has been developed in (Tofts et al. 1997). To obtain lesion-like intensity values, the lesion to white matter contrast is extracted from a data set of a patient with Multiple Sclerosis. Because these intensities vary over different positions within the white matter, different gray value transformations are proposed

for a more realistic appearance. Another approach to obtain appropriate intensity values for a physical phantom is to use material with relaxation times similar to that of human tissue. For example, Yoshimura et al. (2003) use carrageenan gel, a polysaccharide extracted from red seaweeds.

A common approach for software phantoms is to use so-called tissue templates, i.e., gray values extracted from patient data or from healthy volunteers (Burgess et al. 2003; Suryanarayanan et al. 2005). Other methods such as the popular BrainWeb phantom (Collins et al. 1998) uses an MR simulator based on a simulation of Bloch equations to provide realistic intensity values (Kwan et al. 1999).

Noise

As described in Section 3.2.1, magnitude MR images have Rician distributed noise within tissue and Rayleigh distributed noise in the background respectively. The BrainWeb phantom uses these distributions during the simulation of MR signal intensity values (Kwan et al. 1999). Other software phantoms often simplify this approach and model image noise by a Gaussian distribution (Noe and Gee 2001), which is also a common assumption in many image analysis methods.

In order to determine the amount of noise, two approaches can be found: MR simulators compute a distribution from an explicit model (Kwan et al. 1999). On the other hand, noise can also be measured directly from an MR scan. For example, Gedamu et al. (2008) estimate the standard deviation of the noise intensity distribution from a mask drawn in the background of a patient data set. Price et al. (1990) draw a region-of-interest in the data set of a physical phantom that consists of a uniform signal-producing material.

Resolution

For quality assessment or for the evaluation of new sequences, spatial resolution can be assessed with physical phantoms by bar patterns, i.e., by an array of altering signal-producing elements and non-signal-producing elements (McRobbie et al. 2003). The length of each bar should be at least twice the slice thickness (Price et al. 1990). A convenient way for software phantoms to model spatial resolution is to downsample a high-resolution data set.

Field Inhomogeneities / Nonuniformity

Similar to the analysis of noise or of the signal-to-noise ratio of a scanner, evaluating field inhomogeneities requires a physical phantom filled with a uniform signal-producing material. In image analysis, several algorithms use a nonuniformity correction as a pre-processing step. A software phantom consisting of a cube with a random intensity distribution multiplied by a parabolic function is used for the evaluation of the popular N3 method (Sled et al. 1998). A similar approach is used to extend the BrainWeb phantom with intensity inhomogeneities.

Motion Artifacts

As described in the previous section, patient motion can have several reasons, including breathing or heart motion as well as minor patient movement during a scan. Each variant causes blurring and ghosting artifacts and is difficult to analyze for a phantom. Sophisticated phantoms for this task should be able to generate a number of motion patterns for each application.

A common approach for physical phantoms is a movable device that can be controlled by some software program. For example, Fitzpatrick et al. (2005) developed a movable platform allowing the simulation of different motion patterns. Different objects can be placed on the device resulting in a range of potential applications. Timinger et al. (2006) propose an MR compatible ventricle phantom of the heart. Silicon tubes serve as model of the vascular structure. To model arbitrary motion patterns simulating heartbeat and respiration, the phantom is placed on a rigid plastic frame that can be mechanically controlled. A low-cost phantom that models cardiac motion is proposed in (Huber et al. 2000).

Instead of actually moving an object within the scanner, software phantoms provide a computational model of the underlying motion. A generic approach to obtain such a model is based on markers placed outside a patient's body during an acquisition that can be easily tracked. McClelland et al. (2006) propose a patient-specific motion model to describe deformations of lung tumors and of adjacent tissue classes during an average respiratory cycle. The model is reconstructed from a patient data set at free breathing by non-rigid registration of a high-resolution reference data of the same patient. Another approach has been implemented by Segars et al. (1999). In their work, a set of surface models is created from gated MR cardiac data of a normal patient and transferred into a 4D software phantom. Non-uniform rational B-splines (NURBS) are used to model heart motion and to permit the generation of arbitrary time-points. A similar approach is used by the same group to model respiratory motion (Segars et al. 2001; Garrity et al. 2003).

Besides the above-mentioned, patient-specific models, several paper propose physiological or biomechanical models to simulate involuntary patient movement. Wu et al. (2004) describe the respiratory motion of patients by a finite state model including three repeating breathing states, each corresponding to a typical action: exhale, end-of-exhale, and inhale. An analytical simulation of cardiovascular and respiratory mechanics is presented in (Kaye et al. 1998).

Partial Volume Artifacts

The size and shape of an object as well as the scan direction are the most prominent parameters influencing the partial volume (PV) effect, i.e., causing the border between tissue classes to be blurred. Therefore, a common approach for physical phantoms is to study objects with differing known volume or scanning them at different spatial resolution. Typical applications include the analysis of MS lesions (Tofts et al. 1997; Ballester et al. 2002) and

of lung nodules (Ko et al. 2003), where the lesion size is often small compared to the slice thickness. A similar approach is presented in (Plewes and Dean 1981). Here, a phantom consisting of a solid block with several holes of varying diameter is used to study contrast loss due to partial volume averaging.

A popular software phantom for the analysis of PV effect in image data of the brain is the BrainWeb phantom (Collins et al. 1998). Other software phantoms use object blurring or model PV artifacts as a separate layer within an object. For example, Shin et al. (2006) use Gaussian filtering of a binary object to simulate the decreased intensity values at the object's boundary. A square divided into three vertically separated regions is used by Noe and Gee (2001) to evaluate a segmentation with an explicit model for PVE. Two regions are considered as pure tissue classes with intensity values drawn from a normal distribution. In between, the third region is computed by linear interpolation.

Summary

To summarize this section, we give a brief overview for each parameter, divided into physical and software phantoms.

Parameter	Physical Phantoms	Software Phantoms
Contrast	Material with relaxation times related to human tissue	Tissue templates
	Post-processing of resulting image data to adjust intensity values	MR simulator (modeling of pulse sequence, magnetization strength, etc.)
Noise	None	MR simulator (e.g., Rician and Rayleigh distribution of modulus image voxels)
		Estimate standard deviation of noise directly from data set
Resolution	Array of altering signal- and non-signal producing bar patterns	Downsampling high-resolution data
Uniformity	None	Multiply signal distribution with mathematical functions, e.g., parabolic
		MR simulator (apply spatial varying perturbation of RF pulse flip angle)
Motion	Movable device producing motion pattern	Record patient-specific motion from external body marker
	Produce motion pattern in movable phantom	Motion described by physiological or biomechanical model

Parameter	Physical Phantoms	Software Phantoms
Partial Volume Effects	Scanned objects of different shapes with known volume	Downsample high-resolution image data
		Convolve (blur) image with Gaussian filter
		Interpolation between two tissue classes modeled as separate layer

3.3. Other Parameters

In the previous sections, we have analyzed parameters affecting the quality of a scanner or the output of an image processing algorithm. Now, we present two additional parameters that did not fit so far, namely *scanner* and *process*. The first parameter subsumes general features such as an update of the scanner software or a new sequence setup. The second parameter describes dynamic processes within a phantom.

3.3.1. Parameter Description

Scanner

Manufacturers of MR scanners offer a range of scanner types for different applications. Thereby, each scanner has its own hardware and software configuration including different field strengths, coils, scan sequences, etc. During the life time of a scanner several of these components will change due to hardware or software upgrades, and modifying such a parameter can lead to changes in the resulting image data. An example, that shows the effect of the parameter 'scanner' is presented in (Han et al. 2006). Here, an evaluation of the reliability of cortical thickness measurements within as well as across different scanner platforms is performed. The results show a variability in the global mean of the cortical thickness across platforms as well as across different field strengths. On the other hand, an upgrade to a newer scanner version did not have a significant effect on the results. Several other work also demonstrate the effect of field strength changes on the resulting image data, e.g., changes from 1.5T to 3T.

Process

Dynamic processes within a phantom comprise a number of different mechanisms. They can be regarded as a subgroup of the category 'simulation approach' defined in Section 2.2.2 (cf. also Fig. 2.3). In this work we focus on processes related to enhancement characteristics after contrast agent injection and on deformations, e.g., due to tissue growth or shrinkage.

Contrast agents are commonly used in MRI to improve the scanned images by altering relaxation times after injection. Thus, the contrast between different tissue classes is

increased in various parts of the body where the agent resides. Today, the paramagnetic gadolinium is one of the most commonly applied contrast agents. Applications include the evaluation of blood vessels, the analysis of infections and inflammations, diagnosis of cancer, or the characterization of different lesion types. For example, dynamic contrast enhanced MRI of the breast has shown to be a sensitive modality for early detection of cancer. Thereby, a series of scans is acquired at different time points, enabling an analysis of the uptake and wash-out characteristics of the tissue. Differences in contrast uptake between normal and pathological tissue provides a basis for differentiation.

As stated above, the second process analyzed in this section is related to deformations. In this work, we focus on deformations of pathological structures such as tumors: The progression of a tumor is a complex process with different stages from an initial avascular phase to invasion and metastasis. A hallmark is the breakdown of normal cellular interaction and control of replication. Angiogenesis is another process during tumor development, forming new blood vessels from existing vasculature in response to chemical signals from a tumor. See also (Hanahan and Weinberg 2000; Swanson et al. 2003) and references therein for a more detailed review. Another process resulting in deformations is caused by intraoperative movement. For example, the brain undergoes deformations during neurosurgery after the skull has been opened (brain shift).

3.3.2. Phantoms

Scanner

We relate this parameter to the overall process of phantom development rather than to a specific attribute. Today, no phantoms are known to model these parameters. One reason might be, that changing the scanner manufacturer or upgrading the hardware or software of a scanner have only a minor impact on the results as already described above. Furthermore, no detailed information are available from manufacturers, e.g., about software upgrades in an MR system.

Process

Simulations of object deformations and contrast enhancements allow the incorporation of clinical and biological knowledge into the phantom development process. Multi-compartment models are commonly used to describe the enhancement of macromolecular contrast agent particles in tissue and thus are an important tool for computer-assisted analysis of dynamic MRI. The Tofts&Kermode model is a popular approach to generate simulated perfusion data sets (Tofts and Kermode 1991). A simulation of contrast enhancement characteristics of different lesion types is proposed in (Brix et al. 1999).

The second investigated aspect in process modeling is related to deformations. In this work, we focus on tumor growth models, which can be classified into two categories: (1) cellular and microscopic models, and (2) macroscopic models.

Cell population models typically start from a small number of proliferating tumor cells and comprise a set of rules describing the evolution of their state and position. Kansal et al. (2000) propose a three-dimensional cellular automaton for solid brain tumor growth. The authors use a delaunay triangulation with a variable grid size for the underlying lattice, which allows for a tumor growth modeling over several orders of magnitude. A cellular automaton as well as a particle based tumor growth model are proposed in (Sierra et al. 2006). The authors put special emphasis on a realistic macroscopic appearance of common pathologies including polyps and myomas, to meet the requirements of a surgical training simulator for hysteroscopy.

Approaches that simulate tumor growth on a macroscopic level typically model soft tissue deformations. Thereby, two main directions can be taken: a biomechanical approach and a computational discrete approach (Delingette 1998). Common models typically consist of reaction-diffusion equations (Clatz et al. 2005). Furthermore, models based on continuum mechanics assuming linear elastic (Wasserman et al. 1996) as well as nonlinear elastic material (Kyriacou et al. 1999) have been described in literature. Here, tumor growth is influenced through internal and external forces that deform the underlying anatomy. Several additional anatomical constraints are introduced to facilitate a realistic tumor expansion and a deformation of surrounding tissue. In the last few years, these physics-based models have also gained increasing popularity in medical image registration due to their ability to constrain the underlying deformation in a plausible manner (Ferrant et al. 2001; Christensen et al. 1996). Thereby, an image is treated as a physical entity, either an elastic solid (Bajcsy and Kovacic 1989) or a viscous fluid (Christensen et al. 1996). The underlying physical principals, i.e., elasticity theory and fluid dynamics respectively, are a branch of continuum mechanics. An overview of registration approaches using physics-based models can be found in (Modersitzki 2004).

Unfortunately, these approaches only model the growth process and thus only one parameter. Phantoms with a model for tissue deformations plus imaging parameters such as noise or the correct contrast between tissue classes are hardly available. Sierra et al. (2006) manually texture the surface of their phantoms with image fragments from previous intra-operative recordings, to provide an image appearance similar to that of an intra-operative scene. A phantom for brain-shift simulations, suited for MR and ultrasound imaging, has been developed by Reinertsen and Collins (2006). This phantom consists of several parts simulating cerebral falx, brain, and vessel structures. To deform the brain, a catheter balloon is placed under the phantom and inflated by injecting water.

Summary

Similar to the other modeling groups, we provide an overview of discussed modeling approaches in tabular form, divided into physical and software phantoms.

Parameter	Physical Phantoms	Software Phantoms
Scanner	None	None
Process	Flow phantoms	Model uptake & wash-out characteristics, e.g., using a multi-compartment approach
	Repeatedly scan object at different degrees of deformation	Deform object based on methods from mathematics, physics, and biology
	Deform object by inflating external (attached) object, e.g. water-filled balloon	

3.4. Discussion

After introducing general aspects of a phantom in the previous chapter, this chapter focused on parameters used in phantom development. We presented a classification into four groups, each modeling a certain aspect of a phantom.

1. Morphology and topology
2. Imaging parameters and artifacts
3. Dynamic processes
4. Other scanner characteristics

A description of each parameter was followed by a discussion of characteristic phantoms proposed in literature.

Today, phantoms are used for many applications, and only some have been discussed in more detail in this chapter. We concentrated on phantoms for major anatomical and pathological structures such as brain and heart as well as lesions. Furthermore, phantoms for the evaluation of segmentation and quantification approaches were presented. This will be a focus in the upcoming chapters.

General

Both, physical and software phantoms have their own characteristic construction methods, and both types have been proposed in literature for the parameters presented in this chapter. Physical phantoms are commonly developed for quality assurance, since they can be placed in a scanner. Unfortunately, building such a phantom is associated with a large manual effort. One has to gather all components and assemble them to form the final phantom. This often results in objects with a simple shape, although realistic anatomical appearances have been proposed as well and are even commercially available. Another issue is chemical and

physical long-term stability: A phantom should not change its parameters over time, e.g., contrast changes due to different relaxation times.

Software phantoms on the other hand do not need any assembling of hardware objects. Their strength lies in the flexible and correct parameter modeling, and many different approaches have been proposed. Nevertheless, building a software phantom is by no means less challenging or less time consuming. Correct models of the imaging parameters need to be extracted. Furthermore, data sets of an investigated body region have to be acquired to extract anatomically realistic morphology and topology, which can require hours of scan time.

A common approach to software phantom design is a direct simulation of the underlying acquisition process. Thereby, the imaging parameters are controlled via mathematical equations. For example, Kwan et al. (1999) introduce an MR simulator, that is also used in the BrainWeb project. Drexler et al. (2004) propose a simplified model of the CT scanning process to generate software phantoms consisting of a homogeneous background with small vessel-like structures. A more sophisticated software package simulating the process of projecting X-rays through an object is available from the CTSim project (CTSim 5.1.2).

Another method to software phantom development is to simulate a data set from a different modality. Kiebel et al. (1997) use high-resolution MR data to generate simulated PET images. They apply a set of transformations including intensity modifications and smoothing. Furthermore, Gaussian noise and a rigid body transformation are added. The resulting PET data are then used to evaluate multi-modal image registration methods.

This approach is especially beneficial for the evaluation of registration algorithms, because both images originate from the same data set with the same artifacts. Moreover, motion artifacts with known ground truth can be easily added. Unfortunately, such a ground truth is not available for many other parameters including morphological as well as imaging parameters.

Although we have discussed major parameters used in phantom design, not all features of the acquisition process and the considered structures can be investigated. Some might not be known to have an influence on the results or might even not be known at all. Others, such as the global parameter *scanner* are not used because they only have a small impact on the resulting phantom, and little information is available about parameter changes, e.g., changes due to scanner upgrades.

Parameter Selection and Grouping

An important constraint in this chapter was the chosen parameters. But why is this selection a correct assumption? The parameters related to morphology and topology (cf. Sec. 3.1) as well as the parameters describing dynamical processes (cf. Sec. 3.3) are selected based on the examination of published work in the field of medical image analysis. Therein, the shape or the volume of an anatomical or pathological structure are important markers for diagnosis and therapy monitoring, as we have already presented in the previous sections.

The selected imaging parameters in Section 3.2 are commonly used in quality assurance (QA) programs of MR scanners (McRobbie et al. 2003). They have a major effect on the resulting image quality and are thus key parameters in QA. Firbank et al. (2000) developed guidelines for QA based on a comprehensive analysis of an MR scanner over the course of one year. They measured various imaging parameters including the signal to noise ratio, image uniformity, and resolution using two different physical phantoms. In a nationwide survey on quality assurance on MR scanners in England with 24 participating hospitals, Koller et al. (2006) found SNR and image uniformity tests to be among the most frequently evaluated parameters in QA programs. A similar study was performed by McRobbie and Quest (2002), monitoring effectiveness and relevance of quality assurance for 17 MR systems from four manufacturers.

The analysis of different parameters in this chapter has shown that each parameter requires its own modeling approach. Furthermore, it is not only related to the considered application, but also to the phantom type, i.e., physical vs. software. Thereby, various methods from physics and biology are used, modeling normal as well as pathological processes in the human body. Other approaches use statistical methods based on training data sets that show typical variations of the modeled parameter. To give an example, evaluating a brain segmentation approach requires a different phantom shape than evaluating a liver segmentation. Furthermore, a segmentation will focus on different parameters than a quantification of the same structure.

Nevertheless, parameters can seldom be considered separately, and changing one parameter will affect others. For example, changing the shape of an object from a cube to a sphere will also influence the resulting partial volume artifacts. The quality of a phantom might even worsen if one parameter excessively changes, e.g., large motion artifacts can impede the analysis of a quantification algorithm or even make it infeasible. Thus, each parameter has to be described by its individual scope, which can change from application to application. Unfortunately, a method is still missing that defines such a range for all application-relevant parameters.

Relation to Design Properties

Even if we might have chosen all relevant parameters for a phantom, the overall performance is still unclear. In Section 2.1.1 we defined general design properties for this task, namely suitability, correctness, and flexibility. But to what extent do the cited phantoms meet these requirements? To provide a more in-depth analysis of this issue, we discuss three examples that were frequently used in this chapter. See also Figure 3.3 for examples of these phantoms.

Tofts et al. (1997). This work proposes a physical phantom for the quantitative analysis of brain lesion volume estimation schemes. Nine cylinders are placed in a container at different positions and with different orientations. Adjustments are made to the

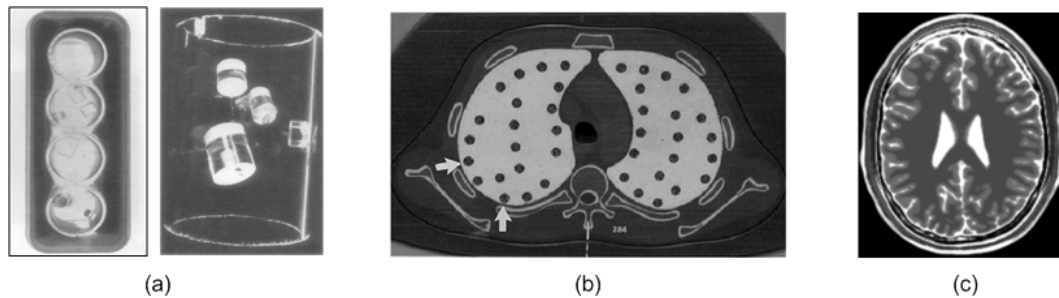


Figure 3.3. Examples of typical phantoms that are frequently referenced in this work. (a) Physical phantom for analysis of MS lesions proposed by Tofts et al. (1997) (left: whole phantom; right: zoom of single annulus), (b) physical chest phantom with drilled wells proposed by Ko et al. (2003), (c) slice of the BrainWeb phantom (T2-weighted data set).

resulting image data to obtain more realistic intensity values. Although the authors state that the phantom is quick to manufacture (about 2h), especially morphology related parameters such as a more complex object shape will be difficult to change. Therefore, the *flexibility* of the phantom is low. On the other hand, the ground truth is available for various parameters. Especially for those relevant to the targeted application, e.g., PV artifacts, contrast, and resolution. Other parameters can only be estimated (noise) or are not considered at all (field inhomogeneities, motion). This results in an average *suitability* and *correctness* of the phantom.

Ko et al. (2003). In a related approach, Ko et al. use a physical chest phantom with wells inside the lungs for the analysis of lung nodules. Each well contains an approximately spherical nodule object. The selected phantom type again reduces the *flexibility* of this approach. *Suitability* and *correctness* also receive only average grades. However, the usage of a chest phantom results in a better quality than the previous approach. For example, the underlying structure lacks of several components such as vessels or bronchi that will affect the results in real patient data.

Collins et al. (1998). The BrainWeb phantom is a software phantom of the brain resulting from a sequence of processing steps. It offers a *great flexibility* for various parameters including the amount of noise or the resolution. The phantom also has a good *suitability* and *correctness*, since its construction is based on a set of scans from one volunteer, affecting morphological parameters such as shape and structure. Additionally, an MR simulator is used to predict image contrast, PV artifacts, noise, and nonuniformities. On the other hand, the phantom is generated from a single individual, reducing the parameter distribution and thus the *flexibility* and *correctness* of this approach. For example, it is not possible to change the shape of white matter or cerebrospinal fluid.

Unfortunately, the descriptions above and the rather fuzzy terms (*good*, *average*, *low*) provide only a qualitative analysis of a phantom. An objective discussion including quantitative

measurements is still missing. This will be addressed in Part II of this work.

Lessons Learned

To conclude, important aspects of parameter modeling for phantoms can be summarized as follows:

1. Not all parameters can be modeled.
2. Each parameter requires an application-specific modeling.
3. Each parameter has a range of suitable values.
4. Modifying one parameter has also an effect on others.
5. A phantom models more than one parameter.

Based on these aspects, we will propose a new approach for software phantom development in Chapter 4 that covers several parameters discussed in this chapter. For each parameter, we present a dedicated model and discuss relations between parameters. An application-specific modeling with appropriate parameter ranges is then proposed in Chapter 5 and Chapter 6.

4. Design and Construction of Software Phantoms

After an overview of current phantom design approaches and modeled parameters in the previous chapters, we now focus on how to develop our own phantoms. Since we aim at the design of hybrid phantoms that propose a separation between object and background, we focus on the parameters required for object design. Our goal is a set of building blocks that can be easily described and exchanged. This will allow us to efficiently design own phantoms, for example within a dedicated software assistant. Therefore, we introduce a modular approach for the phantom design process in the first part of this chapter. Our novel method is suited to describe any phantom, including both physical and software phantoms. Three main tasks are distinguished: object design, background design, and object incorporation, where each task has its own characteristics and design properties.

In Section 4.2, we then propose models for all major parameters such as object position, shape, or intensity values. A formalized description for each parameter is given based on our modular design approach. Based on the developed set of building blocks, we introduce a software assistant for the development of software phantoms in the third part of this chapter. We propose an easy-to-use tool that allows us to interactively combine the parameter models described before.

4.1. Modular Phantom Design

Phantoms have become a standard way to validate the quality of measurements during the development and evaluation of new imaging devices and algorithms. Especially software phantoms are an integral part of solid test specifications for new image analysis methods. However, most approaches today facilitate a merely ad hoc design dedicated to the underlying application. A general design approach is still missing.

In this work, we propose a modular description of the phantom design process, characterized by a set of modules with different functionalities. We aim at a system that enables an information fusion of several parameters, allowing for a sound analysis and comparison of different design approaches. Thereby, our goal is to provide a systematic description of an arbitrary physical or software phantom.

Three main steps are required to determine the overall phantom development process:

The design of a suitable object, the design of a related background, and the incorporation of one or more of these objects into the background. Each step consists of a number of processing steps, providing a detailed description of required parameters and modeling schemes. See Chapter 3 for an overview of characteristic parameters used in phantom design. Furthermore, each phantom has its own targeted application domain consisting of a specific task, a body region, and an imaging protocol. A similar notation is also used by Udupa et al. (2006) to specify the application domain of image segmentation evaluation methods.

In this work, the object of a phantom mainly determines the task of the targeted application, whereas the background typically specifies the body region. For example a phantom dedicated to the evaluation of liver tumor (object) segmentation methods using a CT scan (background). If the background is merely used as image background and consists of a homogeneous material or a constant gray level, the object already determines the body region. For example, Reinertsen and Collins (2006) propose a brain-shaped object with a vessel structure for the simulation of brain-shift. In their work, the brain is placed in a liquid filled acrylic plastic container.

4.1.1. Modules

To formalize the overall phantom design process, we propose a structured development based on functional units that contribute to the overall system. Each of these so-called *modules* consists of a set of properties and behaviors, allowing to extract or combine information. We also denote the description of a module a *model*. A module M_t of type t is then defined as a tuple

$$M_t := \langle H, x, f, \theta, y \rangle \quad (4.1)$$

M_t	Module of type t , $t \in \{parameter, fusion, state\}$
H	hypothesis
x	input value(s), $x = \{x_1, \dots, x_N\}$
f	transition function
θ	parameters of transition function
y	output value.

The first parameter determines the overall assumptions of the module. In other words, each module is developed based on a certain hypothesis H . For example, this could be a characteristic object intensity value. Nevertheless, this provides only a description. A discussion about the correctness of the underlying assumptions is not given, but will be proposed in Part II. Besides the hypothesis, a module consists of one or more input values x , which can be image data or other external sources, as well as the output of another module. A module has one output value y . A transition function $f(x, \theta)$ with a set of parameters θ is used to transform the input. Therefore, it also provides a description of the module's task. For example, a module dedicated to the simulation of cardiac motion based on non-uniform rational B-splines (f) is proposed by Segars et al. (1999). Therein, patient data (x) are used to extract a 4D motion model of the heart (y). A graphical representation of a module is

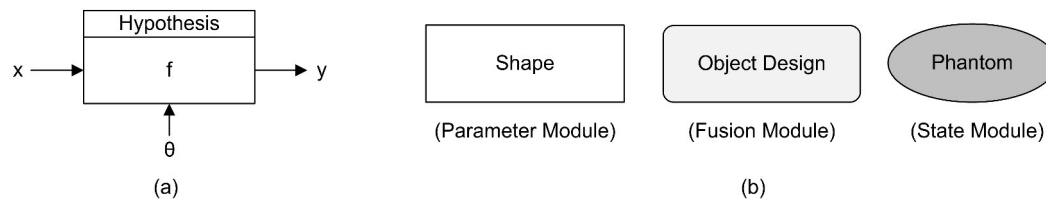


Figure 4.1. Illustration of a module. (a) General module layout: Input of a module are parameters θ as well as input values x such as image data. The transition function f is used to compute the output y . (b) Differentiation of different module types.

shown in Figure 4.1.

To summarize, our modules are somewhat related to components in component-based software engineering (Brown and Wallnau 1998; Broy et al. 1998), and can be described as follows:

Module

A module is a part of a phantom that provides a model for a certain task or parameter.

In contrast to the definition given in Equation 4.1, our definition above is a rather informal description similar to the definitions of quantification or segmentation given in Chapter 2. Nevertheless, it covers the main features of a module, i.e., being a part of a phantom with a clear interface allowing a common description as well as a flexible replacement. Thereby, only modules that actually involve a concrete model are used. Parameters that can only be estimated, e.g., noise or partial volume effects for a physical phantom, are thus not used as part of the phantom specification (cf. Fig. 4.3-4.5). For example, an explicit noise model is not specified in the phantom by Ko et al. (2003) and the corresponding module is thus not given in the phantom description (cf. Fig. 4.4). Tofts et al. (1997) add additional Gaussian noise to their phantom after image acquisition. In this case, we list the associated module in brackets (cf. Fig. 4.3).

Three different module types are distinguished (cf. Fig. 4.1):

1. Parameter Module
2. Fusion Module
3. State Module

A *parameter module* delineates the most elementary module. It represents a parameter model and can be further divided into groups as proposed in Chapter 3. We distinguish between imaging parameters, morphological and topological parameters, and other parameters such as general scanner properties and the modeling of processes. A *fusion module*

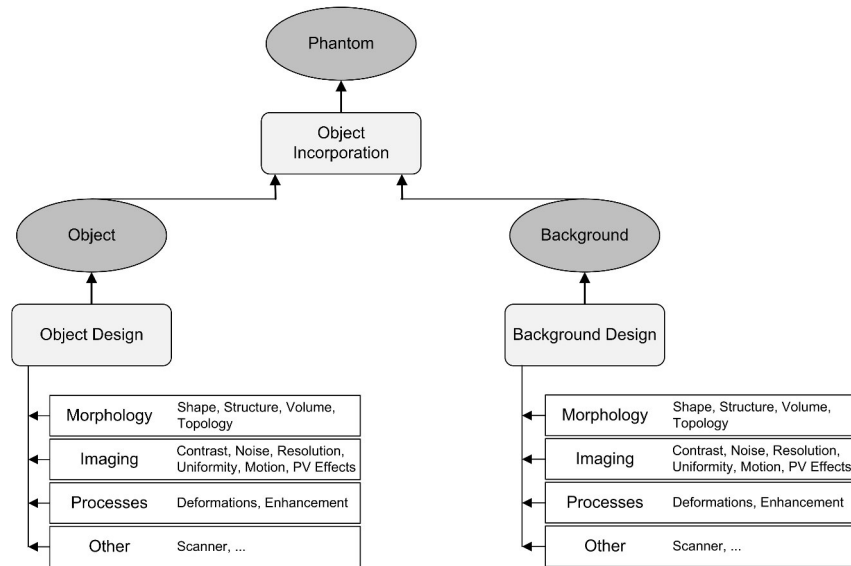


Figure 4.2. A modular approach to phantom development.

combines different parts of a phantom. For example, the *Incorporation* module is supplied by the two state modules object and background (cf. Fig. 4.2). Finally, a *state module* describes one of the three main outputs during phantom development, namely the object, the background, or the final phantom.

4.1.2. A General Phantom Description

After defining the components required during phantom design as well as their interactions, we now combine the different modules within a common representational format. The three main steps, i.e., object and background design and incorporation, are integrated as fusion modules. Thereby, object and background design consist of several parameter modules. The result of the design process is a state module, i.e., object or background. The fusion module Incorporation takes these two states as input and combines them to the final phantom. To simplify the workflow, no additional parameter modules are used in this stage. However, please note that more than one object might be involved during object incorporation as discussed above. Figure 4.2 illustrates our resulting approach with the required modules.

To summarize, our approach offers two ways to describe a phantom: (1) A rather high-level description providing a quick overview especially including the phantom type, the targeted application, and the main hypotheses and modeled parameters, and (2) a detailed analysis of all modules used during phantom development. We have developed an easy-to-use tabular description based on the high-level description that will be used throughout this work. To demonstrate the general applicability of our method and to illustrate our description, we exemplarily show three typical phantoms that were already used in Chapter

Phantom Description	
Author :	P.S. Tofts et al.
Reference:	P.S. Tofts et al. "An oblique cylinder contrast-adjusted (OCCA) phantom to measure the accuracy of MRI brain lesion volume estimation schemes in multiple sclerosis". Magn Reson Imaging 15:183-192, 1997.
<div style="border: 1px solid gray; padding: 2px; display: inline-block;">Application Domain</div>	
Phantom Type :	Physical
Task :	Compare accuracy of brain lesion volume estimation schemes
Body Region :	Head
Imaging Protocol:	1.5T MR, 5mm slice thickness
<div style="border: 1px solid gray; padding: 2px; display: inline-block;">Main Steps</div>	
<u>Object</u>	
Hypothesis :	Compact acrylic cylinders of varying diameters
Parameters :	Shape, Volume, Topology, Contrast, (Noise), Resolution
<u>Background</u>	
Hypothesis :	Water-filled container plus acrylic annuli
Parameters :	Shape, Structure, Volume, Topology, Contrast, (Noise), Resolution
<u>Incorporation</u>	
Hypothesis :	Cylinders glued to the inside of an annulus, which is then immersed into the container
Parameters :	-

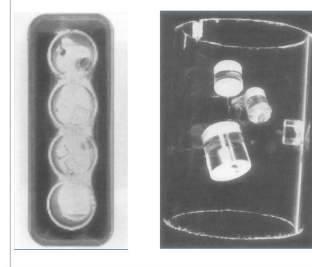


Figure 4.3. Description of the physical phantom proposed by Tofts et al. (1997).

3 to review phantom design properties (cf. Sec. 3.4). Both, physical and software phantoms are used. See also Section 4.4 for a discussion of our approach.

Figure 4.3 gives a description of the phantom developed by Tofts et al. (1997). The objects of this physical phantom are several cylinders placed in a container at different positions and with different orientations. A container filled with water is used as background. Adjustments are made to the resulting MR images for more realistic intensity values. The second phantom is based on the work of Ko et al. (2003) and consists of wells drilled into a chest phantom. Approximately spherical objects are inserted into small wells simulating lung nodules. See Figure 4.4 for an overview. Finally, Figure 4.5 summarizes the phantom described by Collins et al. (1998). The BrainWeb software phantom uses a number of processing steps to simulate an MR image data set of the brain. Thereby, several parameters

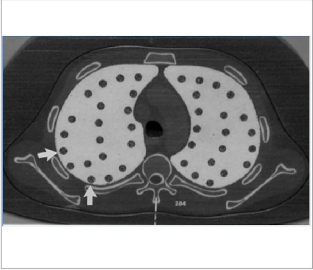
Phantom Description	
Author :	J.P. Ko et al.
Reference:	J.P. Ko et al. "Small Pulmonary Nodules: Volume Measurements at Chest CT – Phantom Study". Radiology 228(3):864-870, 2003.
	
Application Domain	
Phantom Type :	Physical
Task :	Compare methods for quantifying pulmonary nodule volume
Body Region :	Chest
Imaging Protocol:	CT
Main Steps	
Object	
Hypothesis :	Spherical plastic objects
Parameters :	Shape, Volume, Topology, Contrast, Resolution
Background	
Hypothesis :	Commercial chest CT phantom
Parameters :	Shape, Structure, Volume, Topology, Contrast, Resolution
Incorporation	
Hypothesis :	Objects inserted into wells drilled into the material of each lung
Parameters :	-

Figure 4.4. Description of the physical phantom by Ko et al. (2003).

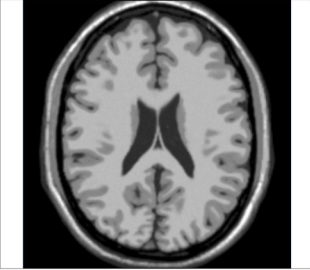
Phantom Description	
Author :	D.L. Collins et al.
Reference:	D.L. Collins et al. "Design and Construction of a Realistic Digital Brain Phantom". Transactions on Medical Imaging 17(3):463-468, 1998.
	
Application Domain	
Phantom Type :	Software
Task :	Simulation of MR images of the brain
Body Region :	Head
Imaging Protocol:	MR
Main Steps	
Object	
Hypothesis :	Data set created by registering 27 scans of the same normal volunteer
Parameters :	Shape, Structure, Volume, Topology, Contrast, Noise, Resolution, Uniformity, PV Effects
Background	
Hypothesis :	Background of acquired data set
Parameters :	Contrast, Noise, Uniformity
Incorporation	
Hypothesis :	Data of volunteer; MR simulator to account for effects of various imaging parameters
Parameters :	-

Figure 4.5. Description of the BrainWeb phantom proposed by Collins et al. (1998).

are addressed including the spatial distribution of different tissue classes as well as imaging parameters such as noise or intensity values.

4.2. Parameter Modeling

After introducing a general phantom description, let us now focus on the required parameters. Our component-based phantom description does not make any specific assumption on the phantom type or appearance and is therefore suited for a large range of applications. Because we focus on applications in the field of neurology and neurosurgery, some parameter models are already tailored to meet the requirements of the targeted phantoms within these areas, e.g., the modeling of object positions. Nevertheless, other applications are possible and a discussion about potential applications is given at the end of the first part of this work. We use the categorization introduced in Chapter 3, i.e., first parameters that describe the overall object appearance (cf. Sec 4.2.1) followed by imaging parameters in Section 4.2.2. In Chapter 5, we will additionally provide examples on how to model parameters for object growth and enhancement characteristics.

In the following sections, each parameter model is formalized by a module description based on the definition given in Equation 4.1.

4.2.1. Morphology and Topology

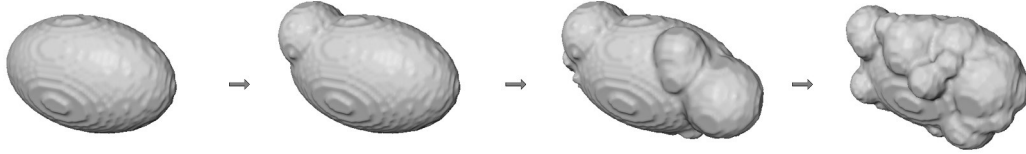
The first parameter group is related to morphological and topological parameters characterizing the overall object appearance. In the following sections, we present models for four different parameters, i.e., shape, structure, volume, and topology.

Object Shape

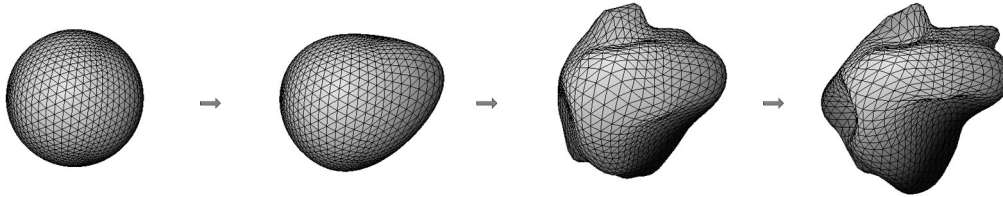
The first investigated parameter is dedicated to the overall object shape. We propose a parameter module $M_{p_{shape}}$ that can be used to build object appearances of varying complexity, ranging from simple shapes to complex models that better reflect anatomical reality. A software assistant that provides a user-controlled design is introduced in Section 4.3.

A common method to shape development is object shapes based on geometric primitives. Mathematical equations provide a standardized and yet flexible approach that is sufficient for many applications. Unfortunately, these phantoms can not capture anatomical variability with typically irregular shapes. An important step towards this goal is to give up this rather strict object definitions with a smooth and homogeneous surface. In (Rexilius et al. 2003), we applied a generic approach that allows the construction of objects from a combination of geometric primitives. Ellipsoids of different size and shape are iteratively placed at random positions on the initial object surface in an iterative fashion. A schematic overview of this algorithm as well as exemplary results are given in Figure 4.6 (top row).

1. Combination of Geometric Primitives:



2. Deformation of WEM Surface:



3. Manual Segmentation of Anatomical Structures:

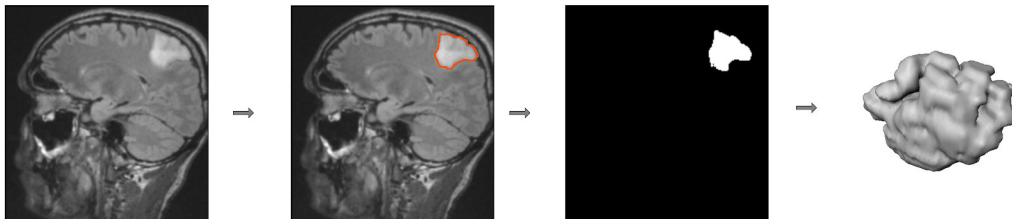


Figure 4.6. Three approaches for object shape design used in this work. (top) Object shape based on a combination of geometric primitives such as ellipsoids, (middle) geometric primitive that is subsequently deformed using a WEM data structure, (bottom) object shape from segmentation mask.

Since an iterative object placement as described above can be time-consuming, we use a further approach that introduces geometric distortions to a given object surface. The 3D representation of the current object is based on a boundary representation (also known as b-Rep), which describes the object as a collection of connected surface elements. Both, the geometric data of primitive geometric entities such as faces, edges, and vertices, and the topological data maintaining the connectivity between these entities are stored. We use a winged-edge mesh representation (WEM) as data structure, allowing a quick traversal between faces, edges, and vertices in 2D and 3D. An arbitrary viewer position in the vicinity of a WEM contour can be interactively selected, which influences all nodes within a sphere around the click point. These nodes will then follow any mouse movement to a certain degree (cf. Fig. 4.6 (middle row)). A comprehensive review of winged-edge data structures can be found in (Baumgart 1975).

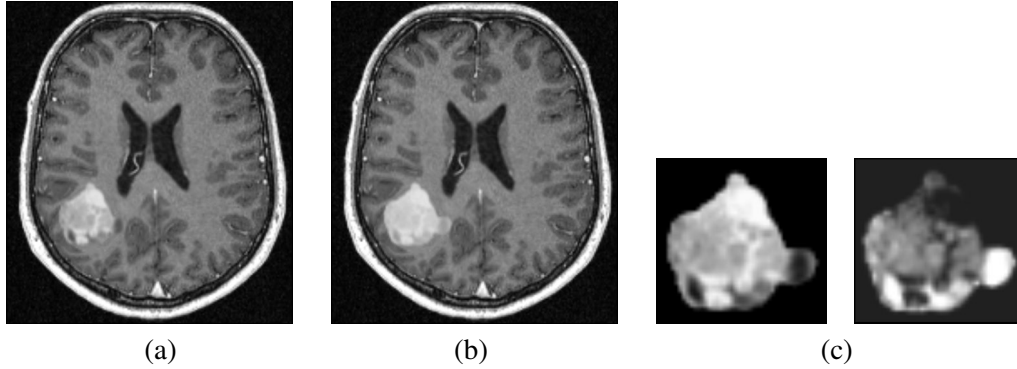


Figure 4.7. Brain tumor phantom with different amount of necrosis. (a) T1gd reference image with incorporated tumor, (b) tumor phantom with necrosis scaled to $\max_{necrosis} = 50\%$, (c) examples for active tumor tissue and necrosis maps.

In (Rexilius et al. 2004), we introduced a third approach to object design for software phantoms, based on manual segmentations of anatomical structures. Different object shapes are generated from segmentations of different patient data sets with similar anatomical structures. For example, Figure 4.6 (bottom row) shows an object resulting from a manual segmentation of a brain tumor. Unfortunately, this method limits anatomical variability to the number of available data sets. Here, additional geometric deformations of the segmentation mask might be allowed, e.g. based on an interactive approach as described above. To summarize, the parameter module $M_{p_{shape}}$ is described by

Parameter	Description ($M_{p_{shape}}$)
H	Compact object shape, i.e., all voxels are connected
x	Zero or one input data depending on selected approach (cf. Fig. 4.6)
f	Manual drawing; deformation of WEM structure; segmentation
y	I_{binObj} (binary object volume)

Object Structure

If an object consists of more than one component, e.g., a brain with white matter, gray matter, and cerebrospinal fluid, each structure, i.e., tissue class, requires its own modeling. This includes a dedicated shape for each tissue class, which can be generated with one of the methods described above. Furthermore, the amount of tissue per pixel has to be determined for all components. Similar to the shape module, this can be a user-defined definition done by an expert. Another approach is to segment an actual patient data set and then define the portions based on the resulting tissue maps. For example, we will develop brain tumor phantoms consisting of two overlapping tissue classes, active tumor tissue and necrosis in Section 5.3. An example is shown in Figure 4.7.

A related approach to generate reasonable object portions is texture synthesis from a sample database extracted from patient data, as proposed for brain tumors in (Prastawa et al. 2009). Unfortunately, this method requires a large amount of reference data to model a sufficient amount of anatomical variability. In this work, we propose a parametric model based on 3D simplex noise introduced by Perlin (2002) to generate different lesion textures, which is frequently used for organ surface-like textures in surgery simulators. Instead of relying on a large pool of reference data, this approach allows us to automatically generate largely heterogeneous object appearances by changing very few parameters. In Section 5.2, we will use this approach to develop textured MS lesion objects. See also Figure 5.6 for a comparison of different real lesions and corresponding phantom results. The resulting parameter module $M_{p_{structure}}$ is given by

Parameter	Description ($M_{p_{structure}}$)
H	An object consists of N tissue classes $t_i, i = 1, \dots, N$
\mathbf{x}	$t_i \in [0, 1], \sum_{i=1}^N t_i = 1$
f	Shape: cf. $M_{p_{shape}}$, tissue portion: user-defined, segmentation result
y	$I_{obj,k}$, for tissue class k ($I_{obj,k} \in [0, 1]$)

Approximation of a Continuous Volume Model

After we have decided for an appropriate object shape, we generate a binary object volume for each considered tissue type. A major focus of the considered applications in the following chapters is on quantitative image analysis. Therefore, an accurate approximation of the correct object volume often is an indispensable requirement. Our design process consists of digitized data, so that a good approximation implies a small size of a single voxel with respect to the whole object volume. We concentrate on medical image data sets used in clinical routine or in studies, using a voxel size about ten times smaller than the in-plane resolution, where a typical in-plane resolution is around 1mm. Although this is a rather heuristic choice, it provides a good trade-off between computational complexity and accuracy.

A more formal description of our approach can be given as follows: We generate a high-resolution binary object volume $I_{obj,k}$ for each tissue type k with signal intensity values $I_{obj}(\mathbf{u}), I_{obj}(\mathbf{u}) = 1$ for $\mathbf{u} \in object$, and $I_{obj}(\mathbf{u}) = 0$ otherwise. Thereby, the actual object is obtained on a lower resolution and is then resampled. Finally, the volume is given by the sum over all object voxels

$$V_{I_{obj}} = \sum_{\Theta} I_{obj}(\mathbf{u}), \quad (4.2)$$

V_{Obj}	object volume
Θ	object domain
$\mathbf{u} = (u, v, w)^\top$	voxel position, $\mathbf{u} \in \Theta$
$I_{Obj}(\mathbf{u})$	intensity value $I_{Obj}(\mathbf{u}) \in \{0, 1\}$ at position \mathbf{u}

We typically consider a 512^3 grid for object delineation. Different object volumes can then be easily generated by specifying a different voxel size of I_{Obj} . The resulting parameter module $M_{pvolume}$ is

Parameter	Description ($M_{pvolume}$)
H	The volume can be approximated by voxel counting
\mathbf{x}	High-resolution object
f	Voxel counting approach to approximate the object volume
y	Volume $V_{I_{Obj}}$

Topology

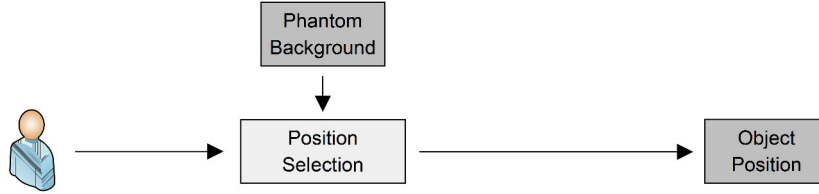
The position of an object within the background is another parameter that has significant impact on the later appearance of the phantom data set. Important attributes are the spatial relation between different objects as well as the actual object position within the background. For example, Multiple Sclerosis lesions predominantly appear in the white matter of the brain. Possible object positions could therefore be restricted to this area. For example, a white matter mask is used for this task in Section 5.2.

In this work, we propose three approaches for object placement. The first method is based on manual object placement. Here, the user selects the final object position (cf. Fig. 4.8 (top row)). To provide this functionality in a structured workflow, we have developed a software assistant, facilitating an interactive manipulation of appropriate positions by simple user interactions. See Section 4.3 for a detailed description.

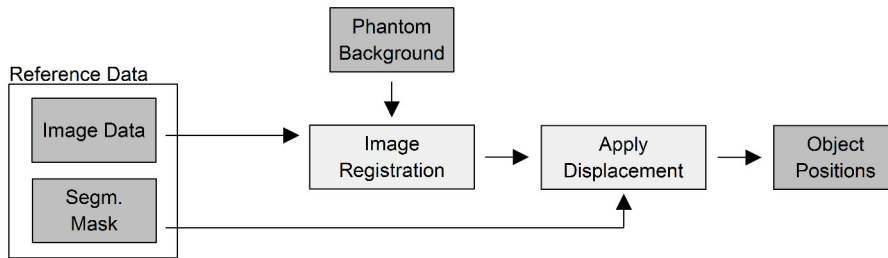
Our second approach uses a reference patient data set plus a segmentation mask of the targeted objects within this data set, e.g., a segmentation of lung nodules. In a first step, the original patient data are aligned with the background image data using an automatic registration algorithm. The resulting transformation is then applied to the segmentation mask so that this information can be used in the coordinate system of the background image. Finally, the center of mass of the segmented objects is used to position objects within the background. See Figure 4.8 (middle row) for an illustration of this approach.

The third approach is related to the previous one. Again, we use a patient data set as reference. However, we do not constrain the segmentation mask to come from a single patient. Instead, we use a number of data sets with segmentations of the targeted structure and create a probability map of cross-subject variability. To this end, we register the training data as well as the segmentations to the reference data. The resulting average map is then used to constrain the spatial object positioning (cf. Fig. 4.8 (bottom row)). This approach is

1. User-defined Object Placement:



2. Object Position from Patient Data



3. Object Position from Probabilistic Atlas

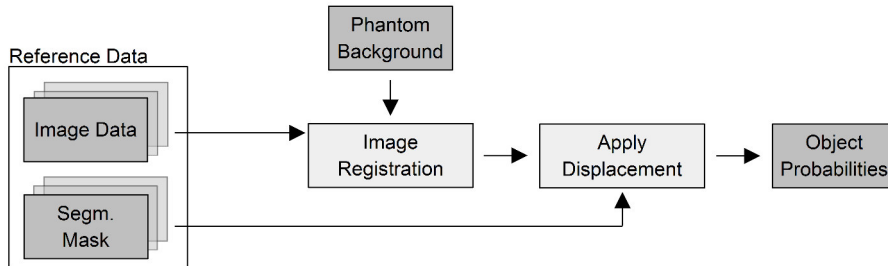


Figure 4.8. Methods for object topology modeling used in this work. (top) User-defined placement of object position, (middle) Object position obtained from reference patient data via registration, (bottom) automatic object placement based on probability map.

used in Section 5.2 to place lesion objects at reasonable positions. The associated parameter module $M_{p_{topology}}$ is described by

Parameter	Description ($M_{p_{topology}}$)
H	Suitable object positions known a priori
\mathbf{x}	Objects and background
f	User-defined or automatically selected object positions
y	List of user-defined object positions within background

4.2.2. Imaging Parameters

The second parameter group is related to imaging parameters and artifacts. Our phantom development process includes the following parameter modules: contrast, noise, PV effects, and resolution.

Determination of Object Intensities

The aim of this parameter module is to generate a volume I_{gv} with intensity values $i_{gv}(\mathbf{x})$ for each modeled tissue type at voxel positions $\mathbf{x} = (x, y, z)^\top$, $\mathbf{x} \in \Theta$. A common approach assumes constant gray values for each tissue class throughout the whole object volume. Although anatomical structures seldom have a uniform intensity on a macroscopic level, this is often a valid assumption for small structures with a compact local appearance in an MR scan. Additional object textures can be added using the methods introduced for the parameter module $M_{p_{structure}}$.

In this work, we use a different intensity model for each tissue class. The models are obtained from a number of training data sets based on segmentations of each investigated tissue class as well as their adjacent anatomical structures. For example, we segment a lesion in the white matter of the brain as well as the adjacent white matter itself (cf. Sec. 5.1). To remove smoothly varying intensity values across the image, an intensity normalization algorithm can be applied to the data, e.g., the N3 algorithm proposed by Sled et al. (1998). The final gray value map I_{gv} for each structure is then obtained using the mean and the intersubject variability as descriptors of the object appearance (Rexilius et al. 2005). The ratio between object and adjacent tissue classes is used as a relative measure for the object intensity value. For example, a hyperintense lesion in the white matter of the brain will result in a high ratio. The parameter module $M_{p_{contrast}}$ for intensity values is given by

Parameter	Description ($M_{p_{contrast}}$)
H	Each tissue class is described by a single intensity value
\mathbf{x}	Number of training data sets
f	Subject-specific mean value of segmented area
y	Resulting gray value map I_{gv}

Noise

Object noise is described by a Gaussian PDF (cf. Eq. 3.2), assuming zero-mean and a standard deviation σ . This results in a new object intensity map \tilde{I}_{gv} with intensity values given as

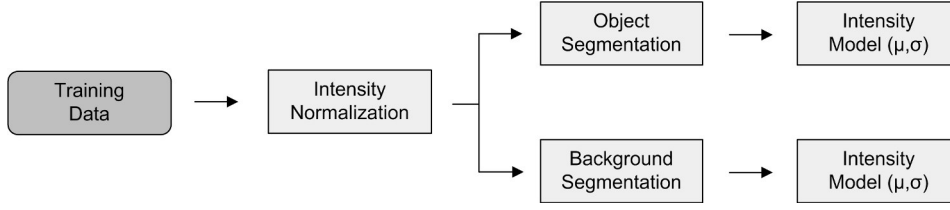


Figure 4.9. Diagram of main components to determine the object intensity values.

$$\tilde{i}_{gv} = i_{gv} + N(\mu = 0, \sigma) \quad (4.3)$$

i_{gv}	initial intensity value
\tilde{i}_{gv}	final intensity value with added noise
$N(\mu, \sigma)$	Gaussian distribution
μ	noise mean value
σ	noise standard deviation.

This parameter module M_{pnoise} is given by

Parameter	Description (M_{pnoise})
H	Noise can be described by a Gaussian PDF
\mathbf{x}	Number of training data sets
f	Combination of intensity value and Gaussian noise (cf. Eq. 4.3)
y	\tilde{i}_{gv}

Partial Volume Effects

One of the major limiting factors for an accurate quantitative analysis are partial volume effects. Especially in structures of the same order of magnitude as the slice thickness of the associated imaging protocol, a dedicated handling of these effects becomes necessary for an applicable phantom.

In order to generate software phantoms with a correct partial volume handling, we propose an approach based on changing the resolution of a binary object volume I_{obj} (Rexilius et al. 2003). Therein, a high-resolution object is downsampled to the same voxel size as the background image using trilinear interpolation, and then reformat the object into the same coordinate system. This results in a probability map $\tilde{I}_{obj} : \Omega \rightarrow \mathbb{R}$ with intensity values $\lambda := \tilde{i}_{obj}(\mathbf{x}) \in [0, 1]$, defining the amount of partial volume at each voxel. A decreasing density can be observed starting from the object core, yielding a typical blurring at the object border. To ensure the correctness of this step, we verify the exact volume of \tilde{I}_{obj} by comparing it with the original volume of $V_{I_{obj}}$. The parameter module M_{ppve} is given by

Parameter	Description (M_{pve})
H	PV effects arise from object downsampling
x	High-resolution binary object I_{obj}
f	Object downsampling
y	Downsampled image \tilde{I}_{obj}

Resolution

A parameter related to the modeling of PV effects presented in the previous section is the object's spatial resolution. Similar to this previous approach, the resolution is defined by object downsampling to the same voxel size as the background image using trilinear interpolation. The corresponding module $M_{presolution}$ is described by

Parameter	Description ($M_{presolution}$)
H	Object resolution is defined by the background resolution
\mathbf{x}	High-resolution binary object I_{obj}
f	Object downsampling
y	Downsampled image \tilde{I}_{obj}

4.2.3. Background Design

In this work, we use volumetric MR data sets $I_{bg} : \Omega \rightarrow \mathbb{R}$ acquired from actual patients or healthy volunteers as background model. Voxel phantoms such as the BrainWeb phantom by Collins et al. (1998) are applied as well. Additional noise can be added similar to the noise model described in Section 4.2.2. A different approach could be a homogeneous background model with a constant gray value. Although this results in a simplified evaluation process, the applicability for the evaluation of image analysis methods is rather low. Today, many image analysis techniques make use of prior information such as the gray value distribution of tissue classes or certain structural characteristics of the underlying anatomy. For example, brain tissue classification schemes are often based on an atlas-based initialization. Furthermore, segmentation methods typically include prior shape information about a regarded structure. In these cases, a software phantom with a homogeneous background model is not an appropriate evaluation tool, due to missing shape and texture information.

4.2.4. Object Incorporation

The final step of our phantom design approach combines the determined object and the selected background model. More specifically, we incorporate the object volume \tilde{I}_{gv} , containing appropriate gray values for each tissue class into the background I_{bg} using a linear weighting function. The amount of partial volume at each voxel, derived from the high-resolution binary object volume as described in Section 4.2.2, serves as weighting factor. A

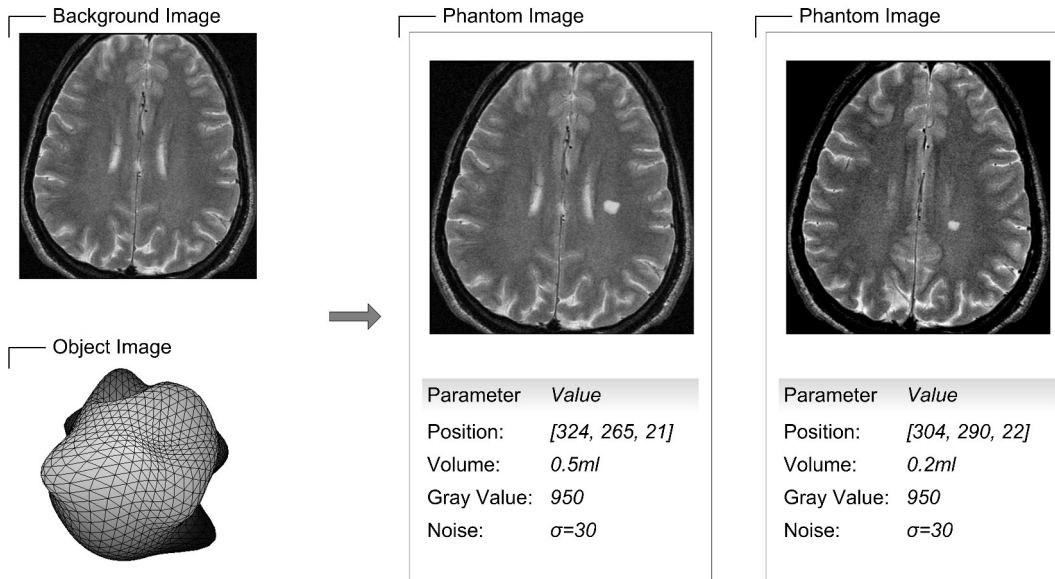


Figure 4.10. Illustration of incorporation step. Two different resulting phantom images as well as the corresponding parameter settings are presented with different object positions and volumes. Please note that the selected object positions can also denote different slices within the data set. In this example, slice number 21 and 22 are used.

phantom data set I_p is then generated by a convex combination at each voxel defined as:

$$i_p = \sum_{k=1}^K (\lambda_k \cdot \tilde{i}_{gv,k}) + (1 - \sum_{k=1}^K \lambda_k) \cdot i_{bg} \quad , \quad (4.4)$$

i_p	intensity value of resulting phantom data set
λ_k	object probability map for tissue class k (cf. Sec. 4.2.2)
$\tilde{i}_{gv,k}$	object intensity value for tissue class k
i_{bg}	background intensity value
K	number of modeled tissue classes

with $\sum_{k=1}^K \lambda_k = 1$. This technique is sometimes also called 'alpha blending' (Shin et al. 2006).

Figure 4.10 provides an overview of the construction process, showing the original background and the object image as well as resulting phantom images and their parameter settings. Different object positions and volumes are used to demonstrate flexibility of our approach.

4.3. A Software Assistant for Interactive Parameter Modeling

The concretization of our hybrid phantom design approach by explicit models for a number of modules gives us the building blocks required to develop a software assistant for the design of software phantoms (Rexilius et al. 2008). The following sections provide an overview of the underlying concepts as well as the different processing tasks.

4.3.1. Design Concepts

Our goal is an easy-to-use tool that can be applied to many different tasks occurring during the design and evaluation of new algorithms in medical image analysis. This way, researchers can generate a data basis for evaluating performance and limitations of their own algorithms. Moreover, the exchange of data sets between research groups is supported, enabling a standardized and objective validation with a set of reference data.

An important aspect for the success of such a software assistant is a clear and easily operated graphical user interface. Our software assistant consists of three different processing steps, related to the design tasks in Section 4.1, namely object design, background design, and object incorporation. For each step, a separate component has been implemented, facilitating a hierarchical encapsulation of the underlying methods. To enable rapid prototyping of software assistants, our software is designed as part of the modular development platform MeVisLab (MeVisLab 1.5).

In addition to the actual graphical user interface, a common data scheme for the description of the resulting image and meta data is an important requirement, e.g., noise variance, voxel size, object volume. Our software supports several image data formats including DICOM, DICOM/TIFF, or the Analyze file format. Further information, e.g., about object intensity values or the object volume, are stored in a separate data structure using the Extensible Markup Language (XML).

4.3.2. Processing Steps

Object Design

The main functionality of the first processing step is the definition of object shapes that are appropriate for the considered application. Ten predefined shapes including spheres and ellipsoids of different geometries can be used. Object shapes that have been used in previous work are available as well (Rexilius et al. 2003; Rexilius et al. 2005). This allows for a direct comparison of new analysis methods with a comparable data pool. Besides the already available shapes, own object files can be loaded as described below. Furthermore, new object shapes can be generated by interactively modifying the surface of a currently selected object. The user interface with an exemplary parameter setting is illustrated in Figure 4.11.

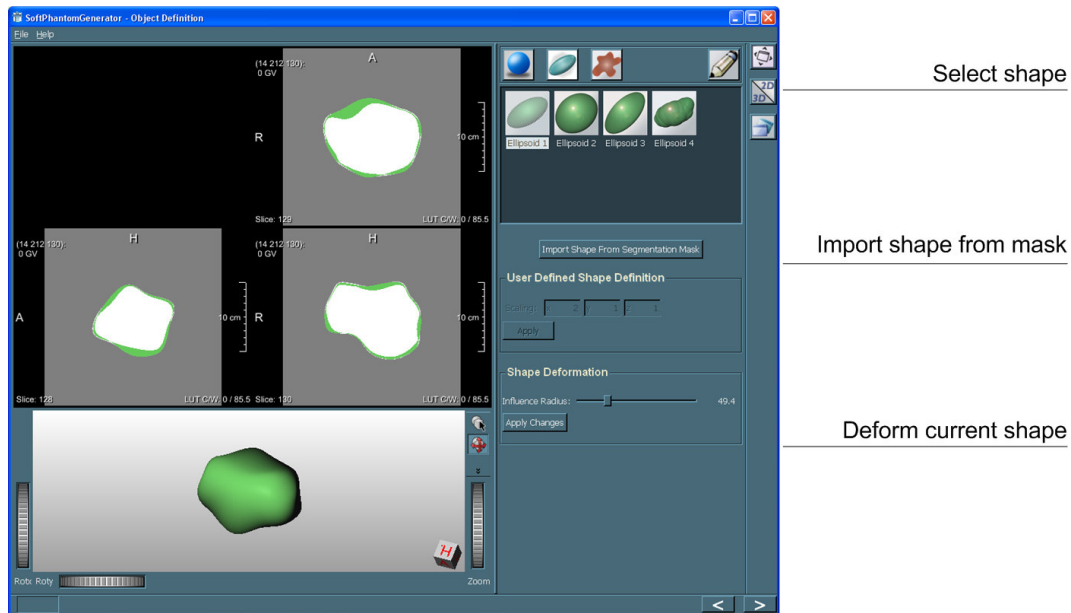


Figure 4.11. First step of our software assistant: Object design.

The software assistant provides 2D as well as 3D representation of a selected object shape. The 3D representation of the current object is based on a winged-edge mesh (WEM) data structure. An implementation of a comprehensive WEM library is readily accessible as part of the standard MeVisLab package.

To tailor a given object shape to the application in mind, the software assistant allows for an intuitive deformation of the object surface. Thereby, the user can interactively select an arbitrary viewer position close to the surface. All nodes in the neighborhood of this dragging position are then influenced by subsequent mouse movements to a certain degree. This way, different local and global object deformations can be achieved, depending on the set influence radius. See also Figure 4.6 for an example.

Defining new WEM object shapes

An important feature for a broad applicability of our software assistant is the ability to use not only the predefined objects that come along with the software, but also new ones. These could be user-defined shapes or even shapes defined by other researchers. So how to import a new object into the application?

In order to generate a new WEM structure, we developed an additional application that can be selected via button from the parameter section of the current object definition step. This application facilitates the design of new shapes from a binary image, e.g., from a manual segmentation of a tumor. We extract an iso surface from the loaded segmentation mask, which stores the result in a WEM. In a subsequent step, a simplification of the mesh is per-

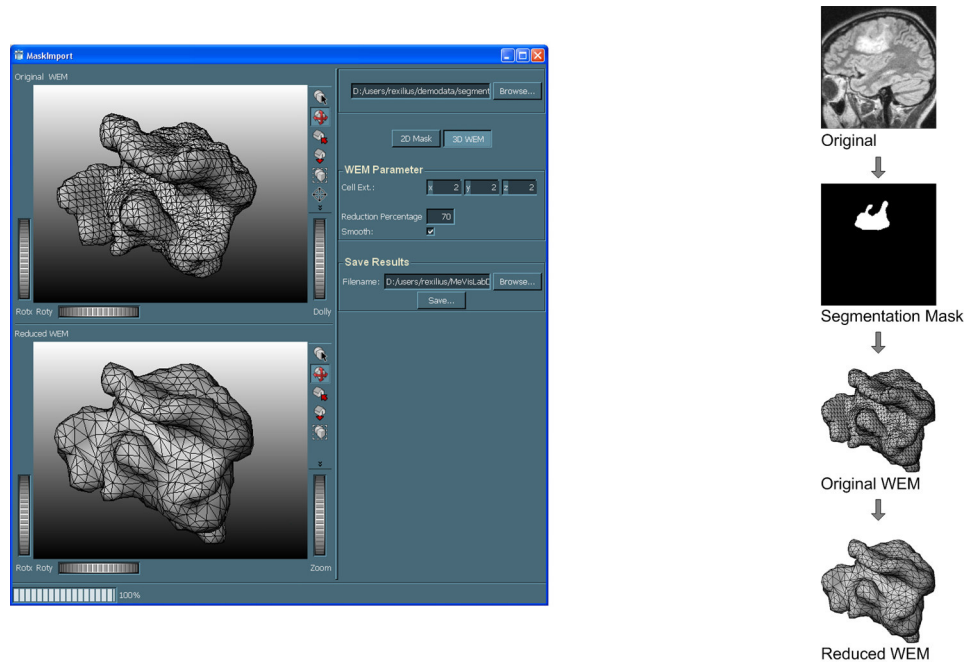


Figure 4.12. Integrating own object shapes from binary segmentation masks. (left) User interface, (right) overview of algorithm steps.

formed, employing a multi-pass edge collapse algorithm. Thereby, all edges are sorted in a priority queue according to an angle criterion in each pass. Then, a sequence of edge collapse transformations is applied. The resulting WEM preserves visually important features of a mesh and greatly reduces the required space for saving the data. Finally, an optional surface smoothing can be applied to the mesh. Figure 4.12 (right) gives an overview of the performed processing steps. The proposed user interface of this application is shown in Figure 4.12 (left).

Background Design

We propose two different approaches for the background design: An image with a constant gray value and predefined image size and voxel size, as well as other, more problem specific image data. Additional Gaussian noise can be added to a selected background. A screenshot of the user interface of this step is presented in Figure 4.13.

Object Incorporation

The last processing step of our software assistant combines object and background. Therefore, the selected object has to be resampled to the same voxel size as the background data set, allowing for an incorporation into this data set at a suitable position. To enable a flexible

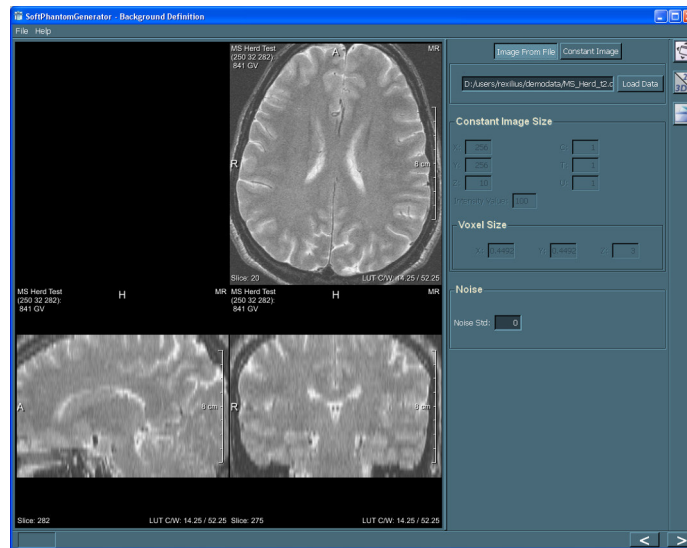


Figure 4.13. Second step of our software assistant: Background design.

workflow, this step facilitates a number of parameter adjustments. An illustration of the user interface is given in Figure 4.14.

Two general workflow scenarios can be distinguished: Either both object and background have been selected in the previous two steps, or an object is selected directly in this last step. For the first case, we have to define a suitable voxel size first – typically about ten times smaller than the smallest resolution of the used background image. An isotropic voxel size that fulfills these demands is automatically computed as a first guess. Then, the user has to define an object volume and the object is downsampled to meet this value. Because the object image data has a high resolution, this is the most time consuming step of the whole application. To avoid any resampling, we added an additional feature that permits the determination of an object shape and volume from previously processed objects. Thereby, an important prerequisite is an equal voxel size of the background data set and the downsampled object data set. Otherwise, the incorporated object volume would not be correct. This constraint is automatically verified for all data sets.

After object downsampling, we can incorporate it into the background. An appropriate object position within the background can be chosen interactively. The amount of partial volume at each voxel serves as weighting factor as described in Section 4.2.4. An extra object shift by up to one voxel within the (x,y,z) -direction can be applied, causing additional partial volume effects. In a final step, proper object intensity values as well as a suitable noise level have to be selected by the user. We also added the possibility to integrate more than one object into the current background.

The resulting software phantom image data can be saved using the MeVisLab-specific DICOM/TIFF format. Further general as well as data-specific information are stored into

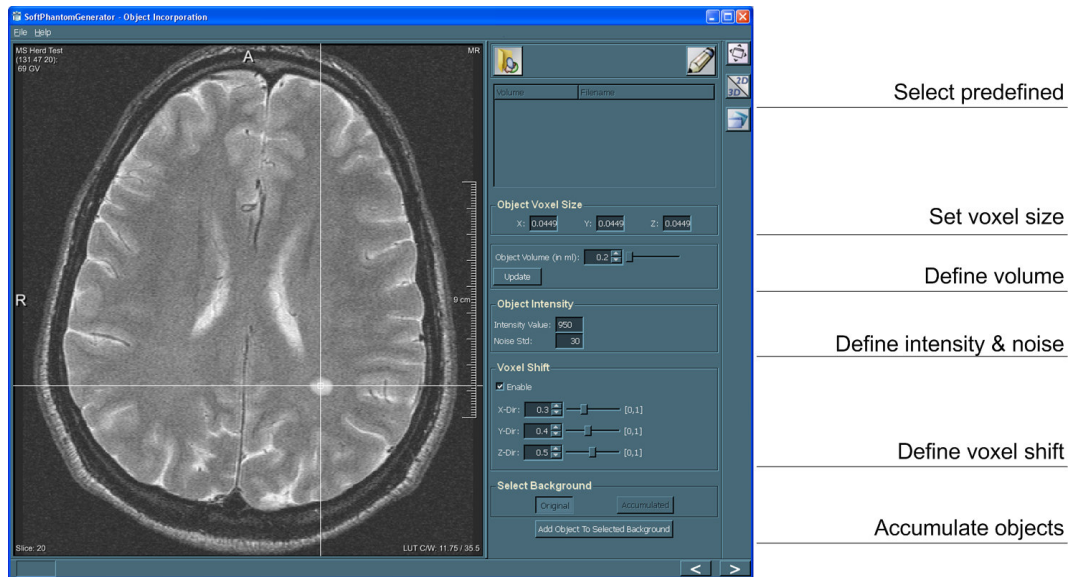


Figure 4.14. Third step of our software assistant: Object incorporation.

an XML structure. Especially the characteristics of an included object are important. We retain information about the volume of each object, its exact position within the background, as well as the gray value and the amount of noise. An evaluation by four field experts with respect to applicability resulted in an overall good rating.

Processing Time

Our software assistant provides an efficient tool that reduces the total time required for phantom design from hours to minutes, depending on the lesion complexity and the number of required lesions in the phantom.

Object Selection For object selection, the user can select an available shape from a list. Nevertheless, this step can become rather time-consuming, if additional manual shape deformations are performed. Although the user is supported in this task, adding several large and small deformations can take up to five minutes. The object design step should be repeated several times and the resulting objects should be stored in advance.

Background Selection This step mainly consists in selecting an appropriate background data set plus additional noise if required. No further processing is done. Therefore, users typically need less than a minute for this step.

Incorporation Several parameters can be changed in this last step such as the voxel size, the object intensity value or the noise level. Furthermore, several objects can be accu-

ulated within one phantom. For a phantom with only a single object, this processing step takes no longer than 2 minutes.

4.4. Discussion

Standard reference data sets are an important tool for the development and evaluation of algorithms in medical image analysis. Furthermore, they can serve as basis for discussions with clinical experts, e.g., about accuracy and precision of current measurement tools. Especially software phantoms of anatomical and pathological structures provide a flexible and cost-effective approach for this task.

Our overall goal is an efficient and reusable design of hybrid software phantoms. To this end, we proposed an approach that allows for a generic specification of phantoms. An integral part is a set of modules providing an in-depth description of parameters used during phantom design. Based on this description, we then presented modeling schemes for various parameters that will be used in the following chapters for software phantom design. This provides the building blocks to develop a software assistant that allows us to combine several parameters in an interactive fashion, adding additional flexibility to the phantom design process.

4.4.1. Modular Phantom Development

In Chapter 2, we first introduced phantoms based on very general design properties, followed by a description of relevant parameters in Chapter 3. Now we focus on how to provide an easy-to-use phantom description. We developed a template data sheet that provides an easy-to-use phantom description and allow for a quick overview of the most important phantom parameters. Figures 4.3-4.5 exemplarily illustrate this description for three phantoms. Both, physical and software phantoms are presented to show the wide applicability of our approach.

Our phantom description can be categorized into three basic steps: Object design, background design, and incorporation. Each step includes a task-specific description of parameters used during phantom development. Thereby, major parts of such a parameter, which we also call a module, are input and output values as well as a transition function to transform the input. Furthermore, a module is based on a certain hypothesis presenting an understandable description of the underlying assumptions. For example, Ko et al. (2003) assume a spherical object appearance in their phantom.

Phantoms have become an inevitable tool for algorithm development and evaluation in medical image processing. Nevertheless, a standardized analysis that allows a comparison between different design approaches has not been proposed yet. Therefore, some publications focus on a domain-specific literature survey of phantoms. For example, Lee and Lee (2006) present a review of phantoms related to radiation dose distribution calculations in the human body. Price et al. (1990) develop a set of standard physical phantoms dedicated

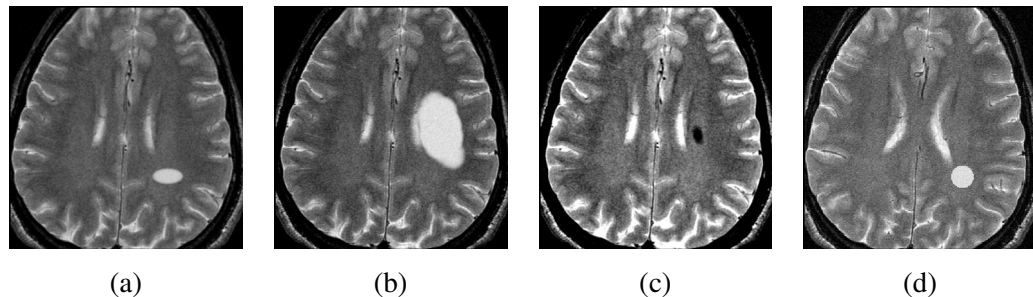


Figure 4.15. Examples of software phantoms for hyperintense brain lesions with wrong or missing object modules. (a) Oversimplified object shape, (b) unrealistic object volume, (c) wrong object intensity value, (d) no consideration of partial volume effects.

to the evaluation of a single imaging parameter such as spatial resolution or uniformity. Unfortunately, these paper merely provide a summary of recent work. Moreover, related work in the field of image analysis as well as an explicit categorization of phantoms has not been proposed yet.

In this work, we present a method that is suited to describe any phantom, and some examples have been given in this chapter. To the best of our knowledge, we provide the first approach towards a comprehensive description of phantom design processes in medical image analysis. Furthermore, we propose a standardized analysis for both physical and software phantoms, that enables a description of modules and of their interactions. We adopt a notation proposed by Udupa et al. (2006), where the domain of a segmentation method is determined by three entities: Task, body region, and imaging protocol (cf. Sec. 8.6.1).

Introducing a general description of the phantom design process also allows direct comparison of different phantoms. For example, our approach enables a quantitative analysis, e.g., about the number of used modules. Furthermore, we can carry out a qualitative analysis of the modules used for phantom design. Thereby, the enclosed field of application allows for a domain-specific comparison. To give an example, Figure 4.15 shows four software phantoms for small hyperintense brain lesions. In each picture, at least one module is not within the correct parameter range or is based on wrong assumptions. For example, an oversized object volume (cf. Fig. 4.15 (b)) or a wrong object intensity value (cf. Fig. 4.15 (c)). Nevertheless, a statement about the actual phantom quality remains difficult. To this end, we propose a new method for phantom validation in Part II of this work.

4.4.2. Our Phantom Approach

Today, phantoms are often based on simple geometric objects with known volumes, especially when using physical phantoms. Additionally, these objects usually consist of a single material or tissue type. In Chapter 2 we introduced three different types of software phantoms: stylized phantoms, voxel phantoms, and hybrid phantoms. We found the last category

to be most suited since these phantoms combine several advantages of the other two classes (cf. Fig. 2.6). In this chapter, we focused on a hybrid method that incorporates objects into tomographic image data of a patient or a volunteer. Based on our modular design approach, we modeled various parameters that are typically used during phantom development. For example, we propose different methods for the development of a suitable object shape. Further aspects of our phantom are an application-specific parameter modeling and a range of suitable values, which were identified as important issues for phantom parameter modeling (cf. Sec. 3.4).

Used Modules

In Chapter 3 we already discussed a number of characteristic modules, which are frequently used in medical image analysis or represent a common choice in scanner quality assurance programs. Our phantom design uses several of these modules to enhance the acceptance of our approach. For example, we assume a commonly used noise model described by a Gaussian PDF. Nevertheless, we also introduce extensions to current modeling approaches such as an interactive method to deform a given surface that is used to change a given object shape. Furthermore, we propose different methods to define the object topology. These are not only based on the common manual object placement at '*typical positions within the human body*', but also obtain a suitable position from reference data such as patient data as well as from probability maps of cross-subject variability. Another specific feature of our phantom is a high-resolution binary object that is used to determine an accurate volume measurement and to model partial volume effects via downsampling. Other than the object design, the background does not contain any concrete parameter modules. Here, we use volunteer image data where the underlying ground truth is not available for most parameters. A similar difficulty occurs for example for imaging parameters in physical phantoms as already discussed in this chapter.

To summarize, we developed specific models for eight parameter modules divided into two categories:

Category	Modules
Morphology and Topology	Shape, Structure, Volume, Topology
Imaging	Contrast, Noise, Resolution, PV Effects

For each module, we provided a detailed analysis plus a tabular summary that consists of input and output values, applied transition functions, and the underlying hypothesis. Compared to the number of relevant parameters introduced in Chapter 3, only *three* are not included in the above list.

Uniformity. A model for intensity nonuniformity will be available for the phantom proposed in Section 5.2. Therein, we use the bias field available from the BrainWeb project.

Motion. The second imaging parameter not included in our modeling approach is a module for motion artifacts, because it is simply not required for our targeted applications.

Process. A general approach for modeling dynamic processes for phantoms is not possible. However, we propose dedicated modules for growth and a simulation of contrast agent enhancement characteristics in the context of brain tumors in Section 5.3 of this work.

Several phantom results based on our modeled parameters are shown in Chapter 5. For example, Multiple Sclerosis lesion phantoms (cf. Fig. 5.2) or brain tumor phantoms in Figure 5.13. To the best of our knowledge, our approach represents the first hybrid software phantom with this amount of object modules.

5. Phantom Examples

In the previous chapter, we proposed new modeling schemes for parameters commonly used in phantom design. This finally enables us to develop our own phantoms. In this chapter, we present three examples. For each phantom, the parameter models proposed in the previous chapter are used to generate a phantom. Extensions are described and examples of the resulting software phantoms are given. Furthermore, the template sheet introduced in the last chapter is applied to provide a compact description and a quick overview of the most important aspects of all phantoms.

Our first example describes the design of phantoms for Multiple Sclerosis (MS) lesions (cf. Sec. 5.1). Therein, an interactive object design of typical lesion shapes incorporated into an MR dataset of a healthy volunteer is used. The software assistant proposed in Section 4.3 is used to design the final phantoms. Section 5.2 then introduces an extension of this manual approach. Here, we propose a fully automatic method to build an arbitrary number of MS data sets. The BrainWeb phantom (Collins et al. 1998) is used as reference data set.

The third example is a phantom for brain tumors (cf. Sec. 5.3). Here, we introduce two modeling scheme for process modules. A biomechanical model is used to simulate tumor growth and the associated deformations inside the brain. Furthermore, a simulation of contrast agent enhancement characteristics is presented. Compared to the lesion objects above, we also develop additional tissue classes to model the amount of edema and active tumor. Again, our software assistant is used to generate the phantoms.

5.1. Interactive Design of MS Lesion Phantoms

Multiple Sclerosis (MS) is a chronic disease affecting the central nervous system (CNS), which includes the brain and spinal cord. It is one of the most common neurological diseases in Central Europe, predominantly affecting young and middle-age adults. Currently, approximately 130,000 persons have been diagnosed with MS in Germany. Around 2,5 million people are affected worldwide. The etiology of MS is still unknown, but current studies point out a multifactorial genesis. A high degree of variability and diversity characterize the clinical signs and symptoms of the disease, that histo-pathologically results from a progressive demyelination and an axonal loss within the CNS (Lassmann 2002).

Currently, no predictable pattern is known for presentation of patients with MS, which makes the diagnosis clinically challenging. Most patients with MS initially have a relapsing-remitting (RRMS) subtype characterized by episodes of neurological dysfunction followed by periods, during which a stabilization or a partial to full recovery of the symptoms can be observed. Over time, the initially relapsing-remitting MS gradually worsens often into a more progressive course of the disease, called secondary progressive (SPMS). Another subtype that can be distinguished is the primary progressive MS (PPMS), characterized by a gradual progression from its onset with no remissions at all (Inglese 2006).

Several therapeutic options are available to patients to treat the disease symptomatically. However, no curative treatment is available for Multiple Sclerosis today, so that there is a considerable demand to develop new and more effective drugs.

Diagnosis of MS and the Role of Magnetic Resonance Imaging

Multiple Sclerosis can be very difficult to diagnose at first and there is no simple test because of the high variability in signs and symptoms that may suggest a number of conditions. Commonly used methods of quantifying MS are neurological disability scales such as the Expanded Disability Status Scale (EDSS) (Kurtzke 1983). During the last years, MRI has become an important imaging modality for understanding and managing several aspects of MS. Today, it is an integral part of standard diagnostic especially for follow-up examinations and plays a primary role as a surrogate marker of drug efficacy in clinical trials (Miller et al. 1998). In more than 95% of patients with MS, abnormalities can be seen on MR images (Compston and Coles 2002). To make use of the advances in MRI techniques, an international panel was convened and recommended revised diagnostic criteria for MS, known as McDonald criteria that formally incorporated MRI (McDonald et al. 2001; Polman et al. 2005).

Typical multispectral image acquisition protocols include proton density (PD-), T2-weighted, and FLAIR (fluid attenuated inversion recovery) sequences, as well as T1 sequences pre and post contrast with a slice thickness of 3-5mm (Traboulsee et al. 2003). T2- and PD-weighted images are very sensitive in detecting MS lesions, where lesions appear as areas of increased signal intensities, predominantly in the white matter of the brain. Chronic abnormalities such as axonal degeneration and the presence of demyelination can be observed as hypointense lesions on T1-weighted images ('black holes'). Enhancing lesions ('active lesions'), found on T1-weighted images after contrast administration, show the inflammatory stage of the disease. Besides established imaging sequences, a number of other modalities have gained attention in the last few years, including MR-Spectroscopy (MRS), functional MRI (fMRI), high-resolution morphometrical imaging combined with accurate atrophy measurements, or diffusion weighted imaging (DWI) and diffusion tensor weighted imaging (DTI) (Miller et al. 1998). Each sequence offers a specific view on different aspects of the disease and each radiology department typically has its own standard protocols.

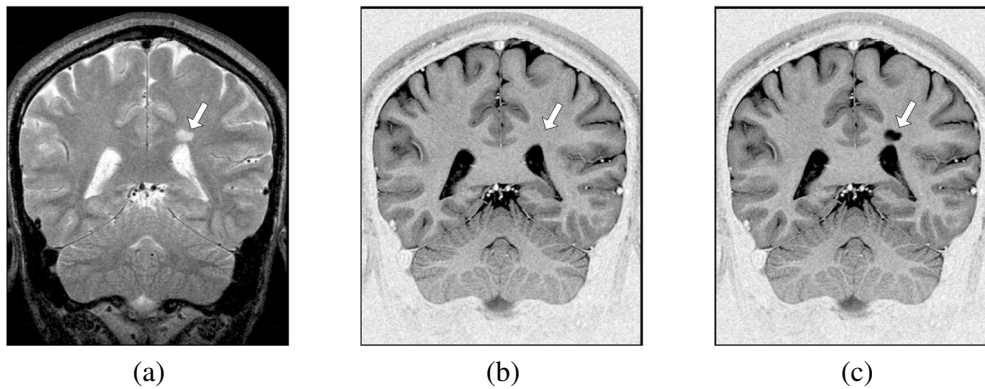


Figure 5.1. Modeling of different lesion types. (a) Standard T2 hyperintense lesion, (b) same volume and position of lesion as in (a) on a T1-weighted image; (c) same lesion object as in (b), but now hypointense on T1-weighted image (black hole).

In this section we develop MS lesion phantoms, that we introduced first in (Rexilius et al. 2003). We propose a manual object design of typical lesion shapes. The lesion objects are incorporated into an MR dataset of a healthy volunteer. In Chapter 6 we will then reuse the resulting software phantoms to investigate the quality of visual assessment in MS lesion volumetry as well as for the evaluation of MS lesion quantification and segmentation approaches.

5.1.1. Morphology and Topology

Shape and Structure

To cover a variety of realistic-shaped MS lesions, we generate different lesion objects such as sphere-like objects and ellipsoidal objects with several deformations. Thereby, we assume that a lesion consists of a single homogeneous tissue class. Each object is constructed from a combination of geometric primitives. To this end, ellipsoids of different size and shape are placed at manually selected positions on the object surface in an iterative fashion. A similar approach has also been used for the design of lung nodule phantoms (Shin et al. 2006).

Volume

Each lesion phantom is defined on a 512^3 grid, i.e., a lesion volume is defined as the number of voxels inside this grid. Different volumes then can be easily generated from a lesion phantom by specifying a different voxel size. We use a voxel size ten times smaller than the inplane resolution of the available MR scan.

Position

The lesion objects are generated manually and placed at typical paraventricular positions in the white matter of the brain. See Figure 5.2 for examples.

5.1.2. Imaging Parameters

Gray value and noise

A lesion object with a reasonable constant intensity value for MS lesions for each available sequence is created, and Gaussian noise is added. The standard deviation of the Gaussian noise is set approximately equal to the noise of the brain scans, estimated from unstructured regions inside the white matter for each available imaging sequence separately. A similar approach is used to generate appropriate lesion gray values. Mean gray values for white matter, as well as for lesion tissue, are computed manually from several patient data sets, and the ratio is used as guideline for the phantom lesion gray values. Different ratios are computed for the available MR sequence. Furthermore, phantom lesion gray values are adjusted based on inspections of patient data sets with MS lesions.

Resolution

The spatial resolution of the images is determined by the underlying reference data.

PV Effects

To account for partial volume effects, we resample the lesion mask using trilinear interpolation after modeling all morphological parameters, i.e., shape, structure, and volume. The resulting probability map defines the amount of partial volume at each voxel.

5.1.3. Background Design and Object Incorporation

The images used as background in this study were acquired from a healthy volunteer (a 28-year-old male) on a 1.5 T scanner (Magnetom Vision; Siemens, Erlangen, Germany). The data-acquisition protocol contained axial and coronal PD-, T2-, and T1-weighted images with an in-plane resolution of $0.449 \times 0.449 \text{mm}^2$ and a slice thickness of 3 mm, matrix of 512×512 , and 34 axial and 51 coronal continuous slices, respectively. All images were acquired in one session with head fixation and without table movement, such that all data sets are perfectly aligned without visible motion artifacts.

The final phantom values $i_{phantom}$ are computed by linear combining all structures available within a voxel, i.e., $i_{phantom} = \lambda \cdot i_{lesion} + (1 - \lambda) \cdot i_{WM}$, $\lambda \in [0, 1]$ (cf. Eq. 4.4).

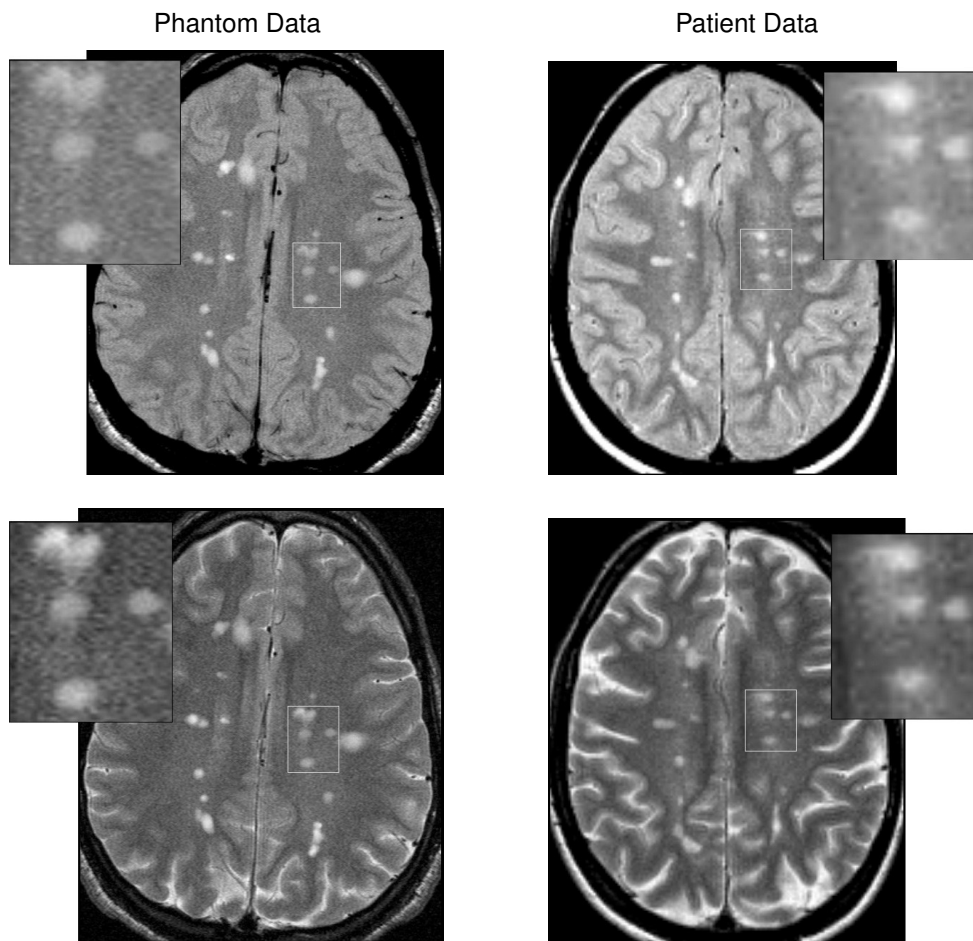


Figure 5.2. Examples of different generated phantom MS lesions showing the potential of our approach on axial PD- and T2-weighted axial images. (Left column) MR scan of a healthy volunteer used in this work with different incorporated MS lesion phantoms, (Right column) a patient's MR scan with several MS lesions.

5.1.4. Resulting Software Phantoms

An summary of the resulting phantom using the proposed template is given in Figure 5.3. Furthermore, Figure 5.1 and 5.2 illustrate the potential of our approach. A comparison of a patient MR scan with several MS lesions and a modeled data set with corresponding lesion types is shown in Figure 5.2. Each lesion object is placed at a position in the volunteer scan that roughly corresponds to a lesion appearance in the patient scan.

Figure 5.1 demonstrates the flexibility of our phantoms. Different lesion types can be generated by simply changing the intensity value of an object. A typical hyperintense lesion within a T2-weighted data set is shown in Figure 5.1 (a). Figure 5.1 (b) and (c) show the

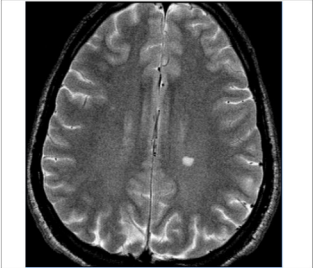
Phantom Description	
Author :	J. Rexilius et al.
Reference:	J. Rexilius et al. "Evaluation of Accuracy in MS Lesion Volumetry Using Realistic Lesion Phantoms". Academic Radiology 12, 17–24, 2005.
	
Application Domain	
Phantom Type :	Software
Task :	Compare lesion quantification schemes
Body Region :	Head
Imaging Protocol:	1.5T MR, 3mm slice thickness
Main Steps	
Object	
Hypothesis :	Compact, high-resolution object of user-defined shape
Parameters :	Shape, Structure, Volume, Topology, Contrast, Noise, Resolution, PV Effects
Background	
Hypothesis :	MR scan of healthy volunteer
Parameters :	-
Incorporation	
Hypothesis :	Voxel-wise combination of object and background gray values
Parameters :	-

Figure 5.3. Summary of the developed MS lesion phantoms.

same lesion in a T1-weighted image with varying intensity values resulting in hyper- and hypointense lesions.

5.2. Automatic Design of MS Lesion Phantoms

A major drawback of the lesion phantoms introduced in the previous section is the overall manual design. Each object shape is constructed from a single geometric primitive with manually added ellipsoids at various positions on the object surface. A feasible object position within the background data is then selected by hand. Furthermore, we assume that a lesion consists of a single homogeneous tissue class, which does not allow to change the object texture.

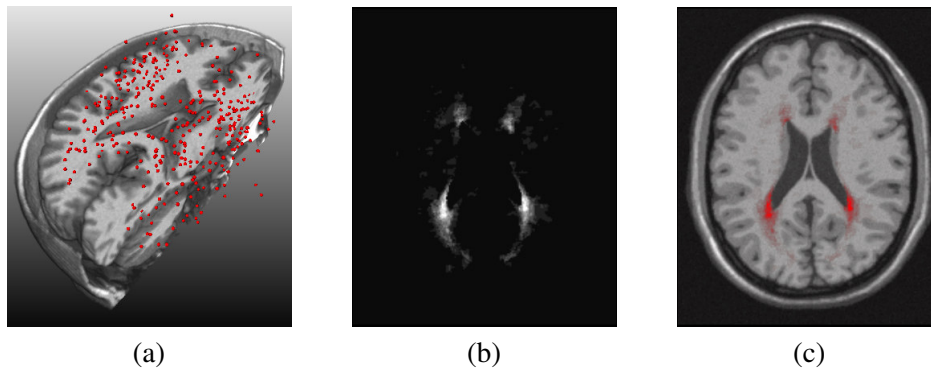


Figure 5.4. Visualization of lesion positions within reference data. (a) The center of gravity (red points) for all reference lesions; (b) statistical position map for one slice; (c) position map overlay on T1-weighted image slice.

In this section, we introduce an extension of the manual approach (Rexilius and Tönnies 2014b). Instead of developing a small amount of hand-crafted lesions, we present a fully automatic method to build an arbitrary number of MS data sets. Several parameters are modeled for each lesion such as shape, volume, or position. The BrainWeb phantom (Collins et al. 1998) serves as reference data set. Our approach can thus be seen as extension of the BrainWeb data by an additional MS lesion class.

5.2.1. Morphology and Topology

Position

MS lesions have a variety of typical locations. In a pre-processing step, we compute a statistical map of lesion positions from a list of patient data sets (cf. Fig. 5.4 (a)-(c)). The BrainWeb data are used as reference coordinate system. The underlying registration process can be described as follows: In the first step, an automatic brain extraction is performed for each patient data set based on an evolving deformable model (Smith 2002). This reduces the degrees of freedom for the following registration steps. Moreover, no information is eliminated since all lesions are located inside the brain. In the next step, we perform a global affine registration followed by a local nonrigid B-spline registration (Klein et al. 2010) to reduce inter-patient shape variabilities. The final map of lesion positions is computed by applying the resulting transformations to the corresponding manual segmentations. An example of the registration results shown as overlay on the BrainWeb data is given in Figure 5.5. This map is then used to automatically place a lesion within the BrainWeb data, using randomly selected positions from the map. Furthermore, a minimum distance between two lesions can be defined.

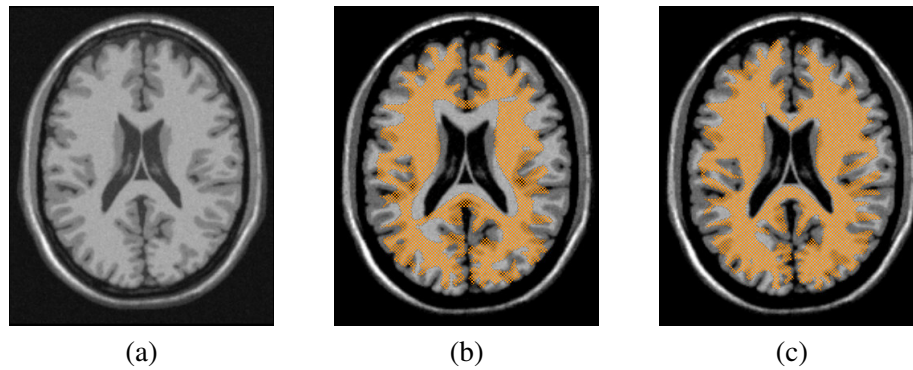


Figure 5.5. Registration results. (a) Reference data. (b) Reference data with patient data overlay (affine registration). (c) Reference data with patient data overlay (affine + B-spline registration).

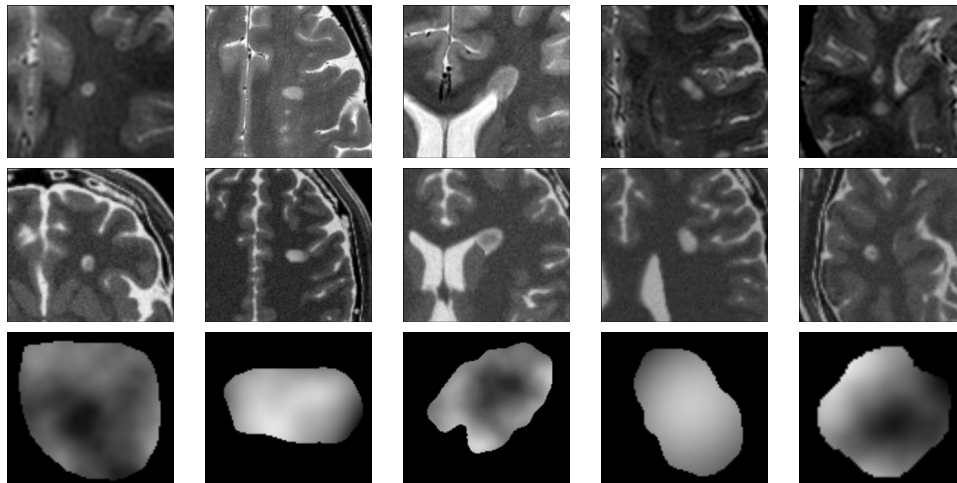


Figure 5.6. Modeling lesion texture using Perlin Noise. (top row) patient data; (middle row) phantom with incorporated lesion; (bottom row) synthetic lesion mask with texture.

Shape

Similar to the lesion positions, their shape also shows a considerable intra- and inter-patient variability. Thus, instead of developing a few handmade shapes, a database consisting of all lesion objects extracted from the registered segmentation masks, as described in the previous section, is used. To reduce the effect of manual segmentation errors, an isosurface is computed from each segmentation mask and further modified by applying a smoothing operation. The result is then transformed back to image space.

Structure

Another important parameter to describe the appearance of an MS lesion is related to its underlying structure or texture. Histologically, MS is characterized by different processes

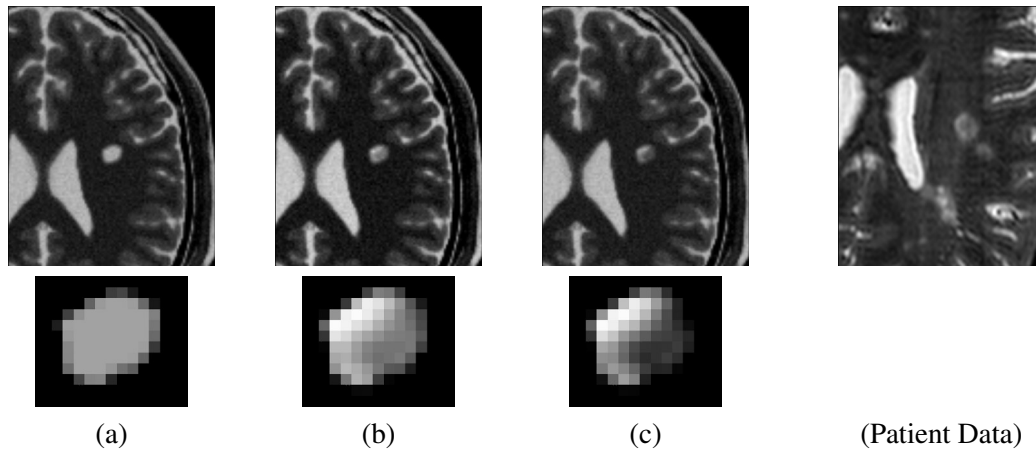


Figure 5.7. Example using scaling of lesion texture. Top row: T2-weighted data with incorporated lesion. Bottom row: corresponding lesion probability mask. (a) Homogeneous lesion; (b) texture values $\in [0.3, 1]$; (c) texture values $\in [0, 1]$.

including inflammation, myelin breakdown, and gliosis. This results in a large heterogeneity in appearance, which is further increased by changing imaging and scanner parameters between data acquisitions. Nevertheless, current methods for MS lesion phantoms use only a single homogeneous tissue class (Tofts et al. 1997; Rexilius et al. 2005).

One way to generate a reasonable lesion structure is texture synthesis from a sample database extracted from patient data, as proposed for brain tumors in (Prastawa et al. 2009). However, we aim at a high-resolution lesion object that is later downsampled to model partial volume effects during image acquisition. A texture extracted directly from patient data can hardly meet these requirements.

In this work, we propose a parametric model based on 3D simplex noise introduced by Ken Perlin (Perlin 2002) to generate different lesion textures. Such an approach is frequently used for example for organ surface-like textures in surgery simulators. Changing the parameter values, i.e. the number of octaves, frequency, and persistence, results in a wide range of different texture samples. Figure 5.6 shows a comparison of different real lesions and corresponding phantom results. Furthermore, the underlying lesion mask is given for each example.

After determining an adequate lesion structure, the resulting texture map is scaled to $[v_{min}, 1]$. Figure 5.7 shows a lesion with an example texture incorporated into the reference BrainWeb data. Different values $v_{min} \in \{1, 0.3, 0\}$ are used to demonstrate the flexibility of our design approach.

Volume

The lesion volume is computed by accumulating all segmented voxels within the segmentation mask. The amount of lesion tissue is used as weight at each voxel, accounting for

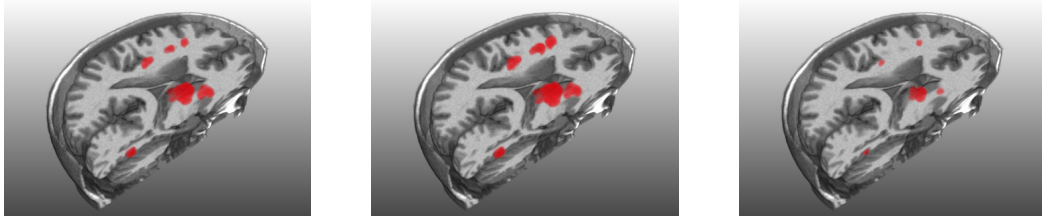


Figure 5.8. Result of automatic lesion volume selection. (left) Random selection of volume $v \in [minV, maxV]$. (middle) $v = maxV$. (right) $v = minV$.

both partial volume effects and textural changes. Moreover, the final volume is randomly selected from a range depending on the original volume. Figure 5.8 shows the result of this approach for different lesions. We use the maximal (Fig. 5.8, middle) and minimal (Fig. 5.8, right) possible volume as well as a random volume selection (Fig. 5.8, left) within the given range for each lesion. This way, the same reference lesion will receive slightly different volumes in different phantom data set. Another approach could be for example to use only lesions from the database that exceed a certain volume threshold.

5.2.2. Imaging Parameters

Gray value and noise

The lesion gray value is defined via the ratio between mean white matter and lesion gray values, extracted from the reference patient data sets. This way, a typical range of values is defined. The final lesion value is then randomly selected from this range. In addition to a gray value, Gaussian noise is added to each lesion. The noise standard deviation is set approximately equal to the reference data noise, selected from a homogeneous region within the white matter.

Resolution and uniformity

The spatial resolution of the images is determined by the underlying BrainWeb reference data. Several inhomogeneity fields are also available from BrainWeb and can be used to add further challenges to the phantom data.

PV Effects

We use the same method that was already proposed for the manual lesion design, i.e., re-sampling the final lesion mask using trilinear interpolation. The resulting probability map defines the amount of partial volume at each voxel.

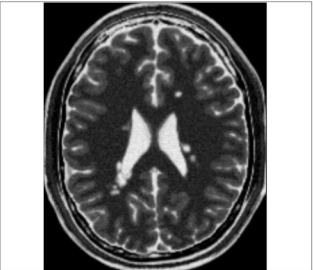
Phantom Description	
Author :	J. Rexilius and K. Tönnies
Reference:	J. Rexilius and K. Tönnies "Automatic design of realistic Multiple Sclerosis lesion phantoms". Workshop Bildverarbeitung für die Medizin, 270-275, 2014.
	
Application Domain	
Phantom Type :	Software
Task :	Compare lesion segmentation schemes
Body Region :	Head
Imaging Protocol:	1.5T MR, 3mm slice thickness
Main Steps	
<u>Object</u>	
Hypothesis :	Compact, high-resolution object of user-defined shape
Parameters :	Shape, Structure, Volume, Topology, Contrast, Noise, Resolution, PV Effects, Uniformity
<u>Background</u>	
Hypothesis :	Brain Web Data (registered data of volunteer, post processing, MR simulator)
Parameters :	Shape, Structure, Volume, Topology, Contrast, Noise, Resolution, PV Effects, Uniformity
<u>Incorporation</u>	
Hypothesis :	Voxel-wise combination of object and background gray values
Parameters :	-

Figure 5.9. Summary of the developed MS lesion phantoms.

5.2.3. Background Design and Object Incorporation

The BrainWeb data are used as background. Several different settings are available for download from the associated website. We use PD-, T2-, and T1-weighted images with a slice thickness of 1mm and an in-plane pixel size of $1\text{mm} \times 1\text{mm}$. Furthermore, a noise level of 3% is selected. Additional noise levels (0%, 1%, 5%, 7%, 9%) as well as slice thickness (1mm , 5mm , 7mm , 9mm) are available from the website.

The developed lesion probability map is used as additional tissue class. Due to registration inaccuracies, it is limited to the white matter mask of the brain. The final lesion phantoms are then incorporated into the MR data sets using an independently computed gray value for each lesion.

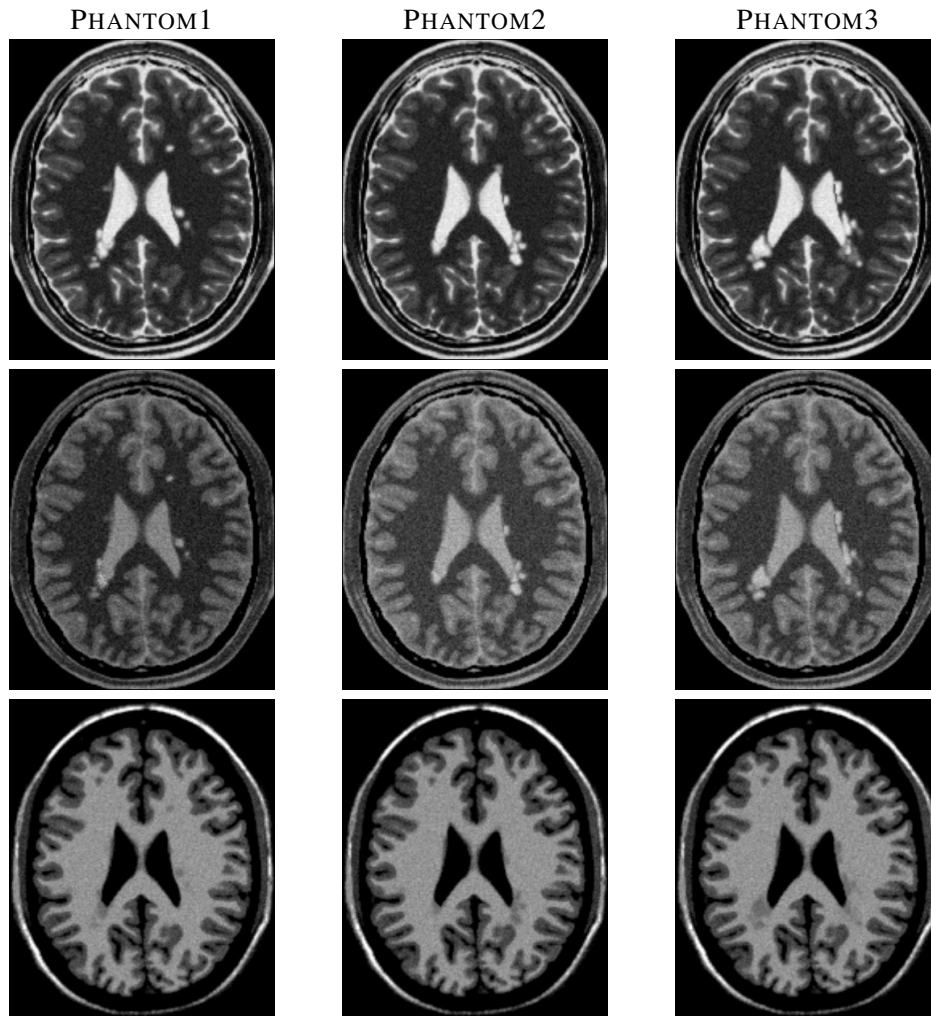


Figure 5.10. Examples of the resulting phantom data sets. Row 1-3: T2-, PD-, T1-weighted data of one phantom.

5.2.4. Resulting Software Phantoms

The tabular phantom description with the most important aspects is given in Figure 5.9. We developed 16 phantom data sets with varying amount of lesions. The associated total lesion load (TLL) has a range from 1.12ml to 7.18ml. An overview is given in Table 6.8. Three examples of the resulting phantoms are shown in Figure 5.10. The overall quality of the data sets is assessed by a domain expert using a rating scale based on fuzzy terms (poor, low, average, high, very high). The detailed results for each data set are given in Table 5.1. Adding a more complex texture model results in a higher rating for all phantoms.

Table 5.1. Visual assessment of the quality of each phantom data set performed by a domain expert. See also Sec. 6.3 for additional information about the phantom data sets.

HOMOGENEOUS LESIONS								
	Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8
Ranking	avg	avg	avg	avg	avg	avg	avg	avg

LESIONS WITH TEXTURE								
	Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8
Ranking	high	high	high	high	high	high	high	high

5.3. Brain Tumor Phantoms

Tumors of the central nervous system (CNS) are considered some of the most lethal and difficult to treat forms of cancer. The main types of treatment are surgery, radiation therapy, and chemotherapy (Osborn and Tong 1999). Characteristic properties that have to be taken into account include the anatomy and vascularity of a tumor, its relation to adjacent structures, as well as neurological functions of underlying areas. Furthermore, a fundamental issue is the accuracy of the calculated qualitative and quantitative parameters, which can have direct impact on treatment planning and therapy monitoring.

Today, MRI is often the imaging method of choice in clinical routine due to its high soft-tissue contrast combined with versatile parameterization alternatives (Thornton et al. 1992). Common acquisition protocols include T2- and FLAIR-weighted sequences, which are sensitive to edema and to tissue infiltration, providing an indication of the extent of low malignant tumor parts. Furthermore, T1-weighted sequences pre- and post-contrast facilitate an estimation of blood-brain barrier dysfunctions, often caused by high malignant tumor parts.

The boundary of a tumor and its volume are often used as objective parameters. However, since brain tumors can largely vary in size, shape, amount of edema, and enhancement characteristics, any analysis used in clinical routine and in multi-center studies has to be carefully evaluated. Varying acquisition protocols and image quality add to complexity of this task. Unfortunately, publicly available reference data sets are hardly available. Furthermore, a ground truth is usually not known for brain tumors.

In this section, we develop brain tumor phantoms, which have been introduced in (Rexilius et al. 2004). A biomechanical model enables us to simulate deformations that occur due to tumor growth inside the brain (cf. Sec. 5.3.3). Additional properties such as the amount of edema at each voxel (cf. Sec. 5.3.2) as well as a simulation of contrast agent enhancement characteristics are provided within the software phantom (cf. Sec. 5.3.5). An extension of our work was proposed by Prastawa et al. (2009). Their phantoms have also been used as training data at the BRATS challenge at MICCAI 2012.

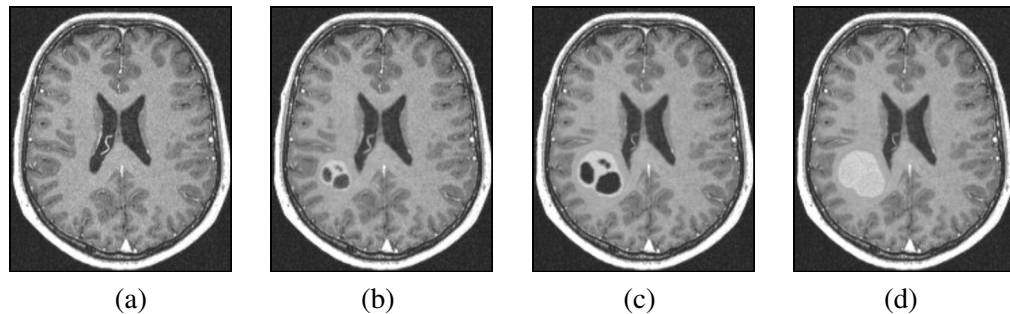


Figure 5.11. Examples of tumor phantoms without simulation of edema that differ only in size and amount of necrosis in comparison to original MR data. (a) Original T1-weighted image post contrast (T1gd) of healthy volunteer, (b) T1gd image with small tumor, (c) T1gd image with large tumor, (d) T1gd image with large tumor and necrotic tissue scaled to 5%.

5.3.1. Overall Phantom Design

To develop a realistic object shape and appearance for brain tumors, we use two different tissue classes: active tumor and necrosis. Furthermore, we introduce an additional class for edema (cf. Sec. 5.3.2). This approach can be considered as an extension of the one-class approach proposed for MS lesions in the previous sections of this chapter.

Both, active tumor tissue and necrosis are obtained from segmentation masks of a real tumor patient, and are resampled onto a high-resolution grid (512^3 grid). The amount of necrosis can be easily changed by appropriate scaling (cf. Fig. 5.11 and 5.12). Suitable object gray values as well as noise characteristics for both tissue classes are determined from the patient data set, and are adjusted to the underlying background data set.

Three-dimensional T1-weighted pre and post contrast MR images from a healthy volunteer are used as background data. All data are obtained from a clinical 1.5T scanner (Siemens Magnetom Vision, Siemens, Erlangen, Germany) with 1.0mm isotropic voxel size, 256×256 matrix (cf. Fig. 5.11 (a)). A simulation of deformations induced by tumor growth is performed to provide a more realistic background model. Furthermore, we introduce an additional tissue class, to simulate edema. The following sections give a more detailed description of these steps. Finally, the objects for each tissue class are incorporated into the volunteer MR data sets as a linear combination of the deformed MR scan and the tumor tissue classes.

5.3.2. Modeling of Edema

In addition to the actual tumor tissue, edema is another important structure that should be taken into account. Brain edema is an inflammatory response to the tumor, which causes the brain around the tumor to swell, and is mostly located in the white matter (Osborn and Tong 1999). Because the brain is located in a confined space and cannot expand, and

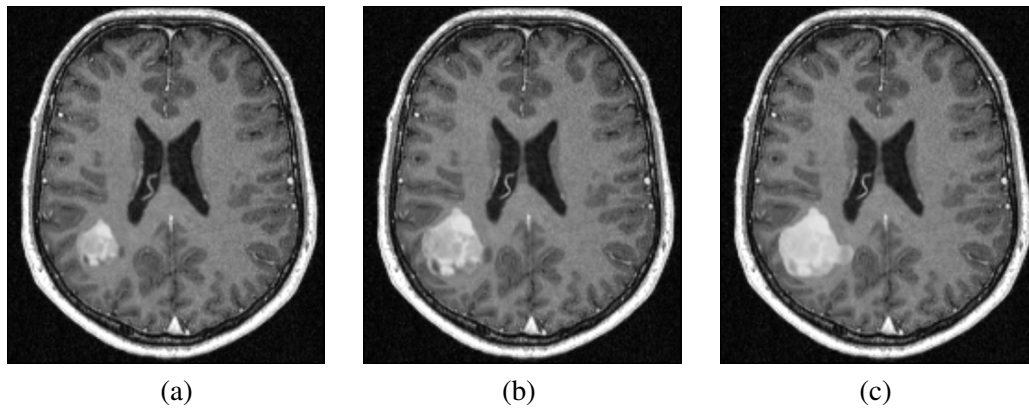


Figure 5.12. Second example of tumor phantoms without simulation of edema, similar to Figure 5.11. (a) T1gd image with small tumor, (b) image with large tumor, (c) necrotic tissue scaled to 5%.

because the fluid that accumulates cannot easily be carried away, an edema can impair the normal brain functions and cause an increase of intracranial pressure. Therefore, an accurate segmentation and quantitative analysis could add valuable information for a physician. For example, a method for tumor segmentation with an explicit model for edema is proposed in (Prastawa et al. 2003).

We simulate the amount of edema at each voxel, using a geodesic distance transformation (Soille 2003) starting from the tumor boundary. The basic idea is to constrain the distance computation to remain within a subset of the image volume. We use a white matter mask, because the edema is usually located in the white matter of the brain. Depending on the resulting distance map we define a region of pure edema and of a mixture between pure edema and normal brain tissue. This way, various degrees of edema dissemination can be simulated. The amount of PV is scaled accordingly. Further PV effects that occur between edema and tumor tissue at the boundary of the tumor are considered as well. The resulting edema probability mask is then added to the structure module as additional tissue class. Suitable intensity values and noise level are again extracted from the associated patient data. See Figure 5.13 for an example.

Unfortunately, the amount of edema is not only influenced by the distance to the tumor boundary. Other aspects such as tumor infiltration should also be taken into account. Furthermore, methods that account for preferred tumor dissemination pathways could provide a more accurate basis for the tumor and edema growth, e.g., maps of the principal diffusivity directions derived from diffusion tensor imaging.

5.3.3. Modeling of Deformations Induced by Tumor Growth

An important aspect in generating a realistic phantom for brain tumors is to simulate the deformation imposed by tumor growth. A fundamental assumption is that surrounding brain tissue is pushed away from the tumor. In order to gain insight into the process of tumor growth, mathematical modeling has become an increasingly important role and various methods have been proposed.

We simulate the three-dimensional tumor growth using a physics-based model. Our approach is based on a linear elastic model that was previously used to capture shape changes of the brain during neurosurgery (Ferrant et al. 2001; Rexilius et al. 2001). Since a rigid model can be assumed for surrounding tissue such as the dura mater, the model is constrained at the boundaries of a brain mask generated by a watershed transform (Hahn and Peitgen 2000), so that motion is restricted to areas inside the brain. Tumor growth is then simulated from an initial start point. Thereby, a brain tumor object is placed at an arbitrary position inside the brain with a given radial displacement $u(\mathbf{d}) = \alpha\mathbf{d}$, $\alpha \in \mathbb{R}^+$ in each direction $\mathbf{d} \in \mathbb{R}^3$. The center of gravity of the tumor object is used as point of origin. The constraints for tumor and brain boundary are then introduced as external forces into the elastic model defined in Equation 5.1 with material parameters $E = 3kPa$ (Young's modulus measured in Pascal) and $\nu = 0.4$ (Poisson's ratio) as

$$E(u) = \frac{1}{2} \int_{\Omega} \sigma^{\top} \varepsilon \, d\Omega - \int_{\Omega} F^{\top} u \, d\Omega. \quad (5.1)$$

Ω	image domain
σ	strain
ε	stress
F	external forces
u	displacement.

Similar parameters have been used for example by Ferrant et al. (2001). The computed constraints for both, tumor and brain boundary are then introduced as external forces into the elastic model. Thus, changes in the shape of the brain are modeled to result in an equilibrium state of energy with a displacement u that minimizes the total potential energy as defined in Equation 5.1.

The resulting equation is solved by a finite element approach (Zienkewicz and Taylor 1987). We chose a structured mesh with a triangulation done using tetrahedra and linear shape functions. To solve the resulting system of equations, we use a fast parallel implementation, which was already implemented as part of a nonrigid registration approach (Rexilius et al. 2001).

Finally, an iterative tumor growth is applied as proposed in Equation 5.2. We start with a deformation restricted to the boundary of a downsampled tumor shape (factor 5). The center of gravity is placed at the same position as for the full-scale tumor, so that even brain

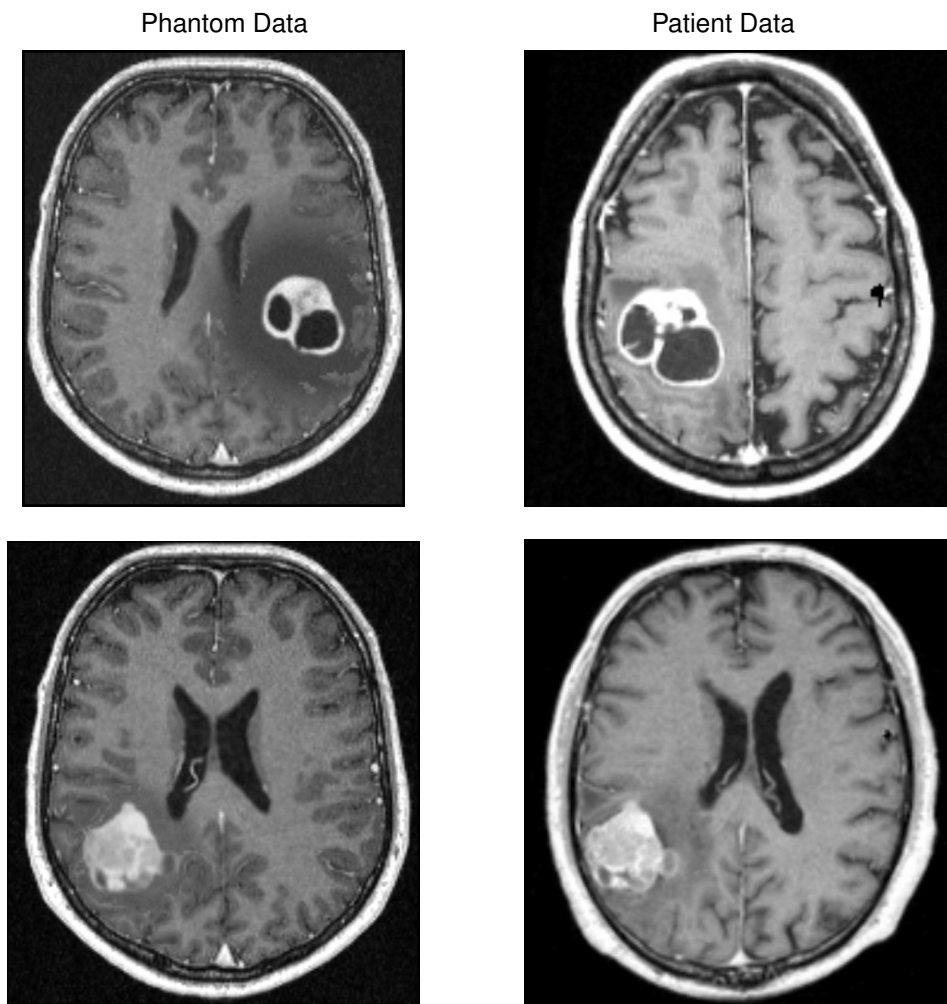


Figure 5.13. Examples of different software phantoms with simulated edema compared to actual patient data. (Left column) MR scan of a healthy volunteer with incorporated tumor object, (Right column) MR scans of patients with a brain tumor.

structures very close to or even inside the full-scale object's boundaries can be pushed away from the tumor. The amount of displacement per iteration varies with the scale factor α as defined above.

$$u_n(\mathbf{x}) = u_{n-1}(\mathbf{x}) + u(\mathbf{x} + u_{n-1}(\mathbf{x})) . \quad (5.2)$$

$u_{n-1}(\mathbf{x})$ displacement at time point n
 $u_{n-1}(\mathbf{x})$ displacement at time point $n - 1$.

The maximum deformation in the full-scale tumor size is set as stopping criterion for the iteration process. The amount of displacement per iteration varies with the scale factor α as

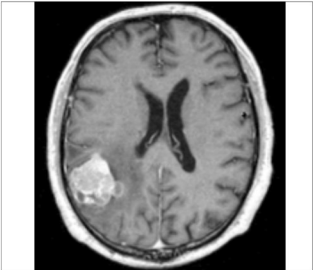
Phantom Description	
Author :	J. Rexilius et al.
Reference:	J. Rexilius et al. "A Framework for the Generation of Realistic Brain Tumor Phantoms and Applications". Proc. MICCAI, 3217 in LNCS, pp. 243-250, 2004.
	
<div style="border: 1px solid gray; padding: 2px; display: inline-block;">Application Domain</div>	
Phantom Type :	Software
Task :	Compare brain tumor quantification schemes
Body Region :	Head
Imaging Protocol:	1.5T MR, 1mm slice thickness
<div style="border: 1px solid gray; padding: 2px; display: inline-block;">Main Steps</div>	
<u>Object</u>	
Hypothesis :	Compact, high-resolution object of user-defined shape
Parameters :	Shape, Structure, Volume, Topology, Contrast, Noise, Resolution, PV Effects, Growth, Enhancement
<u>Background</u>	
Hypothesis :	MR scan of healthy volunteer
Parameters :	Growth
<u>Incorporation</u>	
Hypothesis :	Voxel-wise combination of object and background gray values
Parameters :	-

Figure 5.14. Summary of the developed brain tumor phantoms.

defined above. The associated parameter module $M_{p_{growth}}$ is defined as follows:

Parameter	Description ($M_{p_{growth}}$)
H	Tumor growth modeled by iterative linear elastic model
\mathbf{x}	Background data set and initial tumor shape
f	Compute background deformations using Eq. 5.1 and Eq. 5.2
y	Pixel-wise displacement map

5.3.4. Resulting Software Phantoms

Figure 5.13 illustrates the resulting software phantoms. Similar to the generated software phantoms for MS lesions, we present a comparison of a patient MR with a brain tumor

and a phantom data set. To evaluate the quality of these phantom data sets, we conducted a visual assessment by two clinical experts. All four image data sets were shown to the participants, i.e., the two patient and the two phantom data sets. The first example (top row in Fig. 5.13) was clearly identified as tumor phantom. Especially the edema region appeared to be too homogeneous. Nevertheless, one expert also rated the patient data as possible software phantom. In the second example, the bottom row in Figure 5.13, both experts rated the tumor phantom as potential patient data set. A summary of the phantom and its major components is given in Figure 5.14.

5.3.5. Simulation of Contrast Enhancement Characteristics

In this section we add an additional parameter model to our brain tumor phantom that allows the simulation of contrast agent enhancement characteristics. Since the focus in this work is not on dynamic MRI, we merely provide exemplary results to showcase the flexibility and potential of our approach.

Multi-compartment models are commonly used to describe the enhancement of macro-molecular contrast agent particles in tumor tissue and thus are an important tool for computer assisted analysis of dynamic MRI. We have used the Tofts&Kermode model to generate simulated perfusion data sets (Tofts and Kermode 1991). This enables us to combine the prediction of contrast agent enhancement and a known ground truth for a quantitative analysis in simulated brain tumors. To apply the Tofts&Kermode model to the tumor phantom, we generate maps of the artificial distribution of physiologic parameters: The permeability of tissue and the extracellular volume fraction that is accessible for the contrast agent. Since most of the active tumor tissue is usually located at the border of the tumor, we assume an increase of the permeability from the center to the border. A very low permeability is assigned to necrotic tissue using the simulated amount at each voxel as a scaling factor. The extracellular volume fraction is assumed to vary only slightly between 0.7 and 0.8. We set a higher value for necrosis than for active tumor tissue. Finally, contrast enhancement is simulated at a 0.5 minutes scan-interval from 0 minutes up to 15 minutes after injection of contrast agent at a dose of 0.1mmol/kg. The associated parameter module $M_{penhancement}$ is

Parameter	Description ($M_{penhancement}$)
H	Enhancement characteristics can be simulated via Tofts&Kermode model
\mathbf{x}	Phantom data set with tumor object
f	Tofts&Kermode model
y	Pixel-wise enhancement characteristic over N time-steps.

Figure 5.15 (b) provides a color-coded visualization of the wash-in and wash-out of contrast agent within the tumor region. The red area corresponds to active tumor, the blue area to necrosis. The corresponding enhancement curves for different positions inside the tumor are shown in Figure 5.15 (c). For example, the simulation of active tumor tissue (red

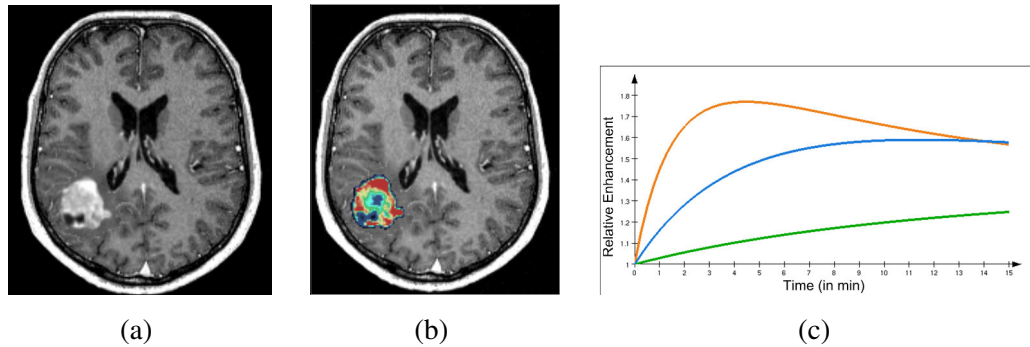


Figure 5.15. Simulated enhancement characteristics. The imaging parameters are adapted to the parameters of the real MR scan (T1w gradient echo, TE=5ms, TR=15ms, FlipAngle=30°). (a) Slice of phantom data set, (b) combined colored visualization of wash-in and wash-out of contrast agent (red: active tumor, blue: necrosis) (c) relative enhancement curves for selected regions inside the tumor.

curve) results in a rapidly increasing curve due to high values in the generated permeability parameter map.

5.4. Discussion

Today, different software phantom approaches are available, and some have already been discussed in the previous chapters. For example, evaluation of segmentation methods is often based on artificial phantoms that consist of geometric primitives. Although such an approach provides a reasonable tool in many cases, a dedicated validation of algorithms using anatomical or other prior knowledge is not feasible, because these artificial phantoms can only provide a very rough approximation of anatomy. Voxel phantoms such as the BrainWeb phantom (Collins et al. 1998) on the other hand, enable a realistic representation of anatomical structures, and have gained importance in medical image analysis. However, only a small number of data sets are available for public use. The current BrainWeb data consist of only one data set with T1-, T2-, and PD-weighted sequences. Twenty additional data sets (only T1-weighted) are available based on the work of Aubert-Broche et al. (2006).

In this work, we develop hybrid phantoms that have proven to be an elegant way to a flexible software phantom design (cf. Sec. 2.2.1). Our phantoms are well suited for tasks such as algorithm development and evaluation, and some examples are given in the upcoming chapter. We propose a phantom design based on the embedding of objects into already available anatomical structures. The object design consists of several parameter modules related to morphology and imaging parameters. Furthermore, our approach allows for an extension to growth-induced deformations of objects. The background is based on actual patient data or well-known software phantom data.

Two phantoms are based on a manual object design. In Section 5.1, we developed MS lesion phantoms consisting of a single tissue class from hand-crafted objects of typical lesion shapes. An extension of this approach is presented in Section 5.3 for brain tumors. Besides a tumor class, we model edema as additional tissue class. Furthermore, object deformations affecting the background data are applied to model tumor growth. Nevertheless, the actual tumor objects are still modeled by hand based on reference patient data. For both applications, our software assistant introduced in Section 4.3 is used.

The phantoms described above provide a high degree of flexibility. However, generating a new phantom requires a long processing time and a large amount of manual work. Therefore, we extended the above methods to a fully automatic approach for MS lesion phantoms in Section 5.2. A statistical map of object positions and an automatic selection of other parameters such as shape or volume enable the generation of an arbitrary amount of phantom data sets. To the best of our knowledge, our approach represents the first method of this type.

6. Applications

After developing several brain lesion phantoms in the previous chapter, we now discuss how such phantoms can be used to evaluate algorithms that frequently occur in clinical practice and trials. We focus on MS lesions and brain tumor phantoms. Four different applications are analyzed: First, we investigate the quality of visual assessment in MS lesion volumetry (cf. Sec. 6.1). Based on a specially developed software assistant, a survey with more than 20 participants is carried out. Typical tasks are evaluated such as estimating volume change between two lesion phantoms by pure visual inspection.

In Section 6.2, we then evaluate common methods for lesion volumetry including manual tracing by three field experts plus voxel counting as well as two semi-automatic methods. More than 50 phantom data sets are generated for this task. Furthermore, we perform an intra-observer study, where all phantom data sets are repeatedly analyzing. Our results demonstrate frequent problems with inter- and intra-observer reproducibility, and the importance of an improved gold standard in lesion volumetry beyond voxel counting.

Finally, the use of phantoms for the evaluation of segmentation methods is discussed in Section 6.3 and Section 6.4. Several algorithms ranging from manual to fully-automatic are analyzed. The Dice similarity coefficient is used to compare the results with the ground truth known from phantom data.

6.1. Visual Assessment in MS Lesion Volumetry

The change of lesion load on yearly T2-weighted MR scans of the brain is widely used in clinical routine and clinical trials. Thereby, the expected change is estimated to be around 10% (Filippi et al. 1998). Today, common measurements are based on manual segmentation followed by voxel counting within the segmentation mask, or even by a simple visual inspection of the available image data. But how reliable are such measurements, and what are their chances and pitfalls?

In this section, we present a survey of a common quantitative measurement used in clinical practice, namely pure visual inspection. Thereby, an assessment heavily relies on the human eye to detect changes, which are then characterized in a qualitative way. To allow for a more quantitative analysis, visual rating scales have been proposed in several medical

disciplines. Nomori et al. (2005) compared visual assessment and a semi-quantitative analysis of fluorodeoxyglucose (FDG) uptake on PET for the evaluation of lung nodules. Three grades were used to describe each nodule. A retrospective analysis of a visual grading from unenhanced CT data for the diagnosis of 30% or higher of hepatic steatosis in living donor liver transplantation candidates was proposed by (Lee et al. 2007). An overview of different visual rating scales used for the analysis of anatomical and pathological white matter changes in the brain is given in (Malloy et al. 2007).

In this section, MS software phantoms of different shape and volume are used to provide a large range of different realistic data sets (cf. Chap. 4). Furthermore, the exact ground truth is known for each modeled lesion. To this end, we developed a software assistant that consists of several steps, comprising different tasks for the user. Each data set contains exactly one lesion object that can be easily detected by the user.

6.1.1. The Software

The user interface consists of two parts: A viewing section and a description and question section. To simplify the viewer handling, we provide several predefined zoom factors. Furthermore, the number of visible slices per viewer can be interactively adjusted. Since we are not interested in an evaluation of object detection, a region of interest is drawn around a lesion. The lower part of the user interface in each step consists of a description of the requested tasks as well as a short documentation of the current step and viewer functionality. Additionally, a box with questions is given.

To participate in the survey, a user has to download the application from a website and install it on her/his computer. A preamble serves to acquire some personal information including age and sex as well as work experience (in years) and area of expertise (radiology, neuroradiology, computer science, etc.). The results are stored in a separate structure that is automatically returned as plain text by email after answering all questions.

Three different aspects are investigated, each consisting of at least two steps with different lesion objects:

(1) Visual inspection.

The first part of the survey consists of five steps (S1–S5). In each step, a set of three lesions are presented, and the user is requested to estimate their relative volumes. In two steps the lesion objects have different shapes, whereas the three other steps comprise a single shape. We asked for the largest/smallest lesion or if all lesions have the same volume respectively. The questions are given as follows:

- Q1 : All lesions have equal volume?
Possible answers: yes, no
- Q2: Which is the largest lesion volume?
Possible answers: left, middle, right

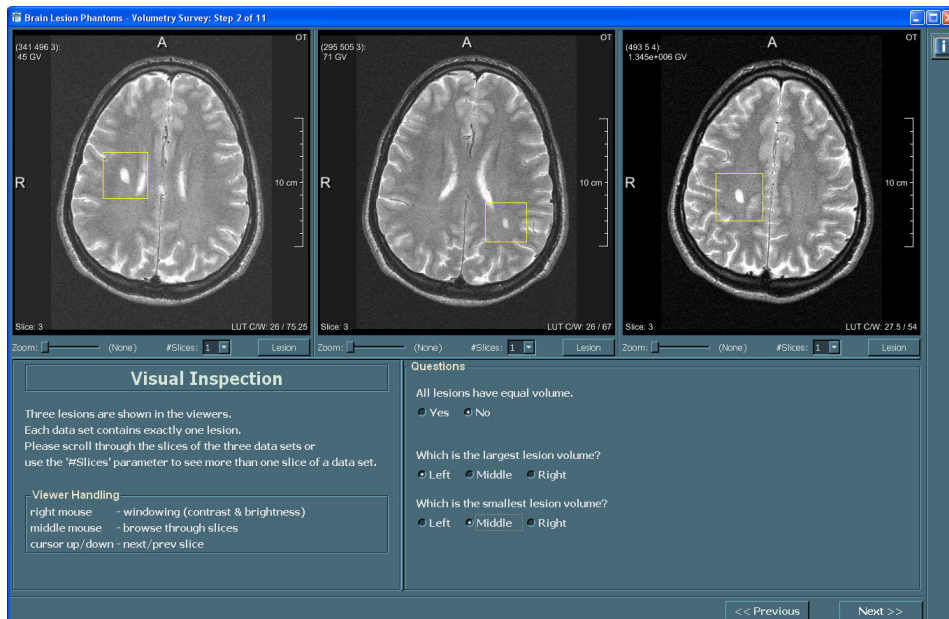


Figure 6.1. First part of the software assistant: Visual inspection.

- Q3: Which is the smallest lesion volume?
Possible answers: left, middle, right

Figure 6.1 shows an example of the corresponding user interface.

(2) Segmented mask volume.

In the second part of the survey, different segmentation masks are presented to the user as overlay over a lesion object (two steps, S6–S7). Besides the above mentioned interaction and viewing properties (zooming, etc.), the transparency of a segmentation mask as well as its color can be adjusted to provide further flexibility. We asked for the largest/smallest segmented mask volume as well as for the best fit of the true lesion volume. The questions are given as follows:

- Q4: Which is the largest segmented mask volume?
Possible answers: left, middle, right
- Q5: Which is the smallest segmented mask volume?
Possible answers: left, middle, right
- Q6: Which segmented mask volume is closest to the true lesion volume?
Possible answers: left, middle, right

Figure 6.2 gives an example of the user interface.

(3) Lesion growth.

Finally, an examination of lesion growth and shrinkage has to be performed. Here,

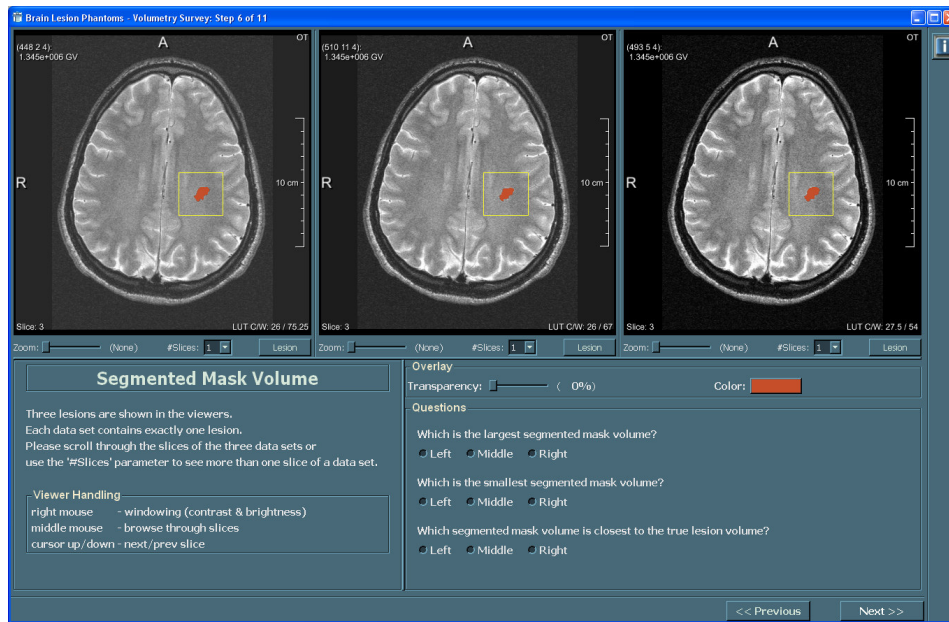


Figure 6.2. Second part of the software assistant: Segmented mask volume.

two lesion objects of different volume are presented to the user in each step (three steps, S8–S10). The image in the left viewer is regarded as original data set, the other image as follow-up data set. The questions are given as follows:

- Q7: Is a volume change visible between original data and follow-up data?
Possible answers: no volume change, volume shrinkage, volume growth
- Q8: Volume shrinkage in percent:
Possible answers: -5%, -10%, -25%, -30%, -40%, -50%, -75%, -100%
- Q9: Volume growth in percent:
Possible answers: 5%, 10%, 25%, 30%, 40%, 50%, 75%, 100%

An example of the user interface is given in Figure 6.3.

6.1.2. Results

Five women and 16 men participated in the survey. The average years of experience of all participants was 4.93 ($min=0.5$, $max=15$). All of them already worked with medical image analysis software before. The largest group of participants were experts in the field of software development and research in medical imaging (19; 90.5%). Among them were 16 computer scientists, two mathematicians, and one physicist. Furthermore, 2 physicians (radiology, neuroradiology) participated in the survey.

Twenty software phantom data sets including different lesion shapes and volumes have

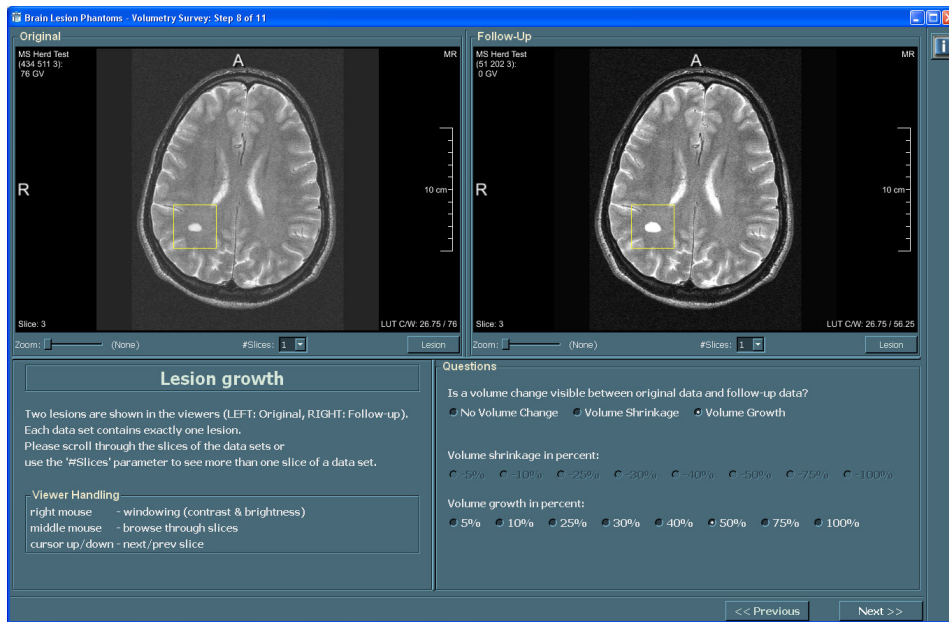


Figure 6.3. Third part of the software assistant: Lesion growth.

been developed for this study. The object volumes range between 0.05ml and 0.6ml. No participant correctly answered all questions in the ten steps of the software assistant. Especially the lesion volume change between two data sets was often miscalculated. In the following, we present the results of the three different tasks of this study.

(1) Visual inspection.

The first part of this study deals with the assessment of relative lesion volumes. Either a single volume for all data sets or three different volumes are used, ranging from 0.05ml to 0.44ml. Thereby, for a step that contains different volumes, the volumetric difference between lesion objects within one step is at least 10%, which is the expected change per year as described above. The underlying volumes as well as the correct answers for each step are given in Table 6.1.

S1: Three different lesion shapes with equal volume are shown in this step. The correct answers, i.e., equal volume for all three lesions, were given by nine participants (42.9%).

S2: A single lesion shape is presented in all viewers. The largest and the smallest volume were correctly estimated by 18 participants (80.7%). No participant voted for an equal volume for all lesions.

S3: Again three different lesion shapes with equal volume are shown. Here, eight participants (38.1%) gave the correct answer.

Table 6.1. Ground truth for each step of task 'visual inspection'.

Step	Volume (in ml)			Questions		
	left	middle	right	Q1	Q2	Q3
S1	0.1	0.1	0.1	yes	–	–
S2	0.36	0.12	0.25	no	left	middle
S3	0.36	0.36	0.36	yes	–	–
S4	0.4	0.44	0.36	no	middle	right
S5	0.05	0.05	0.05	yes	–	–

S4: The three viewers show a single lesion shape. The largest as well as the smallest volume were correctly estimated by five participants (23.8%). Five participants voted for an equal volume for all lesions.

S5: In this task, a single lesion shape with equal volume is presented. Most participants (19, 90.5%) voted for the correct answer.

(2) Segmented mask volume.

In the second part of this survey, segmentation masks are overlaid on the lesion objects. A volume of 0.36ml is chosen in step 6, and a volume of 0.12ml in step 7. See Table 6.2 for the underlying ground truth of the segmented mask volumes and the correct answers.

Table 6.2. Ground truth for each step of task 'segmented mask volume'.

Step	Lesion Volume (in ml)	Mask Volume (in ml)			Questions		
		left	middle	right	Q4	Q5	Q6
S6	0.36	0.59	0.53	0.56	left	middle	middle
S7	0.12	0.26	0.20	0.23	left	middle	middle

S6: The largest mask was correctly identified by 13 participants (61.9%), the smallest mask by only two participants (9.5%). Furthermore, nine participants (42.9%) selected the correct mask volume closest to the true lesion volume.

S7: For this task, ten participants (47.6%) correctly identified the largest mask. The smallest mask was correctly identified by no participant (0%). Nevertheless, 18 participants (85.7%) selected the correct mask volume closest to the true lesion volume.

(3) Lesion growth.

The change of lesion volume between two data sets has to be estimated in the last part

of the survey. Table 6.3 provides the true volumes as well as the requested growth rates. Our results show that this last task has been the most difficult one.

Table 6.3. Ground truth for each step of task 'lesion growth'.

Step	Volume (in ml)		Questions		
	left	right	Q7	Q8	Q9
S8	0.4	0.6	Growth	–	50%
S9	0.4	0.28	Shrinkage	-30%	–
S10	0.1	0.125	Growth	–	25%

S8: Only two participants (9.5%) correctly estimated the volume growth. Two participants (9.5%) estimated no volume growth at all. No one decided for a volume shrinkage. Furthermore, no participant overestimated the volume growth (0%). The mean estimated volume growth was 21.2% ($min = 0\%$, $max = 50\%$).

S9: The correct amount of volume shrinkage was estimated by one participant (9.5%). No one underestimated the amount of shrinkage. Furthermore, no one estimated a volume growth. Nevertheless, four participants (19.0%) estimated no volume growth at all. The mean estimated volume shrinkage was -13.1% ($min = 0\%$, $max = -30\%$).

S10: In this step, most participants (12; 57.1%) estimated no volume growth. One participant selected the correct volume growth (4.8%). The mean estimated volume growth was 3.8% ($min = 0\%$, $max = 25\%$).

6.1.3. Discussion

The survey has demonstrated the capabilities of our software phantoms. Typical drawbacks of visual rating schemes have been identified. For example, only few participants were able to correctly estimate the largest or smallest lesion, if three different lesion volumes were shown. This also implies an overall low performance for a differentiation between different segmentation masks. Furthermore, a comparison of the change in lesion volumes between two data sets clearly revealed the importance of computer assistance for the quantitative analysis of lesions. Although most participants were at least able to give the correct trend in the data, i.e., growth or shrinkage, the volume change was typically underestimated.

6.2. Accuracy in MS Lesion Volumetry

Besides visual inspection of clinical data, a manual or (semi-)automatic assessment of the volumetric lesion load often is used as an objective parameter. A fundamental issue is the

accuracy of the calculated lesion volume because this can increase the impact of lesion volumetry on diagnosis and therapy monitoring of the disease. Several methods have been proposed to quantify lesion burden, ranging from manual tracing of each lesion by experts to semiautomated and fully automated methods. A common method used in many clinical trials that evaluate cancer treatments is the Response Evaluation Criteria In Solid Tumors (RECIST) (Therasse et al. 2000), which uses a one-dimensional measurement of the tumor size, the diameter, to approximate the volume. Molyneux et al. (1998) evaluate the performance of manual outlining and a semi-automatic contour technique for the segmentation of MS lesions from 16 patient data sets. The authors show that the semi-automatic method results in a more robust quantification of the lesion load. A related approach is proposed in (Ashton et al. 2003), where manual tracing is compared with two semi-automatic methods. Nevertheless, both studies apply a simple voxel counting without considering PV effects during volume calculation, i.e., the volume is computed as the number of voxels within a segmentation mask multiplied by the volume of a single voxel (Clark et al. 1998; Joe et al. 1999; Kaus et al. 2001). An illustration of this method is shown in Figure 8.2.

The aim of this section is to introduce a new approach for the validation of MS lesion volumetry. Because of the absence of a ground truth in patient data, we generate software phantom data sets of MS lesions with known exact volumes (Rexilius et al. 2005).

Extensive experimental studies over a broad range of different lesions are carried out manually by domain experts. Furthermore, a three-dimensional region growing combined with voxel counting within the segmentation mask as well as a robust semi-automatic volumetry approach based on Bayesian classification with explicit PV modeling are used for comparison.

6.2.1. Software Phantoms for the Evaluation of MS Lesion Quantification

The software phantoms used in this study are developed based on the approach proposed in Section 5.1. The images used as background in this study were acquired from a healthy volunteer (a 28-year-old male) on a 1.5 T scanner (Magnetom Vision; Siemens, Erlangen, Germany). The data-acquisition protocol contained axial and coronal PD-, T2-, and T1-weighted images with an in-plane resolution of $0.449 \times 0.449 \text{mm}^2$ and a slice thickness of 3 mm, matrix of 512×512 , and 34 axial and 51 coronal continuous slices, respectively. All images were acquired in one session with head fixation and without table movement, such that all data sets are perfectly aligned without visible motion artifacts.

The lesion objects are generated manually and placed at typical paraventricular positions in the white matter of the brain. To cover a variety of realistic-shaped MS lesions, we generate three different lesion objects: a sphere-like lesion (L1), an ellipsoidal lesion (L2), and an elongated lesion containing several deformations (L3). We did not use regular-shaped cylinders or spheres because this would not lead to a realistic lesion appearance. A 3D surface rendering of these shapes is shown in Figure 6.4.

In addition to various shapes, six different lesion volumes are chosen: 0.05, 0.1, 0.2,

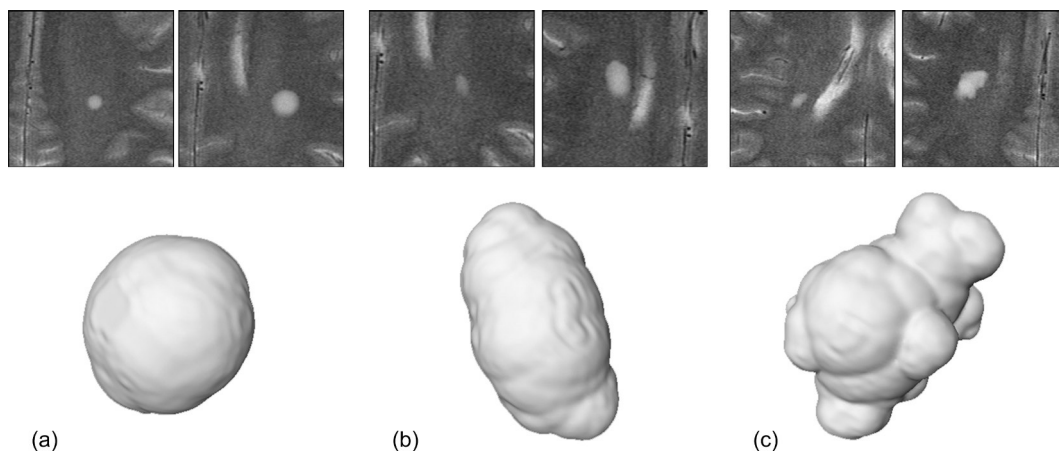


Figure 6.4. Examples of software phantoms for MS lesions with different shapes and sizes. The top row shows a slice of the resulting phantom in 2D, the bottom row shows the employed lesion objects in 3D. (a) Sphere-like, (b) ellipsoidal, and (c) irregular shape containing several deformations.

0.4, 0.7, and 1.0ml, resulting in a total amount of 18 different lesion objects. The voxel size for the largest volume (1.0 ml) was set to an isotropic voxel size of 0.05 mm, resulting in 8,000,000 voxels. Voxel sizes for the remaining volumes are calculated accordingly.

6.2.2. Setup for Manual Expert Analysis

Each data set is analyzed by three experts. In order to provide an intuitive but still powerful tool for the manual analysis of the provided phantom data sets, we developed an application with a graphical user interface based on the research and development platform MeVisLab (MeVisLab 1.5). Therein, an expert is able to trace the boundaries of a lesion, shown as overlay on the original slices. The actual volume is then computed using the voxel counting approach. In addition to basic drawing functionalities, the user may adjust the lookup-table and simultaneously view several neighboring slices. Furthermore, it is possible to change between available sequences during outlining a lesion on one slice. Because we want to analyze volumetric results and not the lesion detection task of different experts, only one lesion is incorporated per data set.

6.2.3. Semi-Automatic Volumetry

Two semi-automatic measurement techniques are evaluated in this work. To provide results corresponding to a popular method for the segmentation of MS lesions in clinical routine and studies, a seeded region growing algorithm is used (Sonka et al. 1998). This technique requires a user to place a "seed" within the lesion using a single mouse click. The region then

successively grows and neighboring voxels are added to the region up to a certain intensity threshold. A compact shape that does not deviate excessively between two neighboring intensity thresholds is used to constrain the process. The lesion volume is computed by voxel counting on the resulting mask similar to the manual approach.

The second algorithm combines a 3D marker-based segmentation and a bimodal histogram analysis with an explicit model for PV effects. Only the T2-weighted images are considered due to their high lesion contrast. In a first step, a cuboid subvolume that contains the entire lesion is selected and resampled in z-direction to an isotropic voxel size. Then, an interactive watershed transformation is applied to generate an over-inclusive segmentation (Hahn and Peitgen 2000). Two different marker types are used. One include marker is placed inside the lesion and up to five exclude markers are used to separate the lesion from other hyperintense structures. The resulting region contains the complete object boundaries including all voxels where PV effects occur.

In a subsequent step, the lesion is classified with a statistical parametric classification algorithm, which are widely used in image analysis. The overall probability density function (PDF) of a voxel is given by

$$P(i) = \sum_{m \in T} P(w_m) P(i|w_m), \quad T = \{background, lesion, PV\}, \quad (6.1)$$

$P(i)$	M -component mixture density
w_m	underlying tissue class in the image
$P(w_m)$	mixing proportions
$P(i w_m)$	component density of given tissue class w_m
i	observed value of a random vector (here: intensity value).

The weights or prior probabilities, $P(w_m)$, which act as scaling parameters of each class PDF, are nonnegative quantities that sum up to one, i.e.,

$$0 \leq P(w_m) \leq 1 \quad (m = 1, \dots, M)$$

and

$$\sum_{m=1}^M P(w_m) = 1.$$

A Gaussian PDF (cf. Eq. 3.2) is assumed for the two pure tissue classes lesion and background.

Because of a typically small object size for MS lesions in the order of the slice thickness and the complexity of tissue boundaries, partial volume (PV) voxels, i.e., a mixture of pure tissue classes, is likely to occur. A common assumption for the modeling of partial volume voxels is a uniform distributed linear mixture of pure tissues because mixture voxels may consist of any fraction of pure tissue. This mixture model can be incorporated directly into the clustering algorithm, and the extended classification process therefore can

be implemented computationally efficient. Model parameters for the partial volume tissue class are defined as a convex combination of the model parameters of the two pure tissue classes, as proposed in (Noe and Gee 2001). The intensity of a voxel is then determined by a weighted sum of lesion tissue with intensity i_1 and background tissue with intensity i_2 as

$$v = a_1 i_1 + (1 - a_1) i_2 .$$

The corresponding PDF is again a Gaussian density function with parameters

$$\begin{aligned} \mu_{mix} &= a\mu_1 + (1 - a)\mu_2 , \\ \sigma_{mix} &= a^2\sigma_1 + (1 - a)^2\sigma_2 . \end{aligned} \quad (6.2)$$

The EM algorithm is applied to estimate appropriate parameters for each class (van Leemput et al. 1999).

Finally, the object volume is computed by

$$V = \left[P(w_{lesion}) + \frac{P(w_{PV})}{2} \right] \cdot V_{voxel} \cdot N \quad (6.3)$$

$P(w_{lesion})$	computed weight for lesion object
$P(w_{PV})$	computed weight for partial volume region
V_{voxel}	voxel volume
N	number of voxels in ROI.

Equation 6.3 assumes a symmetric PV distribution, so that only 50% of the computed weight $P(w_{PV})$ are added to the total volume.

6.2.4. Results

We evaluated 54 phantom MR data sets generated from a brain scan of a normal volunteer with exactly one MS lesion phantom per data set as described above. Lesion objects were placed at typical paraventricular positions in the white matter of the brain. Eighteen different lesions were generated for this study, consisting of six different volumes (0.05, 0.1, 0.2, 0.4, 0.7, and 1.0 ml) for each of three generated lesion shapes (L1, L2, and L3). The maximum diameters of the three lesion shapes are L1 = 75.4mm (L2 = 7.6mm, L3 = 7.2mm) for the smallest lesions (0.05ml), and L1 = 713.5mm (L2 = 18.4mm, L3 = 19.3mm) for the largest lesions (1.0ml). Each lesion phantom is incorporated into both, axial and coronal orientations at corresponding positions. To simulate inter-examination variability with respect to partial volume artifacts and apparent lesion size, an additional set of images is created by randomly shifting each lesion in z-direction on the axial images.

The MS lesions were manually traced by three experienced raters using the software assistant described above. Furthermore, the two semi-automatic volumetry approaches introduced in the previous section were used for comparison. The required interactions for

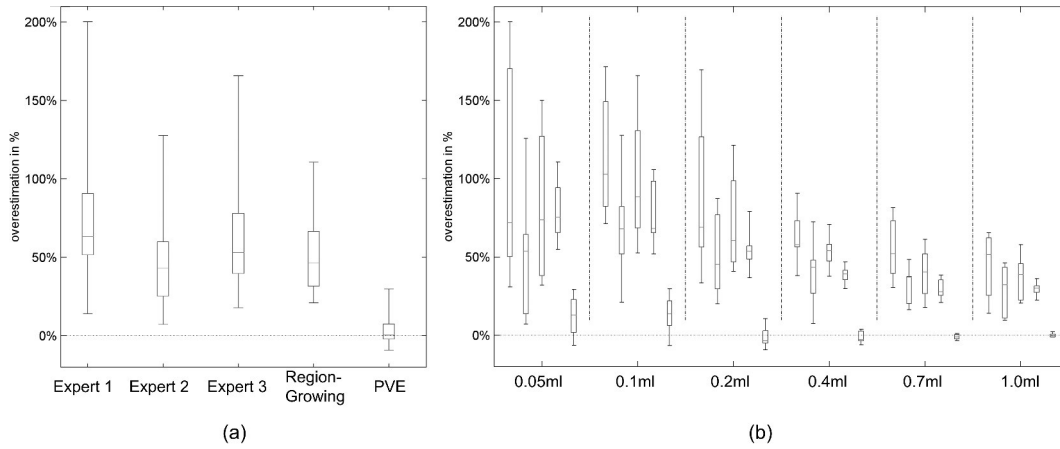


Figure 6.5. Results of manual and semi-automatic volume measurements in a boxplot. (a) Overall results calculated for each rater and algorithm separately. PVE refers to the semi-automatic partial volume analysis. (b) Results for each available volume, i.e., experts 1 to 3 and both semi-automatic approaches.

both algorithms were performed by Expert3.

Figure 6.5 (a) illustrates the overall error for each rater and the two semi-automatic methods in percentage of the true volume in a boxplot. A detailed analysis for each volume is given in Figure 6.5 (b). The overall median overestimation as well as the median overestimation computed for small lesions ($<0.3\text{ml}$) and intermediate lesions ($>0.3\text{ml}$) for each method is shown in Table 6.4. It can be clearly observed that all experts overestimated the true lesion volume, all to a comparable amount. The overall median overestimation for manual evaluation of the three experts ranges between 42.9% and 63.2%. The variability decreases with increasing volume size, because small changes already cause a significant relative error for small volumes. For small lesions the median overestimation over all experts is 73.2%, and 45% for intermediate lesions. No significant shape effect has been observed.

Similar results are obtained using the semi-automatic region growing approach combined with voxel counting. Here, the overall median overestimation is 46.3%. The semi-automatic approach with dedicated PV modeling, on the other hand, provides far more accurate results with a low error margin especially for intermediate lesions. Here, the overall median overestimation is 0.4%. Especially the deviation from the median is much smaller than for both voxel counting based methods.

To evaluate the intra-observer variability, repeatability analyses of the axial phantom data sets for each method were carried out by one rater, assessing each data set ten times. Figure 6.6 and Table 6.5 show the mean and the variance of the measured volumes. The manual approach was found to be highly variable especially for small lesions, whereas both semi-automatic methods provide a high reproducibility.

Table 6.4. Median overestimation of computed lesion volume for each expert and the two semi-automatic volumetry methods; #data sets: $n = 27$ (small and intermediate lesions), $n = 54$ (total).

Method	Median Overestimation (%)		
	Small Lesion	Intermediate Lesion	Total
Manual, expert 1	87.7	57.2	63.2
Manual, expert 2	58.3	35.5	42.9
Manual, expert 3	78.0	47.0	52.9
Region growing	66.5	31.6	46.3
Partial volume analysis	7.3	-0.9	0.4

Table 6.5. Systematic error and intra-observer variability trials for all proposed methods, assessing each data set ($n = 18$) ten times.

Method	Mean Error \pm SD (%)	
	Small Lesion	Intermediate Lesion
Manual, expert 3	91.4 \pm 32.3	56.5 \pm 17.0
Region growing	81.1 \pm 2.8	38.4 \pm 1.6
Partial volume analysis	14.7 \pm 1.0	1.3 \pm 0.4

6.2.5. Discussion

Manually outlined contours can provide a good estimation of visible lesion boundaries, although results may vary depending on the experts' common training. However, volumetric results using the common gold standard based on voxel counting within the segmentation mask show massive overestimation with poor interobserver and intraobserver reproducibility, even for lesions of intermediate size. Here, the partial volume effect is the largest source of systematic error. The computed volumetry results clearly depend on lesion size because, for small lesions, relatively more voxels are affected by partial volume.

This also has direct impact on the accuracy of other voxel-counting based methods, and one cannot expect significant improvement for these methods in general. Moreover, the negative effect on total lesion load measurements becomes evident. In this work, a common region-growing algorithm was used as an example. Our results show median overestimation similar to the manual approach, even if the semiautomatic method has a slightly lower error. However, a higher degree of automation could improve the reproducibility of the method and reduce the amount of training needed for inexperienced observers.

To accurately compute lesion volumes, algorithms with a model for partial volume ef-

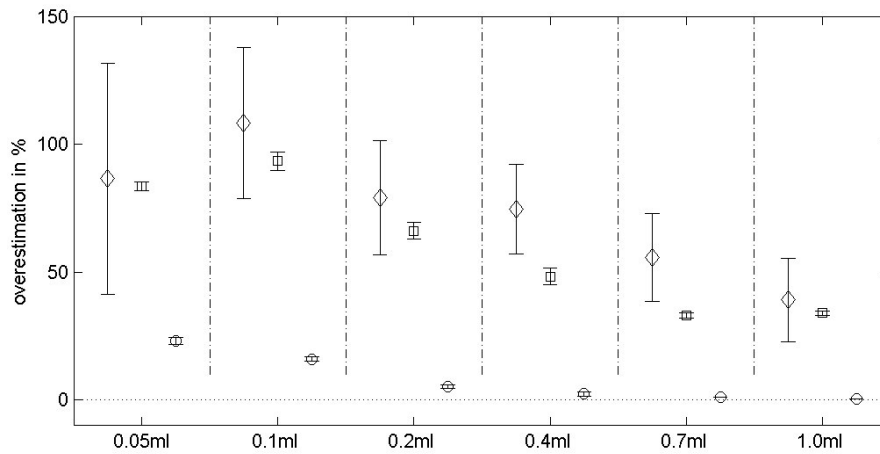


Figure 6.6. Results of an intra-observer study for all proposed methods, assessing each axial data set ten times. (\diamond , manual analysis; \square , region growing; \circ , PV analysis).

facts have to be taken into account. The proposed partial volume analysis method is robust for intermediate and large lesions.

Although correlation between MRI and clinical findings remains difficult, volumetric analysis of the lesion load has become an important issue and an active research field. However, analysis of the common gold standard in MS lesion volumetry among three experts shows a large median overestimation. The overall maximum was approximately 200% for one expert. An intraobserver study also showed large variability, even for a single rater. No manual tracing underestimated or met the true volume.

Similar results were obtained for a semiautomatic approach based on region growing. Although this method generates reproducible lesion segmentations, the actual volumetry is still based on voxel counting and thus is error prone. Because accuracy is an important factor for the clinical relevance of a method, results clearly indicate the importance of an improved gold standard in lesion volumetry beyond voxel counting. New measurements that accurately address partial volume artifacts are likely to correspond better to clinical findings. Therefore, our phantoms can provide a basis for comparison and testing of current and new approaches.

6.3. Segmentation of MS Lesions

Prior to a quantitative analysis of volumetric lesion load as discussed in the previous section, detection and segmentation are important pre-processing steps for diagnosis and treatment monitoring of MS lesions. An extensive description of lesion segmentation methods including both semi-automatic and automatic algorithms can be found in (Mortazavi et al. 2012). To evaluate a lesion segmentation method, using patient data and associated manual seg-

mentations, which serve as a surrogate ground truth, is a common approach. However, a large data pool is required to account for major anatomical variations and pathologies. Furthermore, exact measurements of parameters such as the volume of an imaged organ are not possible with patient data sets. Today, only few data sets that include both patient data plus manual expert segmentations are freely available for the community. An exception for example is the Lesion Segmentation Challenge workshop at the MICCAI conference (Styner et al. 2008).

Instead of another patient data set with unknown ground truth, we propose software phantoms for this task in this section. In Section 6.3.1, manual and semi-automatic methods are tested. A fully automatic approach is evaluated in Section 6.3.2.

6.3.1. Manual and Semi-Automated Segmentation

The Algorithms

Two algorithms are analyzed that were already used in Section 6.2. The first method is a manual expert segmentation. Here, a domain expert interactively outlines the lesion boundary. The second algorithm is a semi-automatic region growing, where a seed is manually placed within the lesion as starting position.

Results

To evaluate the two segmentation approaches, we use the same lesion phantoms that were proposed for lesion volume estimation (cf. Sec. 6.2). A total number of 18 phantom MR data sets is analyzed. For each data set (MS brain scan of normal volunteer), one MS lesion phantom is placed in the white matter of the brain. Three different lesion shapes were generated and six different lesion volumes are chosen for each shape (0.05, 0.1, 0.2, 0.4, 0.7, and 1.0ml) resulting in 18 different lesion objects.

To measure the overlap of the segmentation results with our ground truth, the Dice similarity coefficient $DSC(A, B) = 2|A \cap B| / (|A| + |B|)$, ($A = AlgorithmResult, B = GroundTruth$) is used (cf. also Sec. 8.2). Since both segmentation algorithm compute a binary lesion mask, we calculate the overlap only for ground truth voxels with $\geq 50\%$ lesion probability.

Table 6.6 gives the results for all phantoms. The expert segmentation usually resulted in lower overlap values compared to the semi-automatic region growing algorithm, independent of the segmented lesion. The values for the manual segmentation range between 0.58 and 0.90 and for the region growing approach between 0.61 and 0.96. Note, that a value of $DSC = 0.7$ is generally regarded as good segmentation. Similar to the volumetric lesion analysis, the computed similarity measure increases with increasing volume size. No significant shape effect is observed. For small lesions ($<0.3ml$), the mean similarity is $DSC_{manual} = 0.75$ and $DSC_{rg} = 0.78$ respectively, while for intermediate lesions the mean values increase to $DSC_{manual} = 0.84$ and $DSC_{rg} = 0.88$.

Table 6.6. Comparison of segmentation algorithms (manual segmentation, semi-automatic region growing) to ground truth using the Dice similarity coefficient (DSC). For the phantom ground truth data, only voxel with $\geq 50\%$ lesion probability are included.

Method	DSC (%)					
	0.05ml	0.1ml	0.2ml	0.4ml	0.7ml	1.0ml
Spherical Object						
Manual	0.85	0.84	0.78	0.83	0.86	0.90
Region growing	0.95	0.85	0.81	0.84	0.88	0.85
Ellipsoid Object						
Manual	0.87	0.75	0.58	0.85	0.85	0.84
Region growing	0.91	0.69	0.61	0.89	0.96	0.81
Deformed Object						
Manual	0.76	0.68	0.61	0.81	0.81	0.82
Region growing	0.80	0.70	0.69	0.83	0.95	0.87

6.3.2. Automated Multi-Spectral Segmentation

The Algorithm

After evaluating two segmentation algorithms that require user interaction, we now analyze a fully automatic approach. We use the lesion segmentation algorithm proposed by van Leemput et al. (2001), which has become quite popular within the medical imaging community. A search for citing papers on IEEE Xplore returned 121 documents (performed on November 6th, 2014). Furthermore, the source code is freely available for download from the EMS website (EMS).

The algorithm performs an intensity-based segmentation from multispectral MR images. Each voxel is classified into one of four healthy tissue classes (white matter, gray matter, csf, and other). Spatial information for each tissue type is incorporated using a Markov random field (MRF). Moreover, MR field inhomogeneities can be corrected. Finally, MS lesions are detected as outliers that are not well explained by a model for normal brain MR images.

Results

We apply the same parameters as proposed on the EMS website (EMS) in Section "Additional intensity and contextual constraints." An overview of parameters and corresponding values is given in Table 6.7. The computed mask contains a value for lesion probability at each voxel.

The reference data used in this work consist of twenty patient data sets (Styner et al. 2008) with manually segmented lesions acquired on a 3T MR scanner. The data-acquisition protocol includes T1-, T2-, and FLAIR-weighted images. The T1-weighted data are used

Parameter	Value
Modalities	$T1, T2, PD$
Order of Bias Field Polynomial Model	4
Type of Polynomial	3D
Mahalanobis Threshold	3
MRF	yes
Intensity Constraints	$i_{T2} \geq gm_{T2}$
Lesion Tissue	wm (white matter)
Other Outliers	no

Table 6.7. Parameters and values of the evaluated segmentation approach proposed by van Leemput et al. (2001) (EMS).

as input for the initial brain extraction algorithm and the global and local B-spline registration. The resulting lesion position map is computed from approximately 500 lesions with a volume range of 0.001ml to 4.38ml (mean=0.25ml). See Section 4 for more details.

Sixteen lesion phantoms have been generated with a total lesion load (TLL) ranging from 1.12ml to 7.18ml. Therein, we developed six phantom pairs that only differ in the texture model (Data3,...,Data8), which is varied from homogeneous to inhomogeneous (cf. Sec. 5.2). Furthermore, four additional phantoms were generated without this restriction (2x homogeneous, 2x inhomogeneous). See Table 6.8 for an overview.

Again the Dice similarity coefficient (DSC) is used to compare the segmentation results with our ground truth. An illustration of the algorithm results is given in Figure 6.7. Since the ground truth as well as the computed lesion mask contain a voxel-wise lesion probability, we threshold the values before computing the overlap measure. Only values with a lesion probability of $\geq 50\%$ are used. The average similarity measure for phantoms with homogeneous lesions is $mean(DSC_{homogen}) = 0.67$ ($min = 0.57$, $max = 0.77$). For phantoms with inhomogeneous lesions this value is reduced to $mean(DSC_{perlin}) = 0.49$ ($min = 0.35$, $max = 0.55$). Even more, no result for the homogeneous phantoms is lower than any result for the inhomogeneous phantoms. The largest difference is 0.23.

One reason for the low segmentation overlap for inhomogeneous lesions is the incorrectly computed amount of lesion tissue by the EMS algorithm. This results in a larger lesion mask compared to the underlying ground truth. See Figure 6.8 for a visualization. The yellow area, i.e., the voxels with $\geq 50\%$ lesion probability, is much larger for homogeneous lesions (cf. Fig. 6.8 (b)) than for inhomogeneous lesions (cf. Fig. 6.8 (d)).

Besides the segmentation overlap, our phantom also enables a comparison of the total lesion load. Instead of using only voxels with $\geq 50\%$ lesion probability, the whole lesion mask is used. The last rows in Table 6.8 give the total lesion load for homogeneous and textured lesions computed by the EMS algorithm (TLL-EMS). The mean overestimation for homogeneous lesions is 161%. For textured lesions, it is 224%.

HOMOGENEOUS LESIONS								
	Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8
TLL (<i>ml</i>)	3.76	5.08	2.06	3.28	3.20	2.66	7.18	5.19
DSC	0.66	0.68	0.71	0.77	0.57	0.65	0.70	0.61
TLL-EMS (<i>ml</i>)	6.57	8.00	4.21	5.60	5.84	4.92	13.52	11.05

LESIONS WITH TEXTURE								
	Data1	Data2	Data3	Data4	Data5	Data6	Data7	Data8
TLL (<i>ml</i>)	2.18	2.86	1.12	2.03	2.13	1.70	4.50	4.60
DSC	0.49	0.52	0.51	0.54	0.35	0.51	0.55	0.43
TLL-EMS (<i>ml</i>)	4.91	6.04	3.29	4.39	4.90	3.92	9.21	7.73

Table 6.8. (1st row) Total lesion load (in *ml*) of lesion phantom (TLL); (2nd row) Dice similarity coefficient (DSC) to compare the computed segmentation overlap with the ground truth; (3rd row) Total lesion load computed by segmentation method (TLL-EMS, in *ml*).

6.3.3. Discussion

Segmentation of MS lesions in brain MR image data is an important step towards diagnosis and patient follow-up, and has been widely investigated in recent years. Several algorithms ranging from manual to fully automatic have been proposed. Evaluation of these approaches is usually performed with both phantom as well as real patient data sets. Phantom data are especially used to detect algorithmic errors in early development stages.

A major point of criticism of phantoms is their overall data quality. Most phantoms have only a very limited anatomical and pathological variability and are thus not as challenging as patient data. This behavior can also be observed with our data: The computed values using our phantoms are about 33.7% to 41.5% higher compared to the values reported in (van Leemput et al. 2001) for patient data sets.

Nevertheless, a phantom-based evaluation is an excellent tool especially in early algorithmic development stages, since each phantom can be adapted to the specific needs of the developer. We have shown that changing parameters can yield lesion objects with increased complexity. In this work, the object structure is modified. The resulting inhomogeneous lesions provide a more challenging task for the segmentation algorithm with a lower segmentation overlap measure and a higher volume overestimation. Furthermore, the values now differ only by 7.2% compared to the values by Leemput et al. reported on patient data sets.

6.4. Segmentation and Quantification of Brain Tumors

In this section, we use our brain tumor software phantoms for the evaluation of a segmentation and quantification approach that has become popular in the field of cancer therapy

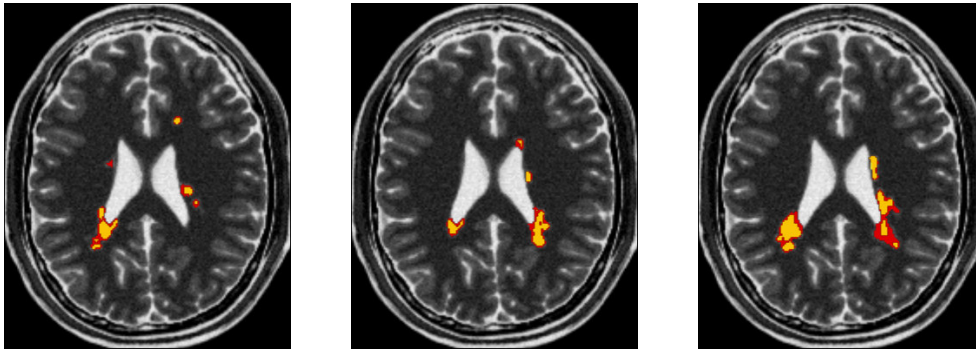


Figure 6.7. Phantoms with segmentation result overlay. The associated T1- and PD-weighted images used for segmentation are shown in Figure 5.10. Red: algorithm result (van Leemput et al. 2001) with $\geq 50\%$ lesion probability, yellow: phantom data with $\geq 50\%$ lesion probability.

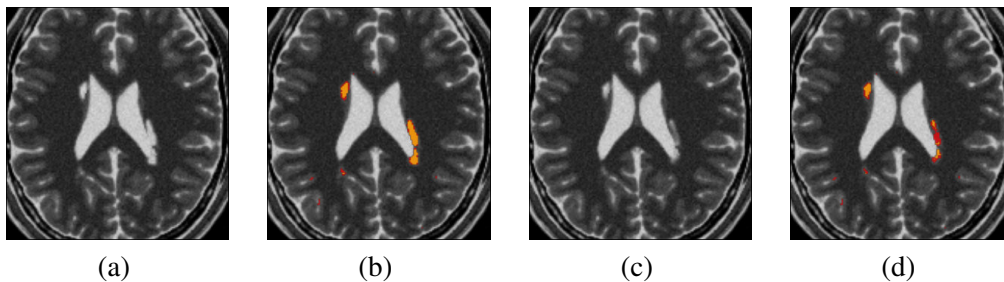


Figure 6.8. Comparison of segmentation results for homogeneous and inhomogeneous lesions. (a) Homogeneous lesions; (b) segmentation overlay on (a); (c) inhomogeneous lesions; (d) segmentation overlay on (c). Red: algorithm result (van Leemput, Maes, Vandermeulen, Colchester, and Suetens 2001), yellow: phantom data. Only voxels with $\geq 50\%$ lesion probability are used.

monitoring, namely OncoTREAT (Bornemann et al. 2007).

OncoTREAT facilitates a semi-automatic segmentation of different lesion types in CT data sets including lung nodules, enlarged lymph nodes, and liver metastases. Furthermore, a segmentation of brain metastases in MR scans included in the software prototype. The underlying algorithm combines an initial seeded region growing and an automatic morphological refinement to exclude adjacent structures with similar gray values. A detailed description can be found in (Kuhnigk et al. 2006). Since this approach requires a rather compact lesion appearance, a new method has been proposed for the segmentation of ring-enhancing tumors by Bornemann et al. (2007), which has shown to result in a more appropriate segmentation for many cases especially of brain metastases. In addition to the different segmentation tools, a fully automatic volumetry is implemented within OncoTREAT. For brain tumors, the software assistant uses a voxel counting approach, i.e., the volume is calculated as the number of voxels within the segmentation mask multiplied by the voxel volume.

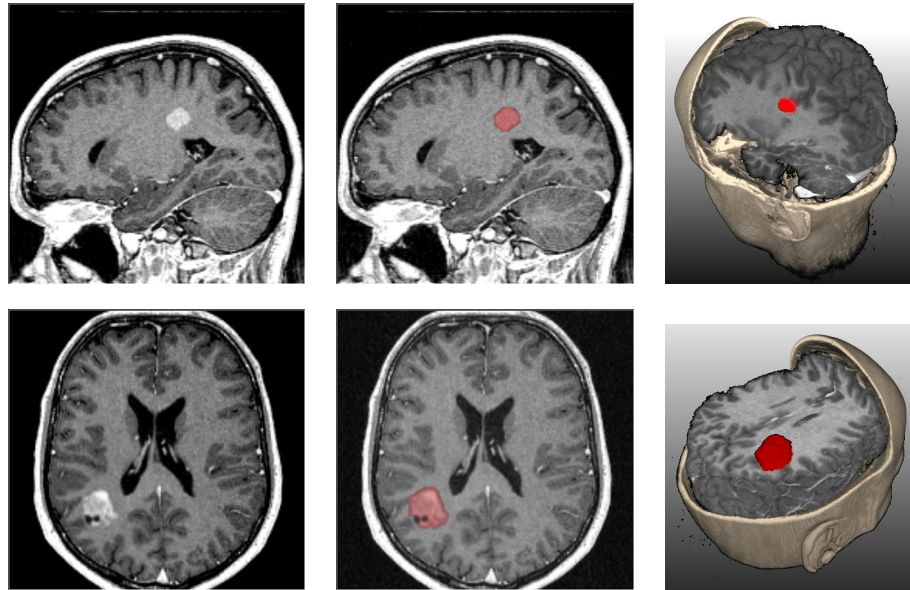


Figure 6.9. Segmentation results of software phantoms Case1 and Case2 using OncoTREAT. From left to right: Slice of T1 post contrast, 2D tumor segmentation overlay, and tumor segmentation in 3D.

Table 6.9. Evaluation of OncoTREAT results for all data sets using the Dice similarity coefficient. Overlap1: $\geq 50\%$ tumor tissue in reference segmentation mask, Overlap2: $> 0\%$ tumor tissue in reference segmentation mask

	Case 1	Case 2	Case 3	Case 4	Case 5
Overlap1	0.90	0.88	0.96	0.89	0.94
Overlap2	0.90	0.87	0.94	0.87	0.91

6.4.1. Results

Five data sets are analyzed representing different tumor shapes, locations, and sizes (Case1, Case2, ..., Case5). T1-weighted post contrast MR image data of a healthy volunteer are used as background, obtained from a clinical 1.5T scanner (Siemens Magnetom Vision, Siemens, Erlangen, Germany; 256x256 matrix; 1.0mm^3 isotropic voxel size). Case1 shows a rather homogeneous tumor shape that is composed only of active tumor. The tumor objects in Cases2 to Case5 contain a model for active tumor tissue and necrosis. Furthermore, in Case3 the amount of necrotic tissue is scaled to 50%. See Figure 6.9 and Figure 6.10 (left) for an example of each data set. The volume and the overlap statistics for all data sets are computed from the combination of both tissue types.

Table 6.9 shows the resulting similarity values for each tumor phantom. For comparison,

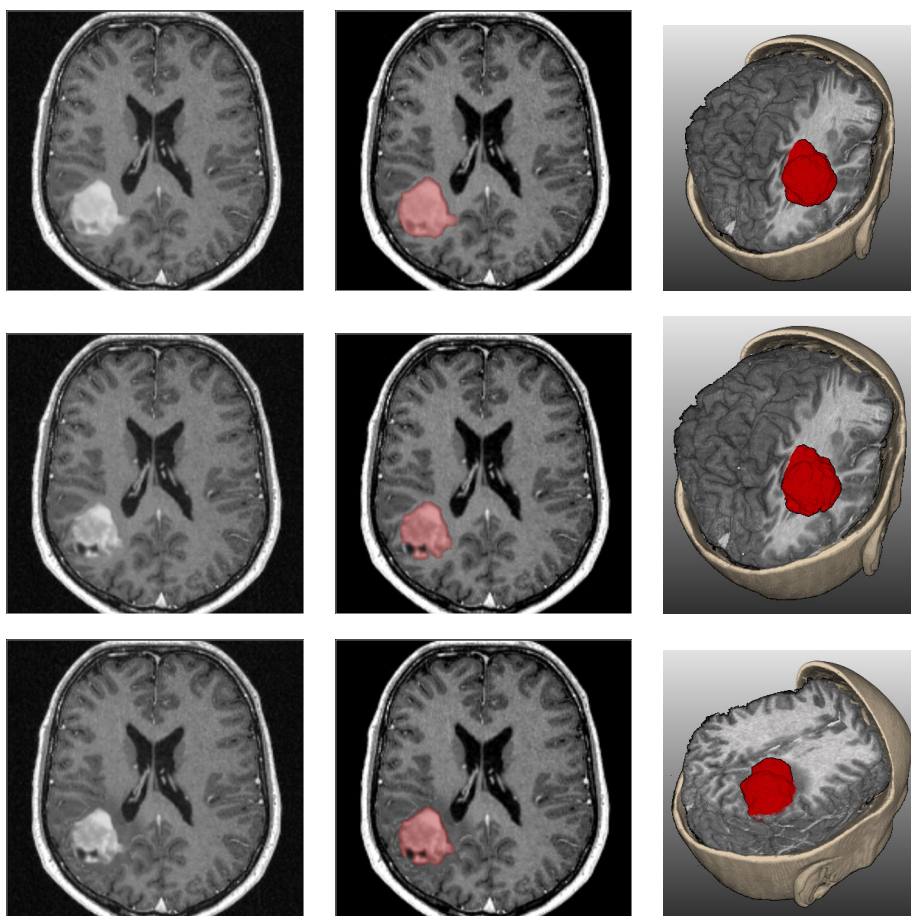


Figure 6.10. Segmentation results of software phantoms Case3, Case4, and Case5 using OncoTREAT. From left to right: Slice of T1 post contrast, tumor segmentation overlay in 2D, and tumor segmentation in 3D.

we used two different reference segmentations: The number of voxels with $\geq 50\%$ tumor tissue in the segmentation mask, and the number of voxels with $> 0\%$ tumor tissue. All cases show a high degree of overlap for both reference segmentation masks. The mean Dice similarity coefficient is $DSC_{mean} = 0.92$ using a manual segmentation as reference, and $DSC_{mean} = 0.90$ for the ground truth reference. The segmentation results also show a good visual correspondence with the underlying ground truth (cf. Fig. 6.9 and 6.10).

Besides an evaluation of the segmentation overlap, we also calculated the tumor volume as well as the volumetric error (cf. Table 6.10). OncoTREAT uses the voxel counting approach as already delineated above, which is still the gold standard in medical image quantification. The resulting volumetric error ranges between -17.24% and 5.263% . Especially Case4 and Case5 underestimate the volume by more than 10% . Here, the segmentation al-

gorithm does not include the whole tumor due to necrosis and PV voxels, which has already resulted in a lower overlap value (cf. Table 6.9).

Table 6.10. Estimated tumor volume and volumetric error for all data sets using the approach implemented in OncoTREAT.

Case	Ground Truth (in ml)	Volume (in ml)	Error (in %)
Case1	2	1.951	-2.450
Case2	8	8.421	5.263
Case3	25	25.771	3.084
Case4	25	20.690	-17.240
Case5	25	22.234	-11.064

6.4.2. Discussion

Data quality was already discussed as important issue in the previous section. The computed similarity coefficients for our phantoms have very high values, ranging between [0.87 – 0.96]. Again, we conclude that our phantoms are not as challenging as typical patient data sets. It is interesting to note that the similarity coefficient is slightly lower using all voxels with $> 0\%$ tumor tissue, especially for large tumors. Here, an important factor are PV effects at the tumor boundary. Furthermore, our software phantoms consist of areas at the tumor boundary that have a higher amount of necrosis, resulting in a lower gray value. Both characteristics yield in a set of voxels that are difficult to identify as tumor tissue on the data set by the segmentation algorithm, and thus are missing in the final tumor mask.

Despite the very accurate segmentation, the tumor volume is not estimated correctly using the voxel counting approach. Similar results were already obtained for the MS lesion phantoms in the previous sections. For the brain tumor phantoms, the inhomogeneous object appearance leads to a volumetric underestimated in most cases.

7. Conclusion Part I

In this chapter, we discuss the main objectives related to phantom development and our contributions.

How to develop phantoms for medical image analysis?

A phantom is an artificial object with known properties used to test certain aspects of an algorithm within a chosen application domain (cf. Chap. 2). Based on this definition, we can derive several requirements for phantom development in medical image analysis.

The first step towards developing a phantom is to decide what kind of phantom is required for the considered task. Therefore, we started our investigation in this work with a general overview of design approaches and proposed a categorization scheme that partitions software phantoms into three categories: stylized phantoms, voxel phantoms, and hybrid phantoms. We focus on hybrid phantoms, which provide a separation between object and background. Based on the knowledge of the underlying application domain and the phantom type, we can focus on the actual design in the next step.

In Chapter 4, we proposed a formalization consisting of three main parts: The design of a suitable object, the design of a related background, and the incorporation of one or more of these objects into the background. Each part of a phantom is based on certain hypotheses and parameters. Therefore, the next step of the phantom design is to choose and to model the relevant parameters. In Chapter 3, we provided a detailed examination of parameters used in phantom development. To help the user in the selection process, we proposed a classification into four groups, each modeling a certain aspect of a phantom. The parameters are selected based on the examination of work in the field of medical image analysis. For example, segmentation algorithms make assumptions about the underlying models for typical parameters such as the volume or the gray value. The shape or the volume of an anatomical or pathological structure are also important markers for diagnosis and therapy monitoring. Furthermore, the selected imaging parameters are commonly used in quality assurance (QA) programs of MR scanners. After modeling all required parameters, we can combine them to our final phantom.

What are the benefits of our approach?

Modular Design. In this work, we aim at a new approach for modular phantom design based on a set of components. To this end, we formalized the overall development process in Chapter 4, which is an essential step towards reusing components in a software assistant.

This yields a flexible standard of the required processing steps and encourages developers to thoroughly investigate all aspects of a phantom before starting the actual design process. Moreover, it provides a straightforward description of phantoms when only limited knowledge about the design process is available. For this task, we developed a template data sheet that is used throughout this thesis. For example, we used it in Part II to initiate a discussion with experts about our phantoms. The template turned out to be an efficient basis for our discussions.

Parameter Reuse. Besides a phantom description, our implemented parameter modules can also be easily changed without affecting other modules. For example, a different object shape does not change the underlying model for intensity values. This eventually saves time, e.g., through reuse of parameter models for new phantoms, and helps to eliminate the risk of design errors. Several parameter models such as noise or volume developed for MS lesion phantoms were later reused to develop our brain tumor phantoms. Moreover, new parameters can be added. For example, the brain tumor phantoms used the MS lesion phantoms as basis and extended them by additional modules such as tumor growth.

Similar to parameter reuse, our design process also allows to develop different classes of phantoms with varying complexity by simple parametric manipulations of one parameter. In this work, we developed two sets of phantoms with equal object positions, shapes, etc., but with different object texture. Thereby, our design process allows to change the texture model without affecting other parameters such as noise. The results have been used to compare how a segmentation algorithm is affected by modifications of the input data. A comparable analysis is not possible with patient data sets. Different appearances for a single object are usually not available or need to be acquired over a long time period. Moreover, such follow-up scans require scanner and imaging parameters equal to the initial data, which might not always be feasible.

To the best of our knowledge, this is the first approach towards a comprehensive description of the phantom design process in medical image analysis. Thereby, a controlled and documented construction of parameter models and phantoms provides an excellent overview of the underlying parameters and assumption.

Software Assistant. Using our component-based phantom design, we introduced an interactive software assistant that provides an easy-to-use tool for phantom development (cf. Sec. 4.3). In this work, we used the software assistant to develop the manual phantoms proposed in Chapter 5. The development time for each phantom is reduced from hours to a few minutes depending on the number of objects per phantom. Different implementations are available, for example to generate a suitable object shape. Furthermore, a set of ten pre-defined shapes is already available.

Fully Automatic Phantom Design. Besides the above mentioned manual approach, we also proposed a fully automatic method for phantom design. Thereby, the component-based design helps us to concentrate on the automatization of one parameter at a time. Instead of following the common design approach and of developing the one *best* phantom, we introduced a method to automatically generate a large set of phantoms. The majority of our parameter models are based on actual patient data such as lesion shape and position. This way, our approach allows us to capture the variability encountered in clinical practice. Furthermore, our design process enables a large applicability of the resulting phantoms as well as of the individual parameter models. In other words, the development of a parameter model should start with an analysis of the already available feature pool.

Phantom Diversity. Our modular design approach introduces new perspectives for a broad range of phantoms in medical image analysis as described above. However, some limitations remain. First of all, we focused on hybrid phantoms. Although this enables a practical separation between object and background design, incorporating an object into this background must be possible. Furthermore, we focused on background consisting of volunteer or patient images. Therefore, we can only cope with pathological structures added to the underlying data. This will lead to rather compact object shapes in most cases. Modeling other structures such as whole organs, e.g., the heart or the liver, is not yet possible. An extension to this approach could be a brain phantom. In this case the background is generated from a patient or volunteer data set, where the brain tissue is cropped out algorithmically, leaving only the skull. The object then consists of brain structures such white matter, gray matter and fluid, which is incorporated into this cavity.

Two clinical applications were analyzed more closely in this work, namely the analysis of Multiple Sclerosis (MS) lesions and of brain tumors. The quality of each phantom is visually confirmed by an expert, even if this can only give a hint of the actual quality. Nevertheless, further applications with different modalities such as CT are possible as well. For example, the analysis of lung nodules and hepatic lesions, or coronary artery plaque detection in angiographic images. A validation including a measure of the overall phantom quality as well as the quality of each module will be the focus of Part II in this work.

Ground Truth for Applications. Besides different phantom types, modalities, etc. our phantom design approach covers the evaluation of a wide range of algorithms that occur in clinical practice and trials. An evaluation study for each algorithm was proposed in Chapter 6. For example, an extensive analysis of MS lesion volumetry over a broad range of different lesions is performed in Section 6.2. Our results clearly show the importance of an improved gold standard in lesion volumetry beyond voxel counting. A similar study would not have been possible with patient data sets.

Furthermore, we evaluated an algorithm proposed by van Leemput et al. (2001). Lesion segmentation and volumetry results were compared with the phantom ground truth data. Both approaches show that our inhomogeneous lesions provide a more challenging task with a lower segmentation overlap measure and a higher volume overestimation. An analysis of

the computed lesion mask showed that the computed lesion probability at each pixel is too high for low-contrast lesion pixels.

Part II.

Phantom Validation

8. Validation in Medical Image Analysis

An important aspect of any development in medical imaging is a thorough analysis of the underlying assumptions and the modeled parameters with respect to their quality. Especially the validity of a method needs to be demonstrated, having considerable impact on its value to the user, and no model should be accepted before it has passed elaborate testing. Moreover, the level of confidence in a method increases, as it passes more tests. The purpose of this chapter is to give an overview of validation in medical image analysis. We discuss current concepts and provide a general overview of terms and definitions. Furthermore, we introduce prominent validation approaches for common fields of application including image segmentation and quantification. Thereby, phantoms used as reference data play an important role.

8.1. Principles and Definitions

Validation is characterized by an iterative process to demonstrate the compliance of a method with a set of given criteria. A method has to pass elaborate testing before it is accepted. Thereby, the level of confidence increases with the number of passed tests. In contrast to the verification of a system that ensures that *a method is built right*, i.e., that it includes all required parameters, validation determines that *the right method is built*. Thereby, validation does not give an answer to the question 'is the method correct?', since a system specification is rarely precise, and it is impossible to test a method under all possible events. Therefore, no method is ever 100% accurate. Instead, validation ensures the degree of required accuracy for its underlying purpose, which can vary from one application to another. As a consequence, this also implies the exact knowledge of a methods' objectives before it can be validated. For example, a segmentation algorithm may have been validated for use in brain tumor segmentation. However, this does not necessarily enable a valid use for the analysis of liver tumors. In this work, we use a definition of validation similar to that given by the American Institute of Aeronautics and Astronautics (1998):

Definition: Validation

The process of determining the degree to which a model is a sufficiently accurate representation of reality from the perspective of its intended uses.

Today, most algorithms are accompanied by some kind of validation study, and several examples are given in this chapter. A key criterion is that a method must model the clinical setting, i.e., a validation should consist of a broad spectrum of actual clinical use cases (Gee 2000). A number of publications focusing on this aspect have been proposed as well, each suggesting a slightly different method with its own terminology (Buvat et al. 1999; Jannin et al. 2006; Udupa et al. 2006). Nevertheless, a widely accepted approach is not yet available. Besides algorithm development in a research setting, commercial products should especially follow a rigorous validation process. For example, the Food and Drug Administration (FDA) provides general principles for the management and control of software validation in medicine. In their guidelines, validation is used to assure software quality by assessing the fulfilled requirements.

In order to assess the degree of compliance with the underlying purpose of a method, e.g., the above mentioned validation of clinical use cases, validation metrics are an invaluable aspect, and we will provide a broad overview of used measures in the upcoming sections. Each application has its own set of examined parameters, depending on the required level of validation, e.g., early development stage vs. late stage. For example, methods used in segmentation and quantification are discussed in Section 8.6. Therein, validation parameters consist of overlap ratios between the analyzed algorithm and the ground truth. Nevertheless, several other parameters often have to be considered as well, including computation time or susceptibility to errors. The availability of common data sets and reference algorithms used as benchmark for comparison are also important contributing factors to an effective validation.

8.2. Phantom-Based Validation

Both physical and software phantoms are used for validation purposes. Thereby, the complexity of the applied phantom depends on the considered application as well as on the current stage of the development process. Simple geometric phantoms such as a square or a sphere are often used during initial development and evaluation phases (Noe and Gee 2001). More complex phantoms are applied in later stages. Unfortunately, phantom development is also a challenging task, and we already analyzed several issues in the previous chapters. For example, anatomical and pathological structures and organs are highly variable with respect to their appearance, size, or shape. Moreover, each parameter requires an application-specific modeling, and modifying one parameter also affects others.

Besides their advantages and weaknesses, an explicit analysis of the quality of phantoms

is still missing. In other words: *'Why does a phantom represent an adequate validation approach for a certain method?'* In Section 2.3, we introduced a rather simple validation approach based on a qualitative analysis with fuzzy terms (good, average, low). Although this method yields a general categorization (cf. also Sec. 3.4), an in-depth analysis of a phantom is infeasible. Today, only few validation approaches are used for phantoms in medical imaging. A common strategy uses visual inspection and expert knowledge about anatomical or pathological structures. (Tofts et al. 1997; Segars et al. 2001; Suryanarayanan et al. 2005). Other publications propose a validation based on visual comparison with patient data sets or with images from literature (Pupi et al. 1990; Tofts et al. 1997; Prastawa et al. 2005). However, a method that provides a systematic analysis of the accuracy and precision of a phantom is not available, and some phantoms are even proposed without any validation. Furthermore, several phantoms have only a very limited parameter distribution. For example the popular BrainWeb phantom is created from a single individual (Collins et al. 1998). To this end, we propose a new phantom validation approach in the next chapter.

8.3. Validation without Phantoms

If a phantom is not available, one has to resort to other sources that allow validating a method. A number of approaches have been proposed, and each technique has its own pros and cons. We roughly distinguish between methods that are either based on expert knowledge or on all sorts of databases.

8.3.1. Expert Validation

An approach particularly suited for early development stages of a new method is the so-called *face validation*. Therein, field experts are asked if the results of a method are reasonable. Face validation provides a quick and rather informal validation. To give an example, a physician's opinion is a suitable way to determine the quality of a new segmentation algorithm during its conceptual phase. It allows for the identification of gross errors such as an under-segmentation or a leakage into adjacent structures. Unfortunately, accurate measurements are difficult to perform with the human eye, leaving the computer as supplementary tool. A further potential uncertainty can result from inferior visualizations. Moreover, reproducibility is often limited. For example, we performed a study to evaluate visual assessment in MS lesion volumetry in Section 6.1.

A related approach to obtain expert knowledge is to perform a comparison to other (valid) methods that have been proposed for the same application, and that are used as expert knowledge instead. This enables an overview of similar algorithms and provides information about important parameters. For example, Hagemann et al. (1999) extract appropriate elasticity constants for a physically-based registration algorithm of the head from a comprehensive literature study. In their work, the mean value calculated from nine different publications finally serves as an estimate of the correct values.

Another way to incorporate expert knowledge into the validation process is based on manual measurements such as manual segmentation of anatomical structures. An overview of related approaches is given in Section 8.6.1.

8.3.2. Databases

Another validation approach is the use of image data and associated segmentations from databases. Thereby, gathering appropriate image data and acquiring segmentations also requires expert knowledge. Nevertheless, this knowledge differs from expert validation as discussed above since it is rather used to assemble the data. Today, several public research resources are available in different fields of medical image analysis such as computer aided detection of calcifications in mammograms. A well-known initiative is the Visible Human Project that was established in 1989 to build a digital library of volumetric image data (Spitzer et al. 1996). The public domain data sets consist of complete, anatomically detailed representations of the human body based on MR and CT scans as well as cryosection images of male and female cadavers.

One of the most popular databases for the evaluation of methods in mammography was introduced by Karssemeijer (1993), containing 40 digitized film-screen mammograms of different clinically relevant cases with associated ground truth. Another, much larger, public database is the Digital Database for Screening Mammography (DDSM), maintained by the University of Florida. The DDSM database contains approximately 2.500 studies (Heath et al. 1998).

In the field of lung cancer detection, a cooperative effort known as the Lung Image Database Consortium (LIDC) was launched in year 2000 to construct a database that contains CT scans from both diagnostic and screening studies (Armato et al. 2004). Various information are stored for each data set in addition to the actual image data including technical scan parameters, patient information, and nodule features (McNitt-Gray et al. 2007). The assessment of lesion boundaries is based on manual outlining performed by expert radiologists.

A related project is the Reference Image Database to Evaluate Response (RIDER) that was initiated in 2004 (Armato et al. 2008). This initiative provides a web accessible public database of images for different organ systems. Currently, RIDER consists of a CT image archive of lung cancer subjects collected longitudinally over the course of treatment. An important aim of this project is the assembly of so-called validated data sets. This includes expert measures such as RECIST as well as associated meta-data including radiologist annotations. Furthermore, RIDER is intended to provide standardized methods for the evaluation of algorithms. To this end, a set of training and test data sets will be established.

An open and freely accessible database related to Alzheimer's Disease has been developed as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI). The database includes 200 elderly controls, 400 individuals with mild cognitive impairment, and 200 individuals with Alzheimer's disease. See Mueller et al. (2005) for an overview of the

initiative.

Besides databases for a specific modality or a specific body part, approaches focusing on the underlying image analysis task have been proposed as well. A synthetic database for the validation of brain image registration and segmentation methods is proposed in (Ens et al. 2009), including image data from 50 patient MR scans (T1-weighted) as well as a labeled brain atlas and an averaged multi-patient brain atlas. The database is generated in a three-step process: The first step consists of a linear transformation to register all data sets to the same coordinate system. In the second step, a non-rigid registration scheme with an elastic regularization is applied to match the labeled atlas data to the data of the average atlas. Additionally, three different methods are used to obtain a registration from the atlas scan onto each of the 50 patient MR data sets. The last step applies the computed deformation fields to transform the labeled atlas data to each patient data set. The resulting deformation fields can be used as basis for the validation of registration approaches. The labeled data sets can be used to evaluate segmentation methods. A major drawback of this approach is the lacking definition of the correct deformation field. Each registration method produces a slightly different deformation and each has its eligibility. However, the right deformation is still unknown. Furthermore, only an indirect and thus less accurate labeling of the patient data is available with this approach.

8.4. Gold Standards

Today, most studies compare the quality of a method with some kind of reference. Such a *gold standard* is presumed to contain the correct result (the *ground truth*) or be at least close to it. For example, segmentation methods typically use manually outlined image data (cf. Sec. 8.6.1). Therefore, establishing an appropriate gold standard is an essential task for validation. In the previous section we already described different types of gold standards. One approach is the use of phantoms since they allow direct access to the modeled parameters (cf. Sec. 8.2). Depending on the underlying application, the ground truth can also be based on manual segmentations or a database consisting of patient data. If no ground truth is available, other methods are required that serve as gold standard. Examples include the STAPLE algorithm by (Warfield et al. 2004) or a comparison by field experts (cf. Sec. 8.3).

But what makes a reliable gold standard? In a seminal paper, Lehmann (2002) proposes five attributes that should be considered:

1. *Reliance*: The development must follow an exactly determined and reproducible protocol.
2. *Equivalence*: The generated data must be equal to real-life data w.r.t. all important parameters.
3. *Independence*: The resulting data must rely on a different procedure than that to be evaluated.

4. *Relevance*: An evaluated algorithm must be self-reproducible, i.e., robust to parameter changes.
5. *Significance*: The number of gold standard images used for evaluation must be sufficiently large.

Based on this characterization, Lehmann (2002) then defines a reference standard as follows:

Reference Standard	Fulfilled Criteria
Gold Standard	(1),(2),(3)
Silver Standard	(1),(2) or (1),(3)
Plastic Standard	otherwise

Properties (4) and (5) are used to define a meaningful evaluation of an image processing algorithm.

In Chapter 2 we defined similar requirements for the design of phantoms, namely suitability, flexibility, and correctness. Based on our classification, a 'perfect' phantom – and thus a good gold standard – should fulfill these three requirements. Especially suitability and flexibility overlap with the attributes that are needed for the gold standard definition by (Lehmann 2002). Particularly equivalence, i.e., the quality of the considered parameters. Nevertheless, the definitions of reliance and independence deviate from our requirements: Although we agree to an exactly determined and reproducible protocol (reliance), phantom design can still include a great amount of manual work, which is explicitly prohibited by Lehmann. Furthermore, independence between training and test data is a crucial aspect during algorithm validation, but should have less importance for phantom design.

In conclusion, we believe that the parameters used by (Lehmann 2002) provide an excellent approach to characterize a reliable gold standard, and to our knowledge this is to date the only attempt to provide a unified classification scheme for this task. However, the three features are somewhat mixed with the task of algorithm validation. The work rather provides an upper bound for a gold standard, and a phantom that meets our requirements is already sufficient.

8.5. General Approaches

After introducing general principles of validation, we now focus on actual methods used in medical image processing. These approaches include general frameworks as well as more specific tasks such as the validation of visual assessment and segmentation. In this section, we provide a survey of methods that are concerned with a standardization of the overall validation process regarding terminology and methodology: Some papers are related to algorithm evaluation (Buvat et al. 1999), or discuss the validation of algorithm performance (Gee 2000). Other papers focus on the validation process itself (Jannin et al. 2006).

Fryback and Thornbury (1991) analyze the efficacy of diagnostic imaging. They propose a hierarchical model outlining the contribution of diagnostic imaging to the patient management process. Six levels are distinguished: (1) Technical efficacy, (2) diagnostic accuracy efficacy, (3) diagnostic thinking efficacy, (4) therapeutic efficacy, (5) patient outcome efficacy, and (6) societal efficacy. Additionally, typical measures for analyses are provided for each level. For example the analysis of ROC curves (Level 2: diagnostic accuracy efficacy) or the number of cases in which an image is judged helpful to making a diagnosis (Level 3: diagnostic thinking efficacy). An important aspect of the presented hierarchy is that demonstrating the efficacy on a lower level is necessary, but not sufficient, to assure efficacy at higher levels. For example, the quality of scanner parameters (Level 1: technical efficacy) has to be assessed before a diagnosis of the resulting patient data can be evaluated (Level 2: diagnostic accuracy efficacy).

Buvat et al. (1999) propose a hierarchical approach to algorithm evaluation. They apply the six levels of efficacy defined by Fryback and Thornbury (1991) to the evaluation of medical image processing. Level 1 is related to validation or feasibility studies of a method; level 2 is associated with the method accuracy by measuring its performance; level 3 measures whether a method helps a clinician; level 4 evaluations determine that a method provides information, which contributes to the appropriateness of patient management; level 5 efficacy corresponds to how a method affects patient outcome; level 6 efficacy is related to the benefit of a method for the society.

Buvat et al. (1999) also present a generic evaluation model (GEM) that identifies required components and provides guidelines for evaluation studies. The main components include an abstract aim, a context, and a hypothesis. The abstract aim results from the method to be assessed and the level of evaluation. The context defines the underlying environment in which a method is evaluated. It also includes the components required to define the evaluation protocol. Finally, the hypothesis is related to tests performed within a context to extract information about the abstract aim. A detailed description including further components required during an evaluation process are given in the paper.

An approach that introduces a model to describe the main components used for validation procedures in medical imaging is proposed by Jannin et al. (2006). It presents one of the few more recent general validation methods in medical image analysis. The model focuses on level 1 and level 2 efficacy. As an example, a relation of these levels to image registration is given. Therein, level 1 efficacy is associated with performance criteria of an algorithm, whereas level 2 efficacy is related to an evaluation at clinically meaningful anatomical or pathological structures. The validation process model is related to the evaluation model of Buvat et al. (1999), where a clinical context is specified as well as a clinical objective, i.e., a hypothesis. The validation process aims at testing this hypothesis. Jannin et al. (2006) base their model on the evaluation of data sets and input parameters, given a reference method that estimates the ground truth. Thus, the model provides a detailed description of common reference-based validation scenarios. An illustration is given in Fig-

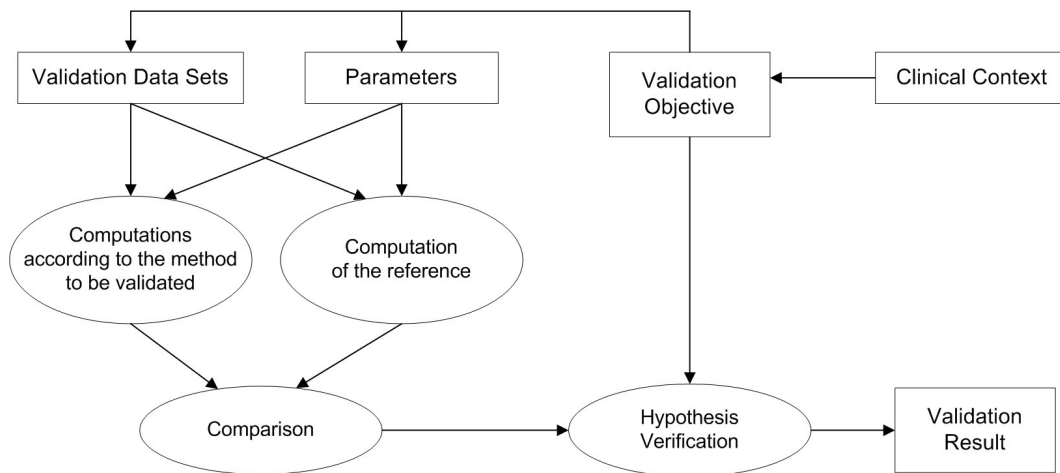


Figure 8.1. Overview of reference-based validation model presented in (Jannin et al. 2006).

Figure 8.1. Besides a detailed description of the process model, Jannin et al. also provide a checklist of components that should be included when reporting a validation study. The presented model is validated on the basis of 38 papers that include validation studies. Results are summarized in a database and are accessible via a website (VMIP).

An approach that is not only related to algorithm validation, but to the whole development process is presented by Thacker et al. (2008). In their work, the authors propose a characterization of algorithms based on the analysis of representative tasks in computer vision including feature detection, stereo vision, or face recognition. Moreover, measuring structural differences in medical images is examined. Central aspects of algorithm performance evaluations are extracted, resulting in a set of key questions such as 'how is testing currently performed?', 'is there a data set for which the correct answers are known?', or 'are there data sets in common use?'. Answering these questions during the development and evaluation of an algorithm is stated to be an important step towards validation.

8.6. Applications

After introducing rather general approaches to validation in medical image analysis, the following sections cope with specific applications.

8.6.1. Segmentation

Segmentation of medical image data is a crucial task for the analysis of normal and pathological processes, where small changes can already be of paramount importance. Today a vast amount of different methods is available for various applications and levels of automation, ranging from manual outlining to fully automatic segmentation. Many approaches

have also been proposed for the validation of these methods. See for example (Udupa et al. 2006) and references therein. However, a widely accepted performance measure is not yet established, making it still difficult to compare two different segmentation methods.

In order to provide a systematic evaluation framework, several authors proposed a categorization of segmentation algorithms into different classes. Zhang et al. (2008) propose a categorization into two major categories: subjective and objective evaluation methods. Thereby, a subjective evaluation is related to visual inspection by a human observer as already described in Section 8.3.1. Further categories are supervised and unsupervised methods, where supervised refers to the use of a ground truth image, and unsupervised can do without one. Udupa et al. (2006) present an evaluation framework consisting of five components: (F1) a specification of meaningful metrics of efficacy, (F2) real life image data, (F3) reference segmentations (ground truth) (F4) a number of available standard segmentation algorithms, (F5) a software incorporating evaluation methods and segmentation algorithms. In their work, an application is determined by three entities:

T: A task.

B: A body region.

P: An imaging protocol.

Each realization of these entities presents a specific application domain $\langle T, B, P \rangle$, and a general statement or a comparison between different application domains cannot be made. Due to the required reference data (component F3), the work of Udupa et al. can be assigned to the category of supervised evaluation methods.

Despite the difficulties to evaluate a segmentation algorithm, most papers provide at least some kind of analysis. Especially applications, where public databases or software phantoms are available, play a pioneer role. For example, a comparison of different brain tissue segmentation methods is performed in (Cuadra et al. 2005; Hahn et al. 2004). Another evaluation of seven algorithms for the same task is presented in (Bouix et al. 2007).

Unfortunately, a comparison of different algorithms is often difficult as most of them are not freely available, and a re-implementation is time-consuming or often even not possible. A more promising approach is to provide a public database with training and test cases where the ground truth for the test data is not publicly available. This way, a research group does not have to release their algorithms to the community. Another approach, that has become popular in the past years, is to arrange an on-site competition of several research teams during a conference. At the MICCAI conference 2007, the first 'Grand Challenge' workshop was held, attributed to the comparison of liver segmentation methods. A summary can be found in (Heimann et al. 2009). Several subsequent workshops for other organs and pathologies such as head or MS lesion segmentation have been carried out since.

Similarity Measures for Supervised Evaluation

Several metrics have been proposed to measure the similarity of two sets of segmented voxels. Two commonly used measures based on the regional overlap are the Jaccard Similarity J (Jaccard 1912) and the Dice coefficient DSC (Dice 1945), given as

$$J(S, T) = \frac{|S \cap T|}{|S \cup T|} \quad (8.1)$$

$$DSC(S, T) = \frac{2|S \cap T|}{|S| + |T|} \quad (8.2)$$

J	Jaccard similarity measure $\in [0,1]$
DSC	Dice similarity coefficient $\in [0,1]$
$ \cdot $	stands for the number of elements
S	set of voxels segmented by algorithm 1
T	set of voxels segmented by algorithm 2 (or ground truth).

Both coefficients are equal to zero if the voxel sets S and T cover disjoint regions, and one if they are identical. Thereby, the Dice coefficient is always larger than the Jaccard metric, except at 0 and 1. Furthermore, both values are related by the function $DSC = 2J/(J + 1)$. A value $DSC > 0.7$ is often reported to indicate 'excellent agreement' between segmentations (Zijdenbos et al. 1994).

A drawback of such measures is that they only consider the number of overlapping voxels and do not take into account their positions. To this end, Pichon et al. (2004) proposed a generalization of these global overlap measures, using the distance between the segmented voxels to the ground truth, defined as:

$$d(x) = \begin{cases} 0, & x \in S \cap T \\ \min_{s \in S} \|x - s\|, & x \in T \setminus S \\ \min_{t \in T} \|t - x\|, & x \in S \setminus T \end{cases} \quad (8.3)$$

d	resulting distance measure
x	voxel $\in S \cup T$
S	set of voxels segmented by algorithm 1
T	set of voxels segmented by algorithm 2.

The mean value and the standard deviation over all voxels with a distance of $d > 0$ are used to measure the segmentation error. Furthermore, the error distance of the worst $f\%$ voxels is used. A related approach is presented by Crum et al. (2006). Here, the authors introduce a fuzzy multilabel overlap value plus an associated error measure based on the overlap distance.

Besides using one specific approach for evaluation, some authors have proposed to calculate the quality of a segmentation method based on a combination of different measures.

Moretti et al. (2000) derive several evaluation criteria based on the BrainWeb software phantom for the analysis of brain tissue segmentation methods. This includes overlap ratios and a distance map from the object's contour. Furthermore, a distance histogram as well as derived features such as the mean and the standard deviation are computed from the distance map. Gerig et al. (2001) developed a tool for validation and comparison of object segmentation. The tool uses several algorithms, again including overlap ratios and distance measures. Both 2D cross-sections with label overlay as well as 3D visualizations are available to illustrate differences between algorithms, or between an algorithm and the gold standard. The tool is freely available from the web.

While the above methods provide a separate analysis for each measure, Cardenesa et al. (2009) propose a multidimensional evaluation technique. To this end, a vector of different similarity measures is assembled and the overall similarity is defined as the l_2 -norm of this vector. Moreover, a principal component analysis (PCA) is applied to allow for a dimensionality reduction and a 2D visualization of the results.

Similarity Measures for Unsupervised Evaluation

As described above, unsupervised methods work without any reference image. Zhang et al. (2008) analyze 16 evaluation methods used in the field of computer vision applications. For example, intra-region metrics such as texture metrics measure the texture uniformity, assuming that a 'smoother' image is preferred. Other metrics use differences between regions, e.g., the inter-region color difference. A third class of measures is related to shape information. Furthermore, a combination of all three classes is presented.

In the field of medical imaging, unsupervised methods have become popular with the introduction of the STAPLE (simultaneous truth and performance level estimation) algorithm by Warfield et al. (2004). The algorithm allows for an evaluation of a set of segmentations by simultaneously estimating performance parameters (sensitivity and specificity) and the hidden true segmentation for each voxel, based on an expectation-maximization (EM) algorithm. STAPLE can also be used to estimate a reference image. Thereby, the algorithm computes the highest consensus between segmentations, which is typically assumed to be close to the ground truth. Details of the algorithm can be found in (Warfield et al. 2004).

Bouix et al. (2007) propose the Williams Index as efficient alternative to the STAPLE method. Considering a set of r raters and a similarity $s(X_i, X_j)$ between rater i and j , e.g., the Jaccard similarity measure, the Williams Index for rater i is defined as

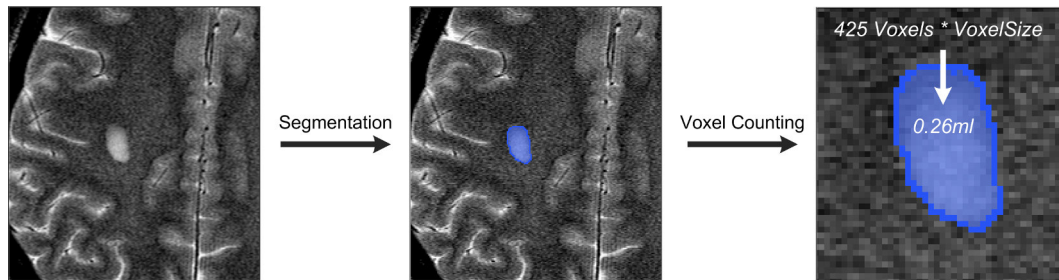


Figure 8.2. Current gold standard in medical image quantification: The voxel counting approach.

$$WI_i = \frac{(r-2) \sum_{j \neq i}^r s(X_i, X_j)}{2 \sum_{j \neq i}^r \sum_{k \neq i}^{j-1} s(X_j, X_k)} \quad (8.4)$$

WI_i	Williams' Index for rater i
r	number of raters
X_i	set of voxels segmented by rater/algorithm i .

If this index is greater than one, it can be concluded that rater i agrees with the other raters at least as much as they agree with each other. In their work, Bouix et al. (2007) found similar results for this measure compared to the STAPLE algorithm.

8.6.2. Quantification

Today, validating quantification methods is largely based on the comparison with results from domain experts. We therefore term these approaches supervised similar to the categories used for segmentation evaluation. Unsupervised methods on the other hand, i.e., a validation without a reference standard, are rather uncommon. One approach could be to perform a reproducibility study, where an object is quantified several times. However, this can only provide information about the precision of an algorithm and not about its accuracy.

Similar to image segmentation, a ground truth is often difficult to obtain especially for small objects such as tumors, and a 'surrogate' ground truth has to be used. The voxel counting method (cf. Fig. 8.2) represents the current gold standard for validation, where the ground truth is defined by counting the number of voxels from an expert's hand labeling. A common issue with this approach is that human observers are known to be quite variable during manual segmentations. Furthermore, partial volume effects are not accounted for.

8.7. Discussion

Validation is gaining growing attention in medical image analysis. Today, several workshops at major conferences also focus on this topic, e.g., the 'Grand Challenge' workshops at the MICCAI conference. Other works include application-specific frameworks (Gedamu et al. 2008; Neumuth et al. 2009) as well as specific tasks such as segmentation (Udupa et al. 2006). However, these are typically not used by research groups different from those of the initial authors. Moreover, validation is rarely the main objective in current publications, but often regarded as an add-on to the development of a new method. Exceptions from literature include procedures suggesting a standardization of the validation process. For example, validation protocols have been presented by Buvat et al. (1999) or Jannin et al. (2006). Unfortunately, a widely accepted method is not yet available.

In this chapter, we have given an overview of current validation concepts in medical imaging. Each application has its own set of validation methods. Nevertheless, we propose a common classification into supervised and unsupervised methods for each application, where supervised refers to a validation based on some reference image. In addition to this categorization, we determined different approaches to validate a method, including validation methods using a phantom and validation methods without a phantom.

If no phantom with known ground truth is available, databases are a common alternative. Thereby, associated segmentations of anatomical structures such as brain tissue classes serve as a surrogate ground truth, and this approach is the gold standard for many applications. A major issue when using databases is the continually evolving technology used to create the data. For example, changes in acquisition protocols, e.g., due to a larger field strength, or new imaging sequences will result in a different set of data in clinical practice. A database used for training and evaluation will have to adapt to these changes, which can result in a huge effort and financial burden. Another issue is the required sample size of the database. If the sample size is too small, an algorithm will adapt to the data and not be able to generalize to new data sets (Gee 2000). A potential solution could be a public repository where data can be freely submitted. However, data sets from different sites can have a different quality, e.g., different amount of imaging artifacts, and merging these data in one database could make the evaluation of algorithms even more challenging. Moreover, databases consisting of patient data do not provide ground truth for many parameters such as the volume or the shape of an imaged organ.

Expert knowledge is another validation approach which does not require a phantom. It is especially useful during early development stages of a method. Unfortunately, expert knowledge is subjective and often not reproducible, and also consensus meetings with several experts can not provide a ground truth.

Validation of Phantoms

In addition to the methods discussed above, phantoms provide an excellent basis for validation. Because the ground truth of the modeled parameters is known exactly, they provide a convenient approach compared to other data that can merely act as surrogate ground truth such as manual segmentations of patient data. Furthermore, the development of phantoms can concentrate on important features of a method, and each application requires its own type (cf. Chap. 2). For example, physical phantoms are often used for quality assurance of medical devices, whereas software phantoms often present a more flexible alternative for algorithm validation.

Today, phantoms are used for many applications in medical imaging. However, developing a phantom can be time-consuming and cost-intensive, and only few are freely or commercially available. Similar to an expert segmentation of a patient data set, the quality of the underlying ground truth is now in the hands of the phantom developer. Moreover, the representativeness of the modeled parameters for the targeted application has to be considered during the design process. Therefore, validating a phantom is particularly important.

Lessons Learned

To summarize, important characteristics of validation in medical image analysis are as follows:

1. Validation is application-dependent.
2. Validation is an iterative process.
3. Each validation step increases the confidence in a method.
4. It is often impossible to validate all aspects of a method.
5. Sufficient data are essential for the success of validation.

In the next chapters, we propose different approaches to validate a phantom.

9. Validation of Phantoms

Phantoms have become a widely accepted tool for validating new algorithms and imaging equipment. Nevertheless, an objective evaluation of the accuracy and precision has not been proposed yet, and most approaches are based on an ad hoc design. In the first part of this chapter, we recapitulate major challenges of phantom validation and identify a link to the overall requirements for phantom development already addressed in Chapter 2. Our approach enables the validation of a single phantom as well as the comparison of different phantoms.

To assess the phantom quality, we then propose a novel iterative validation approach. Each iteration consists of a separate validation method that increases or decreases the confidence in the phantom. For example, one application is a user-study to analyze the detection performance of lesion phantoms. Furthermore, the effect of parameter changes is investigated. Thereby, we propose a new method based on a multi-criteria decision making technique, the so-called Analytic Hierarchy Process, to select a parameter of high importance.

Our overall validation approach consists of three major steps: method selection, method validation, and finally phantom validation. Two attributes are considered, namely the suitability of a method and its correctness. A validation function is proposed that combines both parameters. An example application will be used in the next chapter to evaluate our approach.

9.1. General Approach

An important characteristic of a phantom is the availability of an exactly known ground truth for the modeled parameters. Therefore, it is widely used as gold standard, i.e., as a reference used to analyze a method. But how to validate a phantom? In Chapter 8 we defined validation as *the degree to which a model is a sufficiently accurate representation of reality from the perspective of its intended uses*. Applied to phantoms, validation assesses the degree a phantom is an adequate reference for the targeted application. Unfortunately, a proper analysis of phantoms is difficult, and often only the developer of a phantom is able to fully understand and test it. Current validation approaches are often based on an expert-based visual comparison with patient data sets or with data from literature. A systematic

analysis has not been proposed yet.

An important step towards phantom validation is to identify the underlying challenges. A first step is to determine core features that need to be examined for successful phantom validation. In Chapter 8, we analyzed validation approaches and provided an overview of methods commonly used in medical imaging. This eventually led us to a summary of the most important characteristics in Section 8.7. These results are now applied to the process of phantom validation. The three most relevant aspects are:

1. **It is often impossible to validate all aspects of a phantom.** (*Correctness*)

A key issue for the validation of phantoms is a commonly lacking parameter correctness. Although the ground truth might be available for some parameters, the correct model and an appropriate parameter distribution for all parameters are often difficult to obtain or not known at all. Therefore, phantoms can only act as a surrogate ground truth, and the simplified model might not be suited for the underlying task. For example, pathological structures such as tumors have a high variability, and surgical resection would be required to determine the true appearance.

2. **Validation is application-dependent.** (*Suitability*)

Each application requires its own phantom and many examples were given in the previous chapters. For instance, a phantom used to analyze contrast enhancement characteristics is different from one used to evaluate a segmentation algorithm. Even a phantom used for one algorithm might not be suited to evaluate another. The challenge is to identify the relevant parameters for an application. Eventually, this will also lead to a different validation procedure for each application.

3. **Sufficient data are essential for the success of validation.** (*Flexibility*)

Developing a physical phantom can require a large manual effort depending on the modeled structure. Even for the more flexible design process of software phantoms, generating a distribution of different models is time-consuming. For example, consider building an adequate model of the gray matter of the brain for a large amount of phantoms. Therefore, only a small number of phantoms or even a single exemplar is typically used during algorithm evaluations in medical image analysis.

To summarize, validating a phantom requires methods that assess three general properties, namely correctness, suitability, and flexibility. Thereby, correctness and flexibility are related to the analyzed phantom, whereas suitability is related to the targeted application. The same parameters were already addressed as general requirements for phantom development in the first background chapter of this work (cf. Fig. 2.1).

Table 9.1. Scale used to characterize the correctness and suitability of a module. Intermediate values between two levels (2, 4, 6, 8) do not get a separate naming.

Explanation	poor		low		average		high		very high
Value	1	2	3	4	5	6	7	8	9

9.2. Processing Steps

A simple validation approach could be to let an expert directly assign a value between zero and one, with one being a phantom of high quality. Unfortunately, expert knowledge is subjective and often not reproducible. In order to derive a more sophisticated approach, let us have a look at the phantom design process. In Chapter 3, we started with an analysis of the parameters typically required for phantoms. For each relevant parameter, we then developed an appropriate model. Finally, the generated lesion objects are incorporated into a given background.

In this work, we propose a novel phantom validation method that is derived from this design process. For this approach we evaluate a number of methods (*Method Selection*) with respect to their quality and relevance for the targeted application (*Method Validation*). The final phantom validation is then determined as a combination of all methods (*Phantom Validation*).

Step 1: Method Selection

The first step of our algorithm is an in-depth analysis of the targeted application. In this step, a set $M = \{M_1, M_2, \dots, M_N\}$ of methods is selected. Furthermore, the importance ($suit : M \rightarrow [0, 1]$, $suit(M_i) = s$) is determined for each considered criteria M_i .

The values are calculated via an expert validation, which provides a quick and informal approach. However, expert decisions are based upon multiple subjective criteria and are heavily affected by accumulated experience. Moreover, expert decisions are based upon multiple criteria and are heavily affected by accumulated experience. In which several experts work together towards common recommendations. A related approach is the comparison with information extracted from published work. A more detailed overview can be found in Section 8.3.

In this work, we propose a set of qualitative criteria given in fuzzy terms: *poor*, *low*, *average*, *high*, *very high*. The corresponding function values are integral numbers from 1 to 9. Thereby, the values 1, 3, 5, 7, and 9 are associated with one of the criteria given above, while the other values (2, 4, 6, 8) characterize intermediate levels (cf. Tab. 9.1). Finally, the results are normalized by the function $f_{expert}(z) = z/9$.

Table 9.2. The main processing steps of the phantom validation process.

<p>Step 1. Method Selection</p> <p>Determine the set of relevant methods and quantify their suitability.</p> <p>Step 2. Method Validation</p> <p>Define measures for each method based on the functions <i>corr</i> and <i>suit</i>.</p> <p>Step 3. Phantom Validation</p> <p>Compute an overall validation measure or a ranking of available alternatives.</p>
--

Step 2: Method Validation

Besides the relevance of a method, the quality of the computed results is another important aspect. Therefore, the second validation step determines the correctness of a method and combines both parameters within a validation function. The correctness is defined similar to the suitability by a function $corr : M \rightarrow [0, 1]$, $corr(M_i) = c$. Another approach could be to use a comparison with reference data. However, adequate reference data are difficult to obtain. For example, the correct shape of a complex object such as the white matter can only be estimated from patient data.

The third attribute to be tested is the flexibility. For example, a module that is very difficult to develop and requires substantial manual work could result in an overall lower validation value than a module that does not require a difficult development process. Unfortunately, estimating the amount of flexibility is difficult. Moreover, even a module that is difficult to create can still be of a high quality, so that this attribute should only be considered as a minor factor during module validation. In fact, we do not include the flexibility parameter during phantom validation.

The validation function $v_{M_i} : M \times M \rightarrow [0, v_{max}]$, $v_{M_i}(c, s) = v$ combines the suitability and correctness values. Important features of this function can be summarized as follows: We aim at a multiplicative aggregation of all validation functions in the phantom validation step (cf. Tab. 9.2), so that a value $v = 1$ represents a somewhat neutral result. To this end, methods with low suitability ($suit \ll 1$) should result in values close to one. In other words, there is no need to heavily weight unimportant methods, independent from their correctness. On the other hand, an important method that is not correctly modeled should result in a low value $v \ll 1$, reducing the confidence in the phantom. Finally, important and correctly modeled methods should result in a value $v > 1$, i.e., we increase the confidence in the phantom. Table 9.3 provides a summary of these features. Based on these properties, we propose a module validation function given by

Table 9.3. Summary of general features of validation function v .

1. Not required criteria do not affect the validation result ($v(c, 0) = 1, c \in [0, 1]$).
2. Important and correctly modeled criteria get a high value ($v(1, suit_{max}) = v_{max}$).
3. Important and not correctly modeled criteria get a low value ($v(0, suit_{max}) = 0$).
4. $v(c, s) > 1$ enhances the confidence in the phantom.
5. $v(c, s) < 1$ decreases the confidence in the phantom.

$$v(c, s) = \frac{v_{max}}{suit_{max}} \cdot s \cdot c - \frac{1}{suit_{max}} \cdot s + 1 \quad (9.1)$$

v	Module validation function
c	Measure of module correctness ($corr(C_i) = c$)
s	Measure of module suitability ($suit(C_i) = s$)
v_{max}	Maximum value of v
$suit_{max}$	Maximum module suitability.

The function proposed in Equation 9.1 reaches its maximum value v_{max} for methods with a maximum relevance and correctness. Due to the normalization function used in Step 1 (Method Selection) $f(z) = z/9$, we select $v_{max} = 3$ as an arbitrary value within the range $1 < v_{max} < 9$. The maximum input suitability is given by $suit_{max} \in [0, 1]$.

The function is derived using the following steps: Taking into account the first and third item in Table 9.3, we initially define a linear function between $v(0, 0) = 1$ and $v(0, suit_{max}) = 0$ as

$$v_1(0, s) = 1 - \frac{1}{suit_{max}} \cdot s, \quad v_1 \in [0, 1].$$

From the first and second item in Table 9.3 we then derive a second linear function between $v(1, 0) = 1$ and $v(1, suit_{max}) = v_{max}$ given by

$$v_2(1, s) = 1 + \frac{v_{max} - 1}{suit_{max}} \cdot s, \quad v_2 \in [1, v_{max}].$$

The proposed module validation function finally results from linear combination of a point in function v_1 with a corresponding point in v_2 . See Figure 9.1 for a visualization.

Step 3: Phantom Validation

The last step combines all validation functions v_{M_i} , resulting in an overall measure for the analyzed phantom. The multiplicative aggregation results in

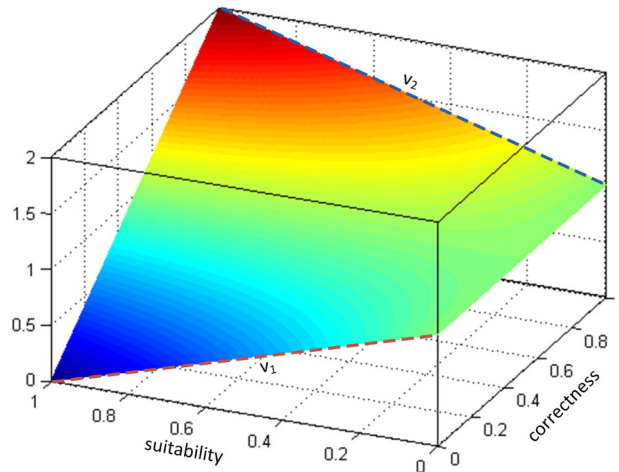


Figure 9.1. Visualization of the validation function introduced in Equation 9.1. The underlying functions v_1 and v_2 are plotted as dashed line (see text for a description).

$$v_{phantom}(P) = \prod_{i=1}^N v_{M_i}(P) \quad (9.2)$$

$v_{phantom}$	Phantom validation function
P	Analyzed phantom
v_{M_i}	Validation function for module M_i , $i = 1, \dots, N$.

A value of $v_{M_i} > 1$ denotes an improved phantom confidence. Thus, evaluating many correct and relevant methods will result in a large value for $v_{phantom}$. On the other hand, a single method with $v_{M_i} \ll 1$ is sufficient to greatly reduce the overall phantom validation result.

If the goal is a comparison of different phantoms, the result can also be a ranking of the available alternatives. Thereby, the aim is to evaluate a set of alternatives describing a separate phantom. The best alternative should receive the highest value.

9.3. Iterative Phantom Validation

Phantom validation is an iterative process that requires several steps to reach a reasonable confidence level. In other words, not only one but several methods should be used to validate a phantom, each increasing or decreasing the confidence. Unfortunately, a proper analysis of phantoms is difficult.

In this section, we introduce an iterative phantom validation approach that combines the results of several separate validation methods to an overall measure of the phantom quality. In Chapter 10 we will discuss these methods in the context of MS lesion phantoms.

9.3.1. Fusion of Parameter Validations

The first method introduces a direct analysis of the phantom based on an expert analysis. The suitability of this method is given by the expert's experience. The correctness value, i.e., the phantom quality, is calculated by fusing the validation of all parameters to a single measure. Two features need to be investigated for each parameter: suitability and correctness.

The Analytic Hierarchy Process

We introduce a validation approach, that allows for an explicit analysis of the relationships between modules. The proposed method is based on the so-called Analytic Hierarchy Process (AHP). AHP is a multi-criteria decision making technique developed by Thomas L. Saaty (Saaty 1977; Saaty 1990). It has become one of the most widely applied tools for a variety of decision situations, e.g., for ranking a set of alternatives or for making the best choice from a number of options. Instead of analyzing all parameters at once, the AHP approach uses pairwise comparisons, which has demonstrated in studies to be an efficient and accurate approach.

A literature review including several areas of applications such as engineering, government, or education can be found in (Vaidya and Kumar 2006). A review focusing on healthcare and medical decision making based on AHP is given in (Liberatore and Nydick 2008). The authors investigate 50 articles, classified into seven categories such as diagnosis, therapy/treatment, or project and technology evaluation and selection.

The AHP approach starts by sorting out all parameter modules that are not important for the targeted application. The definition of a hierarchy of different levels is an essential aspect of the AHP. Thereby, modules of an upper level are only influenced by modules of the adjacent lower level. Based on this hierarchy, the suitability of the remaining modules is examined. A frequently used method in psychology is to use pairwise comparison, since it has turned out to be easier and more accurate to choose between two elements than simultaneously between all available alternatives. The AHP uses the same approach. In other words, a pairwise comparison is performed of (1) all object modules and of (2) all background modules. Then, the two parameter sets are averaged, meaning that we weight the object and background design equally important. Finally, the resulting values are normalized.

The comparisons of object and background modules can each be organized into a matrix. For N modules, this yields an $N \times N$ matrix C with

$$C = \begin{pmatrix} c_{11} & \dots & c_{1j} & \dots & c_{1N} \\ \dots & \dots & \dots & \dots & \dots \\ c_{i1} & \dots & c_{ij} & \dots & c_{iN} \\ \dots & \dots & \dots & \dots & \dots \\ c_{N1} & \dots & c_{Nj} & \dots & c_{NN} \end{pmatrix}, \quad (9.3)$$

C Comparison matrix for N modules
 c_{ij} Matrix entry delineating the comparison between element i and j .

The matrix has the following properties:

- $c_{ij} > 0$, $i, j = 1, \dots, N$.
- $c_{ij} = 1/c_{ji}$, $i, j = 1, \dots, N$.
- $c_{ij} = 1$, $i = j$.

Before we now derive the actual module suitability from C as proposed in the AHP approach, we need to define an adequate judgment scale. The AHP commonly uses a ratio scale consisting of values from one to nine (Saaty 1990) as given in Table 9.4. A value of one denotes two modules of equal importance, whereas nine indicates the highest suitability difference between two modules. A description for all values of this scale is given in Table 9.4.

To give an example, let us consider a phantom developed to evaluate liver segmentation algorithms. Thereby, we define the object contrast to be more important than the position parameters, and parameters modeling processes will be much less important. The corresponding matrix entries are then: $c_{img,pos} = 3$, $c_{pos,proc} = 7$, $c_{img,proc} = 7$. The resulting matrix M will then be

$$C = \begin{pmatrix} 1 & 3 & 7 \\ 1/3 & 1 & 7 \\ 1/7 & 1/7 & 1 \end{pmatrix}.$$

After assembling the comparison matrix C , we can compute the suitability for all modules at a given hierarchy level. The AHP derives the weights for each module from an eigenvalue analysis of C

$$C \cdot p = \lambda_{\max} \cdot p \quad (9.4)$$

C Comparison matrix (cf. Eq. 9.3)
 λ_{\max} Maximum eigenvalue
 p Eigenvector (priorities) corresponding to λ_{\max} .

The eigenvector p corresponding to the largest eigenvalue of V , normalized by dividing by its sum (i.e., $\sum_{i=1}^N p_i = 1$), contains the suitability for each module. If all modules have equal importance, the resulting matrix V is singular and the maximum eigenvalue is $\lambda_{\max} = N$. In

Table 9.4. The 1 – 9 fundamental scale as defined in (Saaty 1990).

Intensity of importance	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective
3	Moderate importance	Experience and judgment moderately favor one activity over another
5	Strong importance	Experience and judgment strongly favor one activity over another
7	Very strong importance	An activity is strongly favored and its dominance demonstrated in practice
9	Extreme importance	The evidence favoring one activity is of highest possible order of affirmation
2, 4, 6, 8	Intermediate values	

this case, each module gets a suitability value of $1/N$. Besides the normalization of the priorities, a further normalization can be performed across various hierarchy levels. Such a composition eventually leads to a global suitability measure of each parameter module, reducing the complexity of upcoming validation steps. For our example above, the resulting suitability measures are $p_{img} = 0.633$, $p_{pos} = 0.3043$, $p_{proc} = 0.0627$.

An interesting feature of the AHP is that it provides an inherent consistency analysis of the performed pairwise comparisons, i.e., of the comparison matrix of a certain level, and thus a quality measure of the determined judgments. A consistency index CI has been proposed by Saaty (1977), which is related to the eigenvalue analysis described above.

$$CI = \frac{\lambda_{\max} - N}{N - 1} \quad (9.5)$$

CI	Consistency index
λ_{\max}	Largest eigenvalue
N	Dimension of comparison matrix M .

A further normalization of this index can be applied, resulting in the so-called consistency ratio CR . This value is computed as the ratio of the consistency index CI and a random index RI

$$CR = CI/RI \quad (9.6)$$

CR	Consistency ratio
CI	Consistency index
RI	Random index.

Table 9.5. Random indices RI as calculated in (Saaty 1977).

n	3	4	5	6	7	8	9	10
RI	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

The random index RI is commonly defined by averaging the CI values of 500 randomly filled comparison matrices based on the AHP fundamental scale (cf. Tab. 9.4). See (Saaty 1990) for a detailed discussion. Table 9.5 shows the values for RI for matrices of order three to ten as given in (Saaty 1977). A high CR value corresponds to a low matrix consistency, and the entries, i.e., the comparisons, should be reevaluated. Consistency ratios of less than 0.1 are usually considered acceptable. A ratio of $CR = 1$ characterizes inconsistent and rather random value selection.

Method Validation

To summarize, the AHP approach computes the degree of suitability for all object and background modules. The next step is now to integrate our results into the validation method. Since the sum of all AHP values is equal to one, we normalize each value by dividing it by the largest output. This way, the most relevant parameter receives $suit_{param} = 1$. The second feature for a parameter, the parameter correctness, is assigned via an expert analysis. The resulting overall phantom correctness is then defined as the minimum of all parameter validations

$$corr_{phantom} = \min(\forall v_{param}).$$

Equation 9.1 is used to calculate the result of each validation function v_{param} . In other words, we choose the lowest combination of correctness and suitability normalized over all parameters. The input values for this validation method are then

Suitability	$suit = \text{expert experience} \in [0, 1]$
Correctness	$corr = \min(\forall v_{param})$

9.3.2. Number Of Parameters

The next validation method is extracted from our development process. Each phantom models a certain amount of object and background parameters. Assuming that a larger number of modeled parameters results in a better phantom, this value is an adequate validation step. Since this approach does not include any statement about the quality of the modeled parameters, we set the importance of this approach to be only moderate ($suit = 4/9$), i.e., we select a value between 'low' and 'average'. The correctness is calculated as ratio of the number of modeled parameters and the number of relevant parameters for the targeted application. To

summarize, the input values for the method validation (Step 2) are

Suitability	$suit = 4/9$
Correctness	$corr = (\#modeled\ params) / (\#relevant\ params)$

9.3.3. Lesion Detection Performance

Another interesting approach is to analyze the object detection performance compared to patient data sets. Here, the user is asked to identify phantom objects in a patient data set that contains similar objects. Unfortunately, this approach does not give any information about the background used in the phantom data sets. It can also only be used in patient data with more than one object. Therefore, we choose the importance of this approach to be moderate ($suit = 6/9$). The correctness should be derived from the computed detection performance.

Suitability	$suit = 6/9$
Correctness	$corr = \text{derived from detection results}$

9.3.4. Segmentation Overlap Measures: Phantom vs. Patient Data

Phantoms have become a widely accepted tool for the evaluation of segmentation algorithms. They have a known ground truth and provide a good alternative to the analysis of patient data sets. A good phantom should be able to predict the performance of an evaluated algorithm on patient images never seen before. Comparing the segmentation results for phantoms and patient data sets is thus a good way to derive the phantom quality.

An important assumption of this approach is that all relevant parameters are modeled. Nevertheless, a poor parameter model can still result in an unrealistic phantom with comparable segmentation overlap. Nevertheless, we believe that this approach is suited to evaluate the quality of a phantom, and therefore assign $suit = 7/9$. The reference values can be extracted from publications related to the segmentation algorithm or from own investigations, i.e., from segmentation of own patient data sets. The correctness values are estimated from the segmentation overlap using the Dice similarity coefficient.

Suitability	$suit = 7/9$
Correctness	$corr = \text{comparison of segmentation overlap}$

9.3.5. Effect of Parameter Changes

Changing the complexity or the value range of a parameter has an effect on the resulting phantom quality. For example, a wrong object position will reduce the applicability of the

phantom. One way to test the effect of parameter changes is to compare the segmentation overlap for different parameter settings with those of patient data. Therefore, we believe that this approach is well suited to act as an indirect measure of phantom quality and assign $suit = 8/9$.

Our aim for this validation method is to compare phantom pairs where certain parameters are changed to some extent. Therefore, we want to use a parameter with high importance. The Analytic Hierarchy Process introduced in Section 9.3.1 is used for this task. An important assumption of this approach is that all relevant parameters are modeled.

Suitability	$suit = 8/9$
Correctness	$corr = \text{comparison of segmentation overlap}$

9.4. Discussion

A good phantom should be able to predict the performance of an evaluated algorithm on images never seen before. In other words, if the phantom is useful, the evaluated algorithm should produce comparable results for both phantom data and patient data. In this chapter, we introduced a novel phantom validation approach. We associate the overall requirements for phantom development given in Chapter 2 with those of phantom validation. Furthermore, our approach is closely related to the phantom development process proposed in Chapter 4.

Our validation approach supports the design process of these hybrid software phantoms already at the initial development phase by an analysis of the required parameters. We propose a method based on a multi-criteria decision making technique, namely the Analytic Hierarchy Process. An important aspect of this method is the use of pairwise comparisons, where each module is compared to all other modules within the same group. This way, dependencies between modules are explicitly modeled. In contrast, two important modules will both receive a high suitability value if analyzed separately.

The AHP method uses an expert analysis of the phantom. This method requires user involvement as well as expert knowledge of the underlying problem domain. Increasing the number of modules also increases the required pairwise comparisons. Analyzing an object including all modules introduced in this work results in $12 * 11/2 = 66$ comparisons, i.e., $N(N - 1)/2$ comparisons. Although this is a significantly higher number compared to a single evaluation of each parameter, we gain an explicit assessment of module dependencies.

Comparison with Related Work

Current phantom validation approaches are based on a visual comparison with patient data sets or apply some expert knowledge for evaluation (cf. Sec. 8.7). Nevertheless, at least some related work can be found that goes beyond these fairly subjective methods.

In Section 8.5, we reviewed several methods that propose a standardization of the overall

validation process. Even though these methods focus on the validation of image processing algorithms, some characteristics are also important for phantom validation. The work by Jannin et al. (2006) introduces a general validation concept (cf. Fig. 8.1). In their work, the authors propose a checklist of roughly 30 components that need to be observed when reporting a validation study. These parameters include the clinical context or the method to be validated. Similar to our approach, Jannin et al. (2006) aim at a formalization of the validation process. Furthermore, they also emphasize the application dependency of validation and delineate the clinical context and the method to be validated as important parameters. However, the authors propose a reference-based process, i.e., the underlying validation data is already known. In our work, we start one step earlier and focus on the analysis of the underlying reference data instead.

Another approach related to phantom validation is the analysis of ground truth data in medical image processing. In Section 8.4, we discussed the work by Lehmann (2002), who proposed five attributes that should be considered for any appropriate gold standard: reliability, equivalence, independence, relevance, and significance. Although this approach provides an interesting attempt towards a classification scheme for reference data, it is not accompanied by an objective measurement for the proposed attributes. Therefore, only a descriptive analysis can be performed, reducing the overall reproducibility between different raters.

Iterative Validation

Validation is an iterative process that requires several steps to reach a reasonable confidence level (cf. Sec. 8.7). Furthermore, it is application-dependent. In this work, we propose a novel iterative validation approach consisting of several evaluation methods. Each method increases or decreases the confidence in a phantom. Our validation approach is divided into three steps: *method selection*, *method validation*, and *phantom validation* (cf. Tab. 9.2). For each validation method phantom-related measures (correctness) and application-related measures (suitability) are analyzed. To combine these two measures, we proposed a validation function (cf. Eq. 9.1). The same function is used for all investigated methods.

In the last step of our validation approach, all validation methods are combined to a final measurement for the analyzed phantom. Our goal is an approach where the phantom quality is expressed in a single value. This allows for a straightforward comparison of different validation results among several raters or among different phantoms. We aim at a function, which ensures that unimportant methods have only a limited influence independent from their correctness. Important methods on the other hand can heavily increase or decrease the confidence in the phantom. To this end, we propose a multiplicative aggregation of all validation functions.

Based on the iterative fusion proposed in this chapter and the chosen validation methods, we believe that our approach allows for an estimation of the phantom quality. Thus, a highly rated phantom will be able to predict the performance on patient data. To the best of our

knowledge this is the first method of this type. Given a phantom we can evaluate to what degree this phantom is appropriate for a certain application, and also assess if one phantom is better suited than another. Moreover, phantom developers and users are supported in analyzing advantages and drawbacks of a phantom in more detail. This was also found useful for our own phantom development in the first part of this work and justifies the additional effort of the proposed validation approach. The next chapter will provide a more detailed analysis on the basis of a validation of our MS lesion phantoms.

10. Validation of MS Lesion Phantoms

In the previous chapter, we proposed a new phantom validation approach based on an iterative analysis. Each iteration consists of a separate validation method, which increases or decreases the confidence in the phantom. In this chapter, we apply our approach to the validation of our MS lesion phantoms proposed in Chapter 5.

Five validation methods are analyzed. For example, we perform an expert-based analysis of the phantom quality. The Analytic Hierarchy Process is used to derive the suitability of all required object parameter modules. Another validation method carries out a user-study to analyze the detection performance of lesion phantoms. Furthermore, the effect of parameter changes is investigated. For each step, a short description of the underlying data and a discussion of the results is provided.

10.1. Fusion of Parameter Validations

The first validation method is a direct analysis of the phantom. We use an expert validation for this method. A detailed description of this validation method is given in Section 9.3.1. The suitability is defined as the expert's experience. In our case, two raters performed the analysis our MS lesion phantoms. Both with several years of expertise in medical image analysis.

The correctness value, i.e., the phantom quality, is calculated by fusing the validation of all parameters to a single measure. Again, two features need to be investigated for each parameter: suitability and correctness. Since our focus in this work is on hybrid software phantoms, only the lesion object is analyzed. Nevertheless, an additional validation is conceivable that also includes the background.

To analyze the parameter modules of our phantoms, the first step classifies the parameter modules into the binary categories relevant and negligible. Modules that have no effect on the targeted application are sorted out. A consensus meeting was held to decide on the remaining parameters. The resulting comparison matrix consists of the following nine modules: shape, structure, volume, topology, contrast, noise, resolution, PV effects, and uniformity. All object parameters are modeled in our approach. Since the sum of all AHP values is equal to one, we normalize each value by dividing by the largest output. This way,

Table 10.1. Expert analysis for each parameter. The suitability values are calculated via the AHP approach described in Section 9.3.1. All parameters are normalized to $[0, 1]$.

Parameter	Suitability		Correctness	
	Rater1	Rater2	Rater1	Rater2
Shape	1.0	1.0	0.78	0.67
Structure	1.0	0.99	0.78	0.67
Volume	0.38	0.41	0.78	0.78
Topology	0.67	0.67	0.89	0.78
Contrast	0.81	0.85	0.78	0.89
Noise	0.33	0.28	0.89	0.89
Resolution	0.10	0.09	1.0	1.0
PV effects	0.29	0.36	0.78	0.89
Uniformity	0.09	0.06	0.78	0.89

the most relevant parameter receives $suit_{param} = 1$ (cf. Tab. 10.1).

The second feature, the parameter correctness, is assigned via an expert analysis. The results of this evaluation are also given in Table 10.1. The values for correctness and suitability are then used as input for the validation function (cf. Eq. 9.1), resulting in a value v_{param} for each parameter module. Finally, the overall phantom correctness for the current validation method is defined as the minimum of all parameter validations, i.e., $corr = \min(\forall v_{param})$.

Evaluation

Both experts rated their experience to be 'high' ($suit = 7/9$). The correctness values are extracted from the analysis of the parameter modules. Therein, the computed values are approximately the same for both raters. Four parameter modules get assigned high suitability values of over 65%, namely shape, structure, topology, and contrast. Furthermore, all modules got assigned a correctness value larger than five, i.e., better than average. The resulting overall phantom correctness, defined as the minimum of all parameter validations, is $corr = 0.78$ for Rater1 and $corr = 0.67$ for Rater2.

	Rater1	Rater2
Suitability	$suit = 7/9$	$suit = 7/9$
Correctness	$corr = 0.78$	$corr = 0.67$

The validation results are for Rater1 $v = 2.04$, and for Rater2 $v = 1.78$.

10.2. Number of Parameters

The first step of the AHP-based validation described above extracts all relevant modules and sorts out the rest. Assuming that the amount of modeled parameters directly corresponds to the phantom quality, we can derive a phantom validation approach. Similar to the AHP-based validation presented above, only the lesion objects are analyzed.

Evaluation

In Section 10.1, we extracted nine relevant parameters. All of them are modeled in our phantom approach. See also Figure 5.9 for an overview. Because we calculate the correctness as ratio of the number of modeled parameters divided by the number of relevant parameters, our phantom receives the highest value $corr = 1$.

An advantage of this approach, compared for example to the comparison of segmentation results, is that the phantom does not necessarily have to be available. A list of all modeled parameters is sufficient. Therefore, we investigated related work on MS lesion phantoms and carried out a similar parameter analysis. The phantom by Tofts et al. (1997) models five parameters resulting in $v = 1.30$, which is a decrease by 35.1% compared to our approach. The work of Melhem et al. (2003) models eight parameters. The validation result decreases by 8.7% to $v = 1.74$.

The resulting values for our validation function are

Suitability	$suit = 4/9$
Correctness	$corr = 9/9$

The validation result for our approach is $v = 1.89$.

10.3. Lesion Detection Performance

The next method to assess the phantom quality, is a human observer study. Therein, the detection performance of lesion phantoms in patient data sets with several real MS lesions is investigated. To carry out the study, we developed a software assistant with an intuitive graphical user interface. It consists of a viewer showing the current image data as well as a simple parameter panel to select the current lesion type. Two different marker types are available. A 'lesion' marker is used for real lesions and a 'phantom' marker for lesion phantoms. Six data sets from different patients were analyzed (Case1, Case2, ..., Case6). Two to three MS lesion phantoms were incorporated in each MR scan at different positions in the white matter (T2-weighted, matrix 256×256 , 3mm slice thickness). The size and intensity values were adjusted to match the real MS lesions in the underlying patient data. Only two different object shapes were used for all data sets. To reduce the detection task for each participant, only the lesions on one slice had to be detected. Nevertheless, the raters

Table 10.2. Results of human observer study averaged over six patient data sets with included lesion phantoms.

	Rater1	Rater2	Rater3	Rater4
Sensitivity (in %)	27.27	23.08	23.08	38.46
Specificity (in %)	67.57	86.00	81.13	56.10

were able to slice through the whole data set, whereas the slice to be rated was highlighted by a rectangle.

Two physicians (Rater1, Rater2) as well as two computer scientists with considerable experience in computational neuroimaging (Rater3, Rater4) participated in the study. All raters were blinded in the number of real MS lesions and lesion phantoms. All used the same computer and monitor devices to carry out the detection task. Figure 10.1 (c)-(f) show the resulting marked lesions for each expert (Case6). The corresponding ground truth is presented in Figure 10.1 (b). The raters correctly marked 26.9% ($min = 23.08\%$, $max = 38.46\%$) of thirteen lesion phantoms. The overall sensitivity was 27.97%, the overall specificity 72.70%. No significant differences were found in the resulting detection rate of all four raters. The values for each participant is separately given in Table 10.2.

Evaluation

The results clearly show the plausibility of our software phantom design approach with respect to clinical image data. A broad range of software phantoms can be generated that are indistinguishable from actual MS lesions for a human observer. The ability to correctly identify lesion phantoms, i.e., the sensitivity, is very low for all four raters ($<30\%$), even though we only used two different object shapes. Three out of four raters were not able to correctly mark more than 25% of all lesion phantoms. Only Rater4 had a slightly better overall detection rate of approximately 40%. However, the corresponding specificity of this rater is very low (56.10%). All participants suspected several real MS lesions to be lesion phantoms (cf. Fig. 10.1). Therefore, we assign a high correctness of $corr = 8/9$ to this validation approach.

Suitability	$suit = 6/9$
Correctness	$corr = 8/9$

The validation result for this approach is $v = 2.11$.

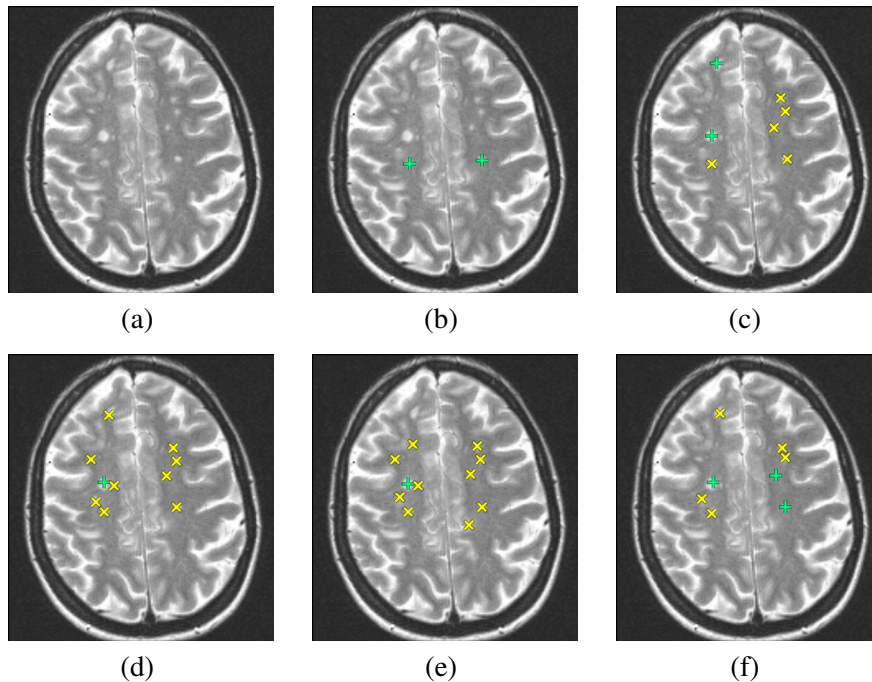


Figure 10.1. Lesion detection performance using a patient data set with additional lesion phantoms (Case6). (\times , real MS lesions (yellow); $+$, lesion phantom (green)). (a) Slice of original image data with incorporated lesion phantoms, (b) ground truth, (c)-(f) Rater1–Rater4.

10.4. Segmentation Overlap Measures: Phantom vs. Patient Data

In this validation step, we propose an evaluation of phantom quality using results of an adequate segmentation method. We apply the lesion segmentation algorithm proposed by van Leemput et al. (2001), which has become quite popular within the medical imaging community with more than 100 citing papers on IEEE Xplore. It is also freely available and therefore well suited as a reference method for comparison. See also Section 6.3.2 for a more detailed overview of the parameter settings. The segmentation results are shown in Table 6.8 (upper table).

Evaluation

To evaluate this validation step, we need to compare our results with those calculated on patient data. In (van Leemput et al. 2001), segmentations of 20 low-resolution patient data sets with an in-plane resolution of 0.9×0.9 and 5 mm slice thickness were performed. Furthermore, three data sets with a slice thickness of 2.4 mm are analyzed. Expert segmentations are used as reference. The best DSC value in the paper is $\max(DSC) = 0.45$ for the

low-resolution data sets and $\max(DSC) = 0.51$ for the high-resolution data sets.

Similar results were recorded by García-Lorenzo et al. (2011). In their work, the authors used the EMS method to segment ten patient data sets (in-plane resolution 0.97×0.97 , 3 mm slice thickness) with a total lesion load (TTL) ranging from 1.0 ml to 47.7 ml. Again, a manual expert segmentation is used as reference ground truth. The calculated DSC values have a large range of [0.31, 0.77] with a mean value of $\text{mean}(DSC) = 0.56$.

Our results are $\text{mean}(DSC) = 0.67$ with $\min(DSC) = 0.57$ and $\max(DSC) = 0.77$. In other words, our mean values differ from those of García-Lorenzo et al. (2011) by 16.4%. Furthermore, our max values differ from those of van Leemput et al. (2001) by 41.5% for the low-resolution data and by 33.7% for the high-resolution data. A moderate correctness is therefore selected for this approach ($\text{corr} = 6/9$). Nevertheless, we believe that this approach is suited to evaluate the quality of a phantom and assign $\text{suit} = 7/9$. The values for our validation function are given by

Suitability	$\text{suit} = 7/9$
Correctness	$\text{corr} = 6/9$

The validation result for this approach is $v = 1.77$.

10.5. Effect of Parameter Changes

Changing the complexity or the value range of a parameter has an effect on the resulting phantom quality. For example, a wrong object position will reduce the applicability of the phantom. One way to test the effect of parameter changes is to compare the segmentation overlap for different parameter settings with those of patient data. Therefore, we believe that this approach is well suited to act as an indirect measure of phantom quality and assign $\text{suit} = 8/9$.

An important assumption of this approach is that all relevant parameters are modeled. Nevertheless, a poor parameter model can still result in an unrealistic phantom with comparable segmentation overlap. Our aim for this validation method is to compare phantom pairs where certain parameters are changed to some extent. Therefore, our goal is to select a parameter of high importance. We choose the structure parameter for the subsequent evaluation, since it got assigned a high suitability value in the AHP analysis in Section 10.3 by two experts.

To assess the effect of parameter changes on the phantom quality, we use the same approach presented in the previous section. First, all phantoms are segmented with the MS lesion segmentation algorithm proposed by van Leemput et al. (2001). The Dice similarity coefficient computed from segmentation mask and ground truth is used to evaluate the results. The calculated values are then compared with segmentation results from real patient data.

In Section 6.3.2 we generated 16 phantoms from two different parameter setups. The

first set consists of a homogeneous lesion structure and was already used in the previous section. The second set includes inhomogeneous lesions using a parametric model to change the lesion texture (cf. Sec. 5.2). The resulting overlap measures and the total lesion load for each data set is given in Table 6.8 (lower table).

Evaluation

Changing the lesion structure to an inhomogeneous texture model has a large effect on the segmentation results. Because of the more complex lesion appearance, we expect a decrease in the overlap measure. In fact, the mean overlap measure decreases by 26.8% to $mean(DSC) = 0.49$ ($min(DSC) = 0.35$, $max(DSC) = 0.55$). Compared to the patient data results of van Leemput et al. (2001), the max values now differ only by 7.2%. We can further evaluate the effect of adding lesion texture using a Wilcoxon matched pairs signed-ranks test. The test results indicate that a significant difference exists between lesions with and without our texture model at the $p \leq 0.05$ significance level. Therefore, this validation approach receives a high correctness value of $corr = 8/9$.

Suitability	$suit = 8/9$
Correctness	$corr = 8/9$

The validation result for this approach is $v = 2.48$.

10.6. Result of Phantom Validation

After analyzing each validation method separately, the final phantom quality is computed by fusing all results as described in the previous sections. In this work, we analyzed five different validation methods. All results have a value $v > 1$ and thus increase the confidence in our phantom. The overall phantom validation value is

$$\begin{aligned}
 v_{phantom} &= v_{expert1} \cdot v_{expert2} \cdot v_{\#params} \cdot v_{detect} \cdot v_{segm} \cdot v_{changeParam} \\
 &= 2.04 \cdot 1.78 \cdot 1.89 \cdot 2.11 \cdot 1.77 \cdot 2.48 \\
 &= 63.71
 \end{aligned}$$

To understand how to interpret the validation result, we compare it with those of a slightly different phantom. Let us assume, that the object structure is not generated by the proposed texture model, but by a checkerboard pattern. In other words, only one parameter model is changed. This will still increase the lesion complexity and result in a lower overlap measure for the applied segmentation approach as described in Section 10.5. The low object quality will be detected by an expert validation. For the given example, we select a low correctness value of $corr_{param} = 1/9$ for both experts. The phantom validation proposed in Section 10.1 will then be $v = 0.48$ ($v = 0.48$) instead of $v = 2.04$ ($v = 1.78$). This will

decrease the result of the iterative phantom validation by 93.7% to $v_{phantom} = 4.03$. Even if only one expert assigns a low rate, e.g., Rater1, the result will decrease by 76.5% to $v_{phantom} = 14.98$.

If the new phantom is evaluated by different raters, interobserver variability is difficult to avoid. Nevertheless, our approach is able to detect the reduced phantom quality using the method described in Section 10.1. Rating the parameter correctness between 'poor' and 'low' with $corr_{param} = 2/9$ instead of $1/9$ for both raters will decrease the resulting phantom value by 84.1% to $v_{phantom} = 10.13$. An even higher value for both raters ($corr_{param} = 3/9$), i.e., a 'low' correctness, will still reduce the phantom validation value by 72.4% to $v_{phantom} = 17.57$.

10.7. Discussion

Phantom validation is an iterative process in which several steps are required to reach a reasonable confidence in a phantom. In the previous chapter, we introduced five different validation methods for this task that were applied to the analysis of our MS lesion phantoms in this chapter. The final phantom quality is computed by fusing the results of all methods. All methods increased the confidence in our lesion phantoms, i.e., all validation results have a value greater than one.

The first method performed an expert validation, analyzing all object parameters. Our approach goes beyond a simple visual rating of a phantom by only looking at it. Instead, we provide an explicit analysis of the relationships between parameter modules based on the Analytic Hierarchy Process, which has become a widely applied decision making technique. Our method uses pairwise comparisons of parameters, which has been demonstrated in studies to be an efficient and accurate approach. Two raters performed the validation for our MS lesion phantoms.

In another validation method, we analyzed the effect of object parameter changes. Again, the AHP approach is applied to select a parameter with high importance. The developed phantoms are then used to analyze a well-known MS lesion segmentation algorithm proposed by van Leemput et al. (2001). The resulting overlap measure for the lesion mask is compared with segmentation results from patient data. Results from published work of two different research groups are used for comparison. Our results indicate that it is worth developing complex parameters, that better reflect the challenges of related patient data.

Our validation yields a single value for the phantom quality, that directly depends on the number of validation methods. However, the result of our iterative approach does not have an upper bound. Adding the result of a new method to our current evaluation in Section 10.6 will change the overall validation and increase or decrease the confidence in our phantom. Thereby, the methods analyzed in this chapter provide a good starting point. To the best of our knowledge, this is the first time a phantom is analyzed to this extent.

Establishing an absolute value for 'good' phantoms is complex because it requires de-

velopers to thoroughly evaluate their own phantoms and many experts to review the results of each validation method. Our approach provides a formalized process for phantom validation, which is an important step in this direction. Using our method, different raters should be able to obtain related results. Our approach can be easily extended by new validation methods to a point where the output value has reached a level, at which the phantom is considered to be sufficiently validated. Moreover, the multiplicative combination of all validation methods has the benefit that even a single method with high suitability and low correctness is sufficient to greatly decrease the built up confidence in a phantom.

11. Conclusion Part II

In this last chapter, we reflect the main challenges of phantom validation and our contributions in this field. The thesis is completed with a discussion of future work.

Why is an analysis of the phantom quality required?

To reach clinical acceptance, and to fully understand a method, dedicated evaluation strategies are required. Today, most algorithms are accompanied by some kind of validation study, where the quality of a method is compared with some kind of reference. Therefore, establishing an appropriate gold standard, that is presumed to contain the correct result (the ground truth) or be at least close to it, is an essential task. Phantoms provide an excellent basis, because the ground truth is known for all modeled parameters.

Nevertheless, tools that analyze the phantom quality are largely unknown. In Chapter 8, we reviewed methods that have been proposed for the validation of image processing algorithms. Several methods use a reference database consisting of patient scans or phantoms. However, only few are actually available for research groups other than the initial developers. Our approach starts one step earlier and focuses on the analysis of the underlying reference data.

A characteristic of validation methods in medical image processing – also important for phantom validation – is a formalized and reproducible analysis. Today, the most common phantom validation approach is a visual assessment by a field expert, which provides a quick and informal method. However, expert knowledge is subjective and often not reproducible.

To analyze the quality of a phantom, we need standardized tools that demonstrate the advantages and drawbacks of a phantom for the targeted application. For example, a bad phantom with wrong object intensity values can heavily effect the output of a segmentation algorithm and reduce the relevance of the calculated results. Furthermore, methods that enable a ranking of different phantoms for the same application are required.

What are the benefits of our phantom validation?

Link to Phantom Design. In this work, we proposed a novel phantom validation method that is closely related to the design process (cf. Chap. 9). Therein, models for the relevant parameters of an object are developed and the results are incorporated into a given background. To support this process already early in the design phase, our phantom validation starts with an analysis of the required parameters. We proposed a method based on a multi-criteria decision making technique, namely the Analytic Hierarchy Process. The AHP approach uses pairwise comparisons of parameters, which has demonstrated in studies to be an efficient and accurate approach. It allows for an explicit analysis of the relationships between modules. This way, phantom developers and users are supported in analyzing advantages and drawbacks of a phantom in more detail, which was also found useful during development in this work. For example, the computed importance values were used in our automatic phantom design proposed in Chapter 5 to focus on the relevant parameters.

Iterative Process. Besides the evaluation of relevant parameters for phantom design, an assessment of the actual phantom is required. Validation is an iterative process that requires several steps to reach a reasonable confidence level. In this work, we proposed a novel iterative approach consisting of several evaluation methods. The methods can be analyzed independent of one another, so that a new method does not have to take previous validation results into account. Our overall phantom validation approach is divided into three steps: *method selection*, *method validation*, and *phantom validation*. Each method increases or decreases the confidence in a phantom through an analysis of the method's quality and its relevance for validation.

Phantom Quality Measurement. Our goal is to express the phantom quality by a single value, which enables a straightforward comparison of different validation results among several raters or among different phantoms. After selecting the validation methods, an analysis of their suitability and correctness is performed. We proposed a validation function that combines these two features for each validation method (cf. Eq. 9.1). The last step of our approach then combines all validation methods to a final measurement.

Our validation process ensures that unimportant methods have only limited influence independent from their correctness. Relevant methods on the other hand can heavily increase or decrease the confidence in the phantom. To this end, we propose a multiplicative aggregation of all validation functions. Methods with low suitability result in values close to one, whereas the output of the validation function is larger than one for important and correctly modeled methods.

Applications. The result of our iterative validation approach does not have an upper bound. Instead, each new method adds further information about the phantom and thus changes the validation value. Thereby, the computed result of our approach provides a good starting point for subsequent validation methods. Furthermore, it enables a comparison with other phantoms that carry out the same methods.

Five validation methods have been proposed in Chapter 9. These methods were then

applied in Chapter 10 for the validation of our MS lesion phantoms. All methods increased the confidence in our phantoms. For example, we compared the segmentation overlap computed for phantom data with those from patient data using the same algorithm. Furthermore, we analyzed the effect of object parameter changes. Similar to supporting the initial phantom design phase, the Analytic Hierarchy Process is used to select a parameter with high importance. Our results show that it is worth developing complex parameters, that better reflect the challenges of related patient data.

In Section 10.5, we compared our results with those of a phantom where only one parameter is modified. Our validation helped to detect a major change in phantom quality. A comparison with phantoms that did not use the same validation methods is also possible if the applied methods have the same relevance.

Standardization of Validation. Our long term goal is a standardization of phantom validation. The iterative approach proposed in this work provides a formalization of the validation process, and the applied methods have shown the capabilities for validation. This way, a phantom that is considered to have a good quality by our approach can enhance the clinical acceptance of evaluation studies for new algorithms in medical image analysis.

Future Research

This thesis has given insights into phantom development and validation and has presented several new ideas. Nevertheless, several extensions of our work are conceivable.

Phantom Development

In Part I of this work, our focus was on the development of software phantoms. A future direction could be to put more emphasis on the automatic design of phantoms. An ad hoc solution that develops a phantom from single patient data set should no longer be needed. Large databases are required to do sufficient testing of an algorithm. Our automatic approach allows to parametrically generate not only tens, but hundreds of phantoms. To this end, families of phantoms should be developed categorized into imaging, normal, and pathological variability encountered in clinical practice. Since our approach uses patient data to capture this variability, additional data are required.

Another future work of our phantom design is to focus on further imaging modalities such as CT or PET. This will particularly require the development of new parameter modules such as a new object noise model. Furthermore, currently missing parameter modules could be added for additional flexibility. New applications within new areas of medical image analysis is another interesting topic. An example is the evaluation of registration methods using phantoms.

Our current phantom design approach covers hybrid software phantoms, which limits the possible objects and thus the clinical applications. An extension of our approach to the automatic development of data sets where the modeled object is an organ, e.g., the heart or the liver, is of great interest. A reasonable software phantom that allows for modeling of parameters within a range extracted from patient data is currently not available.

On-site competitions of several research teams during a conference have become popular in the last years, e.g., the Grand Challenge workshops at the MICCAI conference. An interesting extension of the current patient data used for training and testing are phantoms. For example, a segmentation challenge using our data sets could add new insights into both the tested algorithms as well as the underlying phantom data.

Phantom Validation

An important aspect of our current validation approach is the user involvement. We rely on rater decisions, which includes some inherent and inevitable subjectiveness. A common method to overcome this issue is a consensus meeting of several experts. For example, Sloane et al. (2003) used seven development stages with different participants to assess the criteria required to purchase specific clinical equipment within an AHP-based approach. Furthermore, the competence of each expert is important. Tsyganok et al. (2011) conclude that the individual expertise of a rater can be neglected for large groups of more than 50 persons. On the other hand, for relatively small groups, expert competence should be taken

into consideration. A different approach could be to remove the manual rating from the validation process. For example, a distance function measuring the similarity between phantom data and a database with a sufficient amount of reference data could be used to compute the correctness of a parameter model.

Besides the already proposed features, several additional parameters could be considered for validation such as the required time and costs to develop a phantom. Another related factor is the amount of manual processing required to assemble a phantom. Although such parameters can affect the development process, they should not change the overall phantom quality and are therefore not considered in this work.

Future work will extend our approach by new validation methods to a level where the confidence in our phantom is determined as sufficient. To demonstrate the wide applicability of our validation, we also plan to investigate phantoms from other groups. Thereby, data availability is an important topic for both phantom development and validation. Today, only few data sets are available or are widely accepted in medical image analysis. A step towards this goal are public data sets within a database plus an easy-to-use web interface. This would allow other researchers to upload their phantoms as well as their validation results for an already available phantom. Additionally, it would allow selecting a suitable phantom for evaluation studies. Thereby, our validation approach can provide a common basis for comparisons. Furthermore, determining the required modules and their suitability provides an excellent platform for discussions.

In Chapter 4, we proposed a formalization of the phantom design process and developed a template that can be used to provide an easy-to-use phantom description. A similar approach could also be beneficial for phantom validation. In this case, the description contains not only the output of the overall validation. Also, all validation steps are included with a short description and the values and meanings for suitability and correctness.

Bibliography

- American Institute of Aeronautics and Astronautics (1998). *Guide for the verification and validation of computational fluid dynamics simulations*. AIAA G-077.
- Armato, S. G., G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, A. P. Reeves, B. Y. Croft, and L. P. Clarke (2004). Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community. *Radiology* 232, 739–748.
- Armato, S. G., C. R. Meyer, M. F. McNitt-Gray, G. McLennan, A. P. Reeves, B. Y. Croft, and L. P. Clarke (2008). The Reference Image Database to Evaluate Response to Therapy in Lung Cancer (RIDER) Project: A Resource for the Development of Change-Analysis Software. *Clinical Pharmacology & Therapeutics* 84(4), 448–456.
- Ashton, E., C. Takahashi, M. Berg, A. Goodman, S. Totterman, and S. Ekholm (2003). Accuracy and Reproducibility of Manual and Semiautomated Quantification of MS Lesions by MRI. *Journal of Magnetic Resonance Imaging* 17, 300–308.
- Aubert-Broche, B., A. C. Evans, and D. L. Collins (2006). A new improved version of the realistic digital brain phantom. *NeuroImage* 32(1), 138–145.
- Aubert-Broche, B., M. Griffin, G. B. Pike, and A. C. E. D. L. Collins (2006). Twenty New Digital Brain Phantoms for Creation of Validation Image Data Bases. *IEEE Trans Medical Imaging* 25(11), 1410–1416.
- Bajcsy, R. and S. Kovacic (1989). Multiresolution Elastic Matching. *Computer Vision, Graphics and Image Processing* 46, 1–21.
- Ballester, M. A. G., A. P. Zisserman, and M. Brady (2002). Estimation of the partial volume effect in MRI. *Medical Image Analysis* 6(4), 389–405.
- Baumgart, B. G. (1975). Winged-Edge Polyhedron Representation for Computer Vision. In *National Computer Conference*, Volume 44, pp. 589–596.
- Bornemann, L., V. Dicken, J.-M. Kuhnigk, D. Wormanns, H.-O. Shin, H.-C. Bauknecht, V. Diehl, M. Fabel, S. Meier, O. Kress, S. Krass, and H.-O. Peitgen (2007). OncoTREAT: a software assistant for cancer therapy monitoring. *International Journal of Computer Assisted Radiology and Surgery* 1(5), 231–242.

- Bouix, S., M. Martin-Fernandez, L. Ungar, M. Nakamura, M.-S. Koo, R. W. McCarley, and M. E. Shenton (2007). On evaluating brain tissue classifiers without a ground truth. *Neuroimage* 36, 1207–1224.
- Brix, G., M. L. Bahner, U. Hoffmann, A. Horvath, and W. Schreiber (1999). Regional Blood Flow, Capillary Permeability, and Compartmental Volumes: Measurement with Dynamic CT - Initial Experience. *Radiology* 210, 269–276.
- Brown, A. and K. Wallnau (1998). The current state of cbse. *IEEE Software* 15(5), 37–46.
- Broy, M., A. Deimel, J. Henn, K. Koskimies, F. Plasil, G. Pomberger, W. Pree, M. Stal, and C. Szyperki (1998). What characterizes a (software) component? *Software - Concepts and Tools* 19(1), 49–56.
- Burgess, A., F. Jacobson, and P. Judy (2003). Mass discrimination in mammography: experiments using hybrid images. *Academic Radiology* 10(11), 1247–1256.
- Buvat, I., V. Chameroy, F. Aubry, M. Péligrini, G. E. Fakhri, C. Huguenin, H. Benali, A. Todd-Pokropek, and R. D. Paola (1999). The need to develop guidelines for evaluations of medical image processing procedures. In *SPIE Medical Imaging*, Volume 3661, pp. 1466–1477.
- Caon, M. (2004). Voxel-based computational models of real human anatomy: a review. *Radiation and Environmental Biophysics* 42(4), 229–235.
- Cardenesa, R., R. de Luis-Garcia, and M. Bach-Cuadra (2009). A multidimensional segmentation evaluation for medical image data. *Computer Methods and Programs in Biomedicine* 96(2), 108–124.
- Chen, C. C., Y. L. Wan, Y. Y. Wai, and H.-L. Liu (2004). Quality Assurance of Clinical MRI Scanners Using ACR MRI Phantom: Preliminary Results. *Journal of Digital Imaging* 17(4), 279–284.
- Christensen, G. E., R. D. Rabbitt, and M. I. Miller (1996). Deformable templates using large deformation kinematics. *IEEE Trans Image Processing* 5(10), 1435–1447.
- Cinti, M. N., R. Pani, F. Garibaldi, R. Pellegrini, M. Betti, N. Lanconelli, A. Riccardi, R. Campanini, G. Zavattini, G. D. Domenico, A. D. Guerra, N. Belcari, W. Bencivelli, A. Motta, A. Vaiano, and I. N. Weinberg (2004). Custom Breast Phantom for an Accurate Tumor SNR Analysis. *IEEE Trans Nuclear Science* 51(1), 198 – 204.
- Clark, M. C., L. O. Hall, D. B. Goldgof, R. Velthuizen, F. R. Murtagh, and M. S. Silbiger (1998). Automatic Tumor Segmentation Using Knowledge-Based Techniques. *IEEE Trans Medical Imaging* 17(2), 187–201.
- Clatz, O., M. Sermesant, P.-Y. Bondiau, H. Delingette, S. K. Warfield, G. Malandain, and N. Ayache (2005). Realistic Simulation of the 3D Growth of Brain Tumors in MR

- Images Coupling Diffusion with Mass Effect. *IEEE Trans Medical Imaging* 24(10), 1334–1346.
- Collins, D., A. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans (1998). Design and Construction of a Realistic Digital Brain Phantom. *IEEE Trans Medical Imaging* 17(5), 463–468.
- Compston, A. and A. Coles (2002). Multiple sclerosis. *Lancet* 359(9313), 1221–31.
- Cristy, M. (1980). Mathematical phantoms representing children of various ages for use in estimates of internal dose. Technical report, Oak Ridge National Laboratory, Oak Ridge TN, USA.
- Crum, W. R., O. Camara, and D. L. G. Hill (2006). Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis. *IEEE Trans Medical Imaging* 25(11), 1451–1461.
- CTSim 5.1.2, h. a. <http://ctsim.org>.
- Cuadra, M. B., L. Cammoun, T. Butz, O. Cuisenaire, and J.-P. Thiran (2005). Comparison and Validation of Tissue Modelization and Statistical Classification Methods in T1-Weighted MR Brain Images. *IEEE Trans Medical Imaging* 24(12), 1548–1565.
- Delingette, H. (1998). Toward realistic soft-tissue modeling in medical simulation. *IEEE Special Issue on Virtual and Augmented Reality in Medicine* 86(3), 512–523.
- Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology* 26(3), 297–302.
- Drexl, J., V. Knappe, H. K. Hahn, K. Lehmann, B. B. Frericks, H. Shin, and H.-O. Peitgen (2004). Accuracy analysis of vessel segmentation for a LITT dosimetry planning system. In *Proceedings of the Scientific Workshop on Medical Robotics, Navigation and Visualization*, pp. 204–212. World Scientific.
- EMS. Homepage at: <https://mirc.uzleuven.be/MedicalImageComputing/downloads/ems.php>. Accessed March, 2014.
- Ens, K., F. Wenzel, S. Young, J. Modersitzki, and B. Fischer (2009). Design of a synthetic database for the validation of nonlinear registration and segmentation of magnetic resonance brain images. In *SPIE Medical Imaging*, Volume 7259, pp. 72593–3–725933–9.
- FDA. General Principles of Software Validation; Final Guidance for Industry and FDA Staff. Homepage at: <http://www.fda.gov>. Accessed October, 2009.
- Ferrant, M., A. Nabavi, B. Macq, J. Jolesz, R. Kikinis, and S. K. Warfield (2001). Registration of 3-d intraoperative MR images of the brain using a finite-element biomechanical model. *IEEE Trans Medical Imaging* 20(12), 1384–1397.
- Filippi, M., M. Gawne-Cain, C. Gasperini, J. vanWaesberghe, J. Grimaud, F. Barkhof, M. Sormani, and D. Miller (1998). Effect of training and different measurement

- strategies on the reproducibility of brain MRI lesion load measurements in multiple sclerosis. *Neurology* 50(1), 238–244.
- Firbank, M. J., R. M. Harrison, E. D. Williams, and A. Coulthard (2000, Apr). Quality assurance for MRI: practical experience. *Br J Radiol* 73(868), 376–383.
- Fisher, H. L. J. and W. S. Snyder (1966). Variation of Dose Delivered by ¹³⁷Cs as a Function of Body Size from Infancy to Adulthood. Technical Report ORNL-4007, Oak Ridge National Laboratory.
- Fitzpatrick, M. J., G. Starkschall, P. Balter, J. A. Antolak, T. Guerrero, C. Nelson, P. Keall, and R. Mohan (2005). A novel platform simulating irregular motion to enhance assessment of respiration-correlated radiation therapy procedures. *Journal of Applied Clinical Medical Physics* 6(1), 13–21.
- Fryback, D. G. and J. R. Thornbury (1991). The efficacy of diagnostic imaging. *Medical Decision Making* 11, 88–94.
- Fu, L., V. Fonov, B. Pike, A. C. Evans, and D. L. Collins (2006). Automated Analysis of Multi Site MRI Phantom Data for the NIHPD Project. In *Proc. MICCAI*, Number 4191 in LNCS, pp. 144–151.
- García-Lorenzo, D., S. Prima, D. Arnold, D. Collins, and C. Barillot (2011). Trimmed-likelihood estimation for focal lesions and tissue segmentation in multisequence MRI for multiple sclerosis. *IEEE Trans. Med. Imag.* 30(8), 1455–1467.
- Garrity, J. M., W. P. Segars, S. B. Knisley, and B. M. W. Tsui (2003). Development of a dynamic model for the lung lobes and airway tree in the NCAT phantom. *IEEE Trans Nuclear Science* 50(3), 378–383.
- Gedamu, E. L., D. L. Collins, and D. L. Arnold (2008). Automated Quality Control of Brain MR Images. *Journal of Magnetic Resonance Imaging* 28, 308–319.
- Gee, J. (2000). Performance evaluation of medical image processing algorithms. In *SPIE Medical Imaging*, Volume 3979, pp. 19–27.
- Gerig, G., M. Jomier, and M. Chakos (2001). VALMET: A new validation tool for assessing and improving 3D object segmentation. In *Proc. MICCAI*, Volume 2208 of LNCS, pp. 516–523.
- Hagemann, A., K. Rohr, H. S. Stiehl, U. Spetzger, and J. M. Gilsbach (1999). Biomechanical modeling of the human head for physically based, nonrigid image registration. *IEEE Trans Medical Imaging* 18(10), 875–884.
- Hahn, H. K., B. Jolly, M. Lee, D. Krastel, J. Rexilius, J. Drexler, M. Schlüter, B. Terwey, and H.-O. Peitgen (2004). How Accurate is Brain Volumetry? A Methodological Evaluation. In *MICCAI*, Number 3216 in LNCS, pp. 335–342. Springer.
- Hahn, H. K. and H.-O. Peitgen (2000). The Skull Stripping Problem in MRI Solved by a Single 3D Watershed Transform. In *Proc. MICCAI*, Number 1935 in LNCS, pp.

134–143. Springer.

- Han, X., J. Jovicich, D. Salat, A. van der Kouwe, B. Quinn, S. Czanner, E. Busa, J. Pacheco, M. Albert, R. Killiany, P. Maguire, D. Rosas, N. Makris, A. Dale, B. Dickerson, and B. Fischl (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage* 32, 180–194.
- Hanahan, D. and R. Weinberg (2000). The Hallmarks of Cancer. *Cell* 100(1), 57–70.
- Heath, M., K. W. Bowyer, D. Kopans, and et al (1998). Current status of the Digital Database for Screening Mammography. In *Proc. of Digital Mammography*, pp. 457–460.
- Heimann, T., B. van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. M. M. Cashman, Y. Chi, A. Córdova, B. M. Dawant, M. Fidrich, J. D. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmüller, R. I. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.-P. Meinzer, G. Németh, D. S. Raicu, A.-M. Rau, E. M. van Rikxoort, M. Rousson, L. Ruskó, K. A. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. M. Waite, A. Wimmer, and I. Wolf (2009). Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets. *IEEE Trans Medical Imaging* 28(8), 1251–1265.
- Henneman, W., J. Sluimer, J. Barnes, W. van der Flier, I. Sluimer, N. Fox, P. Scheltens, H. Vrenken, and F. Barkhof (2009). Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. *Neurology* 72(11), 999–1007.
- Hoehne, K. H., B. Pflesser, A. Pommert, K. Priesmeyer, M. Riemer, T. Schiemann, R. Schubert, U. Tiede, H. Frederking, S. Gehrman, S. Noster, and U. Schumacher (2003). *VOXEL-MAN 3D Navigator: Inner Organs. Regional, Systemic and Radiological Anatomy. Innere Organe. Topographische, Systematische und Radiologische Anatomie*. Springer-Verlag Electronic Media, Heidelberg.
- Hoffman, E. J., P. D. Cutler, W. M. Digby, and J. C. Mazziotta (1990). 3-D phantom to simulate cerebral blood flow and metabolic images for PET. *IEEE Trans Nuclear Science* 37(2), 616–620.
- Huber, M. E., M. Stuber, R. M. Botnar, P. Boesiger, and W. J. Manning (2000). Low-cost MR-compatible moving heart phantom. *Journal of Cardiovascular Magnetic Resonance* 2(3), 181–187.
- Inglese, M. (2006). Multiple sclerosis: New insights and trends. *American Journal of Neuroradiology* 27(5), 954–7.
- Jaccard, P. (1912). The distribution of flora in the alpine zone. *The New Phytologist* 11(2), 37–50.

- Jannin, P., C. Grova, and C. R. Maurer (2006). Model for defining and reporting reference-based validation protocols in medical image processing. *International Journal of Computer Assisted Radiology and Surgery* 1(2), 63–73.
- Joe, B., M. Fukui, C. Meltzer, Q. Huang, R. Day, P. Greer, and M. Bozi (1999). Brain Tumor Volume Measurement: Comparison of Manual and Semiautomated Methods. *Radiology* 212, 811–816.
- Kansal, A. R., S. Torquato, G. R. Harsh, E. A. Chiocca, and T. S. Deisboeck (2000). Simulated Brain Tumor Growth Dynamics Using a Three-Dimensional Cellular Automaton. *J Theor Biol* 203(4), 367–382.
- Karssemeijer, N. (1993). Adaptive noise equalization and recognition of microcalcification clusters in mammograms. *Int. Journal of Pattern Recognition and Image Analysis* 7(6), 1357–1377.
- Kauffmann, C., P. Gravel, B. Godbout, A. Gravel, G. Beaudoin, J.-P. Raynauld, J. Martel-Pelletier, J. P. Pelletier, and J. A. de Guise (2003). Computer-Aided Method for Quantification of Cartilage Thickness and Volume Changes Using MRI: Validation Study Using a Synthetic Model. *IEEE Trans Biomedical Engineering* 50(8), 978–988.
- Kaus, M., S. K. Warfield, A. Nabavi, P. M. Black, F. A. Jolesz, and R. Kikinis (2001). Automated Segmentation of MR Images of Brain Tumors. *Radiology* 218(2), 586–591.
- Kaye, J. M., F. P. Primiano, and D. N. Metaxas (1998). A three-dimensional virtual environment for modeling mechanical cardiopulmonary interactions. *Medical Image Analysis* 2(2), 169–195.
- Kazemi, K., H. Moghaddam, R. Grebe, C. Gondry-Jouet, and F. Wallois (2011). Design and construction of a brain phantom to simulate neonatal MR images. *Computerized Medical Imaging and Graphics* 35(3), 237–250.
- Keil, S., C. Plumhans, F. Behrendt, S. Stanzel, M. Suehling, G. Mühlenbruch, A. Mahnken, R. Günther, and M. Das (2009). Automated measurement of lymph nodes: a phantom study. *European Radiology* 19(5), 1079–1086.
- Kiebel, S. J., J. Ashburner, J.-B. Poline, and K. J. Friston (1997). MRI and PET Coregistration—A Cross Validation of Statistical Parametric Mapping and Automated Image Registration. *Neuroimage* 5, 271–279.
- Klein, S., M. Staring, K. Murphy, M. Viergever, and J. Pluim (2010). elastix: A Toolbox for Intensity Based Medical Image Registration. *IEEE Trans. Med. Imag.* 29(1), 196–205.
- Klink, F., T. Hoffmann, A. Boese, M. Skalej, and K.-H. Grote (2014). Additive Manufacturing of Anatomical Phantoms Based on Medical Imaging Data Sets. In *Proceedings of the 3rd International Conference on Design Engineering and Science (ICDES 2014)*, Volume 2, pp. 129–133.

- Ko, J. P., H. Rusinek, E. L. Jacobs, J. S. Babb, M. Betke, G. McGuinness, and D. P. Naidich (2003). Small pulmonary nodules: volume measurement at chest CT-phantom study. *Radiology* 228(3), 864–870.
- Koller, C. J., J. P. Eatough, P. J. Mountford, and G. Frain (2006). A survey of MRI quality assurance programmes. *British Journal of Radiology*.
- Kuhnigk, J. M., V. Dicken, L. Bornemann, A. Bakai, D. Wormanns, S. Krass, and H.-O. Peitgen (2006). Morphological Segmentation and Partial Volume Analysis for Volumetry of Solid Pulmonary Lesions in Thoracic CT Scans. *IEEE Trans Medical Imaging* 25(4), 417–434.
- Kurtzke, J. (1983). Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (edss). *Neurology* 33, 1444–1452.
- Kwan, R. K. S., A. C. Evans, and G. B. Pike (1999). MRI Simulation-Based Evaluation of Image-Processing and Classification Methods. *IEEE Trans Medical Imaging* 18(11), 1085–1097.
- Kyriacou, S. K., C. Davatzikos, S. J. Zinreich, and R. N. Bryan (1999). Nonlinear elastic registration of brain images with tumor pathology using a biomechanical model. *IEEE Trans Medical Imaging* 18(7), 580 – 592.
- Lassmann, H. (2002). Mechanisms of demyelination and tissue destruction in multiple sclerosis. *Clin Neurol Neurosurg* 104(3), 168–71.
- Lazebnik, R. S., B. D. Weinberg, M. S. Breen, J. S. Lewin, and D. L. Wilson (2002). Three-dimensional model of lesion geometry for evaluation of MR-guided thermal ablation therapy. *Academic Radiology* 9(10), 1128–1138.
- Lee, C. and J. K. Lee (2006). Computational Anthropomorphic Phantoms for Radiation Protection Dosimetry: Evolution and Prospects. *Nuclear Engineering and Technology* 39(3), 239–250.
- Lee, S. W., S. H. Park, K. W. Kim, E. K. Choi, Y. M. Shin, P. N. Kim, K. H. Lee, E. S. Yu, S. Hwang, and S.-G. Lee (2007). Unenhanced ct for assessment of macrovesicular hepatic steatosis in living liver donors: Comparison of visual grading with liver attenuation index. *Radiology* 244, 479–485.
- Lehmann, T. M. (2002). From plastic to gold - A unified classification scheme for reference standards in medical image processing. In *Proc. SPIE Medical Imaging*, Volume 4684, pp. 1819–1827.
- Liberatore, M. J. and R. L. Nydick (2008). The analytic hierarchy process in medical and health care decision making: A literature review. *European Journal of Operational Research* 189, 194–207.
- Liney, G. P., M. Sreenivas, P. Gibbs, R. Garcia-Alvarezand, and L. W. Turnbull (2006).

- Breast Lesion Analysis of Shape Technique: Semiautomated vs. Manual Morphological Description. *J Magn Reson Imaging* 23(4), 493–498.
- Liney, G. P., D. J. Tozer, and L. W. Turnbull (1999). A simple and realistic tissue-equivalent breast phantom for MRI. *J Magn Reson Imaging* 10(6), 968–971.
- Luft, A. R., M. Skalej, D. Welte, R. Kolb, and U. Klose (1996). Reliability and exactness of MRI-based volumetry: a phantom study. *J Magn Reson Imaging* 6(4), 700–704.
- Malloy, P., S. Correia, G. Stebbins, and D. H. Laidlaw (2007). Neuroimaging of White Matter in Aging and Dementia. *The Clinical Neuropsychologist* 21(1), 73–109.
- Mattila, S., V. Renvall, J. Hiltunen, D. Kirven, R. Sepponen, R. Hari, and A. Tarkiainen (2007). Phantom-based evaluation of geometric distortions in functional magnetic resonance and diffusion tensor imaging. *Magn Reson Med* 57(4), 754–763.
- McClelland, J. R., J. M. Blackall, S. Tarte, A. C. Chandler, S. Hughes, S. Ahmad, D. B. Landau, and D. J. Hawkes (2006). A continuous 4D motion model from multiple respiratory cycles for use in lung radiotherapy. *Medical Physics* 33(9), 3348–3358.
- McDonald, W., A. Compston, G. Edan, D. Goodkin, H. Hartung, F. Lublin, H. McFarland, D. Paty, C. Polman, S. Reingold, M. Sandberg-Wollheim, W. Sibley, A. Thompson, S. van den Noort, B. Weinshenker, and J. Wolinsky (2001). Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of Neurology* 50(1), 121–127.
- McNitt-Gray, M. F., S. G. Armato, C. R. Meyer, A. P. Reeves, G. McLennan, R. Pais, J. Freymann, M. S. Brown, R. M. Engelmann, P. H. Bland, G. E. Laderach, C. Piker, J. Guo, D. P. Qing, D. F. Yankelevitz, D. R. Aberle, E. J. R. van Beek, H. MacMahon, E. A. Kazerooni, B. Y. Croft, and L. P. Clarke (2007). The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation. In *Proceedings of the SPIE*, Volume 6514, pp. 65140K1–65140K8.
- McRobbie, D. and R. Quest (2002). Effectiveness and relevance of MR acceptance testing: results of an 8 year audit. *The British Journal of Radiology* 75, 523–531.
- McRobbie, D. W., E. A. Moore, and M. J. Graves (2003). *MRI from Picture to Proton*. Cambridge University Press.
- Melhem, E. R., E. H. Herskovits, K. Karli-Oguz, X. Golay, D. A. Hammoud, B. J. Fortman, F. M. Munter, and R. Itoh (2003). Defining Thresholds for Changes in Size of Simulated T2-Hyperintense Brain Lesions on the Basis of Qualitative Comparisons. *American Journal of Roentgenology* 180, 65–69.
- MeVisLab 1.5. Homepage at: <http://www.mevislab.de>. Accessed January, 2008.
- Miller, D., F. Barkhof, J. Frank, G. Parker, and A. Thompson (1998). Measurement of atrophy in multiple sclerosis: pathological basis, methodological aspects and clinical relevance. *Brain* 125(8), 1676–1695.

- Miller, D., R. Grossman, S. Reingold, and H. McFarland (1998). The role of magnetic resonance techniques in understanding and managing multiple sclerosis. *Brain* 121(1), 3–24.
- Modersitzki, J. (2004). *Numerical Methods for Image Registration*. Oxford University Press Series: Numerical Mathematics and Scientific Computation.
- Molyneux, P., P. Tofts, A. Fletcher, B. Gunn, P. Robinson, H. Gallagher, I. Moseley, G. Barker, and D. Miller (1998). Precision and reliability for measurement of change in mri lesion volume in multiple sclerosis: a comparison of two computer assisted techniques. *J Neurol Neurosurg Psychiatry* 65, 42–47.
- Moore, J., M. Dramgova, M. Wierzbicki, J. Barron, and T. Peters (2003). A High Resolution Dynamic Heart Model Based on Averaged MRI Data. In R. Ellis and T. Peters (Eds.), *Proc. MICCAI*, Number 2798 in LNCS, pp. 549–555.
- Moretti, B., J. M. Fadil, S. Ruan, D. Bloyet, and B. Mazoyer (2000). Phantom-based performance evaluation: Application to brain segmentation from magnetic resonance images. *Medical Image Analysis* 4(4), 303–316.
- Mortazavi, D., A. Z. K. AZ, and H. Soltanian-Zadeh (2012). Segmentation of multiple sclerosis lesions in mr images: a review. *Neuroradiology* 54(4), 299–320.
- Mueller, S. G., M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett (2005). The Alzheimers Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America* 15(4), 869–877.
- Neumuth, T., P. Jannin, G. Strauss, J. Meixensberger, and O. Burgert (2009). Validation of knowledge acquisition for surgical process models. *Journal of the American Medical Informatics Association* 16, 72–80.
- Noe, A. and J. C. Gee (2001). Partial Volume Segmentation of Cerebral MRI Scans with Mixture Model Clustering. In R. L. M.F. Insana (Ed.), *Information Processing in Medical Imaging: 17th International Conference, IPMI 2001, Davis, CA, USA, June 18-22, 2001, Proceedings*, Volume 2082, pp. 423–430.
- Nomori, H., K. Watanabe, T. Ohtsuka, T. Naruke, K. Suemasu, and K. Uno (2005). Visual and semiquantitative analyses for f-18 fluorodeoxyglucose pet scanning in pulmonary nodules 1 cm to 3 cm in size. *The Annals of Thoracic Surgery* 79, 984–988.
- Osborn, A. G. and K. A. Tong (1999). *Handbook of Neuroradiology: Brain and Skull, 2nd edition*. Mosby-Year Book.
- Perlin, K. (2002). "improving noise". In *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2002)*, pp. 681–682.
- Petersik, A., B. Pflesser, U. Tiede, K. H. Höhne, and R. Leuwer (2003). Realistic Haptic Interaction in Volume Sculpting for Surgery Simulation. In *Surgery Simulation and*

- Soft Tissue Modeling, Proc. IS4TM 2003*, Number 2673 in LNCS, pp. 194–202.
- Pichon, E., A. Tannenbaum, and R. Kikinis (2004). A statistically based flow for image segmentation. *Medical Image Analysis* 8(3), 267–274.
- Pikus, L., J. H. Woo, R. L. Wolf, E. H. Herskovits, G. Moonis, A. F. Jawad, J. Krzajka, and E. R. Melhem (2006). Artificial Multiple Sclerosis Lesions on Simulated FLAIR Brain MR Images: Echo Time and Observer Performance in Detection. *Radiology* 239(1), 238–245.
- Plewes, D. and P. Dean (1981). The influence of partial volume averaging on sphere detectability in computed tomography. *Physics in Medicine and Biology* 26(5), 913–919.
- Polman, C., S. Reingold, G. Edan, M. Filippi, H. Hartung, L. Kappos, F. Lublin, L. Metz, H. McFarland, P. O’Connor, M. Sandberg-Wollheim, A. Thompson, B. Weinshenker, and J. Wolinsky (2005). Diagnostic criteria for multiple sclerosis: 2005 revisions to the mcdonald criteria. *Annals of neurology* 58(5), 840–6.
- Prastawa, M., E. Bullitt, and G. Gerig (2005). Synthetic Ground Truth for Validation of Brain Tumor MRI Segmentation. In *Proc. MICCAI*, Number 3749 in LNCS, pp. 26–33.
- Prastawa, M., E. Bullitt, and G. Gerig (2009). Simulation of Brain Tumors in MR Images for Evaluation of Segmentation Efficacy. *Medical Image* 13(2), 297–311.
- Prastawa, M., N. Moon, E. Bullitt, K. van Leemput, and G. Gerig (2003, Dec.). Automatic Brain and Tumor Segmentation. *Academic Radiology* 10, 1341–1348.
- Price, R. R., L. Axel, T. Morgan, R. Newman, W. Perman, N. Schneiders, M. Selikson, M. Wood, and S. R. Thomas (1990). Quality assurance methods and phantoms for magnetic resonance imaging: Report of AAPM nuclear magnetic resonance Task Group No. 1. *Medical Physics* 17(2), 287–295.
- Pupi, A., M. T. D. Cristofaro, A. R. Formiconi, A. Passeri, A. Speranzi, E. Giraud, and U. Meldolesi (1990). A brain phantom for studying contrast recovery in emission computerized tomography. *European Journal of Nuclear Medicine and Molecular Imaging* 17(1-2), 15–20.
- Reinertsen, I. and D. L. Collins (2006). A realistic phantom for brain-shift simulations. *Medical Physics* 33(9), 3234–3240.
- Rexilius, J., H. K. Hahn, H. Bourquain, and H.-O. Peitgen (2003). Ground Truth in MS Lesion Volumetry - A Phantom Study. In *Proc. MICCAI*, Volume LNCS 2879, pp. 546–553. Springer.
- Rexilius, J., H. K. Hahn, J. Klein, M. Lentschig, and H.-O. Peitgen (2007). Multispectral Brain Tumor Segmentation based on Histogram Model Adaptation. In *Proc. SPIE Medical Imaging*, Volume 6514, pp. 65140V–1–65140V–10.

- Rexilius, J., H. K. Hahn, M. Schlueter, H. Bourquain, and H.-O. Peitgen (2005). Evaluation of accuracy in MS lesion volumetry using realistic lesion phantoms. *Academic Radiology* 12, 17–24.
- Rexilius, J., H. K. Hahn, M. Schlueter, S. Kohle, H. Bourquain, J. Boettcher, and H.-O. Peitgen (2004). A Framework for the Generation of Realistic Brain Tumor Phantoms and Applications. In C. Barillot, D. Haynor, and P. Hellier (Eds.), *Proc. MICCAI*, Number 3217 in LNCS, pp. 243–250. Springer.
- Rexilius, J., H. Handels, A. Navabi, R. Kikinis, and S. K. Warfield (2002). Automatic Nonrigid Registration for Tracking Brain Shift during Neurosurgery. In *Workshop Bildverarbeitung für die Medizin*, pp. 135–138. Springer.
- Rexilius, J., O. Konrad, and H.-O. Peitgen (2008). A software assistant for the design of realistic software phantoms. In *Proc. SPIE Medical Imaging*, Volume 6914, pp. 69144Y–1–69144Y–10.
- Rexilius, J. and K. Tönnies (2014a). Automatic design and validation of software phantoms for Multiple Sclerosis. *International Journal of Computer Assisted Radiology and Surgery*. submitted.
- Rexilius, J. and K. Tönnies (2014b). Automatic design of realistic Multiple Sclerosis lesion phantoms. In *Workshop Bildverarbeitung für die Medizin*, pp. 270–275.
- Rexilius, J., S. K. Warfield, C. R. G. Guttmann, X. Wei, R. Benson, L. Wolfson, M. Shenton, H. Handels, and R. Kikinis (2001). A Novel Nonrigid Registration Algorithm and Applications. In *Proc. MICCAI*, Number 2208 in LNCS, pp. 923–931. Springer.
- Reynolds, H. M., P. R. D. R. F. Uren, J. F. Thompson, and N. P. Smith (2007). Mapping Melanoma Lymphoscintigraphy Data onto a 3D Anatomically Based Model. *Annals of Biomedical Engineering*.
- Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology* 15, 234–281.
- Saaty, T. L. (1990). How to make a decision: The Analytic Hierarchy Process. *European Journal of Operational Research* 48, 9–26.
- Schlüter, M., O. Konrad-Verse, H. K. Hahn, B. Stieltjes, J. Rexilius, and H.-O. Peitgen (2005). White Matter Lesion Phantom for Diffusion Tensor Data and Its Application to the Assessment of Fiber Tracking. In *Proc. SPIE Medical Imaging*, Volume 5746 of *Medical Imaging: Image Processing*, pp. 835–844. SPIE.
- Segars, W. P., D. S. Lalush, and B. M. W. Tsui (1999). A realistic spline-based dynamic heart phantom. *IEEE Trans Nuclear Science* 46(3), 503–506.
- Segars, W. P., D. S. Lalush, and B. M. W. Tsui (2001). Modeling respiratory mechanics in the MCAT and spline-based MCAT phantoms. *IEEE Trans Nuclear Science* 48(1), 89–97.

- Sekhar, A., M. Sun, and B. Siewert (2014). A tissue phantom model for training residents in ultrasound-guided liver biopsy. *Academic Radiology* 21(7), 902–908.
- Shattuck, D. W., S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy (2001). Magnetic resonance image tissue classification using a partial volume model. *NeuroImage* 13(5), 856–876.
- Shin, H., M. Blietz, B. Frericks, S. Baus, D. Savellano, and M. Galanski (2006). Insertion of virtual pulmonary nodules in CT data of the chest: development of a software tool. *European Radiology* 16(11), 1432–1084.
- Sierra, R., M. Bajka, and G. Szekely (2006). Tumor growth models to generate pathologies for surgical training simulators. *Medical Image Analysis* 10(3), 305–316.
- Sijbers, J. (1998). *Signal and Noise Estimation from Magnetic Resonance Images*. Ph. D. thesis, Vision Lab, University of Antwerpe.
- Sled, J. G., A. P. Zijdenbos, and A. C. Evans (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Medical Imaging* 17(1), 87–97.
- Sloane, E. B., M. J. Liberatore, R. L. Nydick, W. Luo, and Q. Chung (2003). Using the analytic hierarchy process as a clinical engineering tool to facilitate an iterative, multidisciplinary, microeconomic health technology assessment. *Computers and Operations Research* 30, 1447–1465.
- Smith, S. (2002). Fast robust automated brain extraction. *Hum Brain Mapping*. *Hum Brain Mapping* 17(3), 143–155.
- Soille, P. (2003). *Morphological Image Analysis: Principles and Applications*. Springer.
- Sonka, M., V. Hlavac, and R. Boyle (1998). *Image Processing, Analysis and Machine Vision.: Analysis and Machine Vision*. Itps Thomson Learning, 2nd edition.
- Spitzer, V. M., M. J. Ackerman, A. L. Scherzinger, and D. Whitlock (1996). The visible human male: a technical report. *Journal of the American Medical Informatics Association* 3(2), 118–130.
- Spitzer, V. M. and D. G. Whitlock (1998). The visible human dataset: The anatomical platform for human simulation. *Anat. Rec. (New Anat.)* 253(2), 49 – 57.
- Styner, M., J. Lee, B. Chin, and et als (2008). 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. In *MICCAI Workshop*, pp. 1–6.
- Suryanarayanan, S., A. Karellas, S. Vedantham, S. M. Waldrop, and C. J. D’Orsi (2005). Detection of Simulated Lesions on Data-compressed Digital Mammograms. *Radiology* 236, 31–36.
- Swanson, K. R., C. Bridge, J. D. Murray, and E. C. Alvord (2003). Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. *Journal of the neurological sciences* 216(1), 1–10.

- Thacker, N. A., A. F. Clark, J. L. Barron, J. R. Beveridge, P. Courtney, W. R. Crum, V. Ramesh, and C. Clark (2008). Performance characterization in computer vision: A guide to best practices. *Computer Vision and Image Understanding* 109(3), 305–334.
- Therasse, P., S. G. Arbutk, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, J. Verweij, M. V. Glabbeke, A. T. van Oosterom, M. C. Christian, and S. G. Gwyther (2000). New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *Journal of the National Cancer Institute* 92(3), 205–216.
- Thornton, A. F., H. M. Sandler, R. K. T. Haken, D. L. McShan, B. A. Fraass, M. L. L. Vigne, and B. R. Yanke (1992). The clinical utility of magnetic resonance imaging in 3-dimensional treatment planning of brain neoplasms. *Int J Radiat Oncol Biol Phys* 24(4), 767–775.
- Timinger, H., S. Krueger, and J. Borgert (2006). MR Compatible Heart Phantom for Medical Research. In *ISMRM*, Volume 14, pp. 1644.
- Tofts, P. S., G. J. Barker, M. Filippi, M. Gawne-Cain, and M. Lai (1997). An oblique cylinder contrast-adjusted (OCCA) phantom to measure the accuracy of MRI brain lesion volume estimation schemes in multiple sclerosis. *Magn Reson Imaging* 15, 183–192.
- Tofts, P. S. and A. G. Kermode (1991). Measurement of the Blood-Brain Barrier Permeability and Leakage Space Using Dynamic MR Imaging. 1. Fundamental Concepts. *Magnetic Resonance in Medicine* 17(2), 357–367.
- Traboulsee, A., D. Li, J. Frank, J. Simon, P. Coyle, J. Wolinsky, and D. Paty (2003). Consortium of ms centers: Mri protocol for the diagnosis and follow-up of ms. Technical report, CMSC MRI Protocol for MS.
- Tsyganok, V. V., S. V. Kadenko, and O. V. Andriichuk (2011). Significance of Expert Competence Consideration While Group Decision-Making using AHP. In *The International Symposium on the Analytic Hierarchy Process (ISAHP)*, pp. 1–7.
- Udupa, J. K., V. R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn (2006). A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics* 30(2), 75–87.
- Vaidya, O. S. and S. Kumar (2006). Analytic hierarchy process: An overview of applications. *European Journal of Operational Research* 169, 1–29.
- van Leemput, K., F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens (2001). Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans Medical Imaging* 20(8), 677–688.
- van Leemput, K., F. Maes, D. Vandermeulen, and P. Suetens (1999). Automated model-based tissue classification of MR images of the brain. *IEEE Trans Medical Imaging* 18, 897–908.

- Verhey, J., A. Ludwig, J. Rexilius, S. K. Warfield, C. Mamisch, R. Kikinis, C. F. Westin, R. Seibel, and O. Rienhoff (2002). Multimodale nicht-rigide Registrierung von Ultraschall- und MR Bilddaten unter Verwendung eines biomechanischen Modells. In *Workshop Bildverarbeitung für die Medizin*, pp. 310–313. Springer.
- Verhey, J., J. Wissler, S. K. Warfield, J. Rexilius, and R. Kikinis (2005). Non-rigid registration of a 3D ultrasound and a MR image data set of the female pelvic floor using a biomechanical model. *BioMedical Engineering OnLine* 4(19), 1–8.
- Visible Heart Project. Homepage at: <http://www.visibleheart.com/>. Accessed December, 2009.
- VMIP. Validation and evaluation in Medical Imaging Processing. Homepage at: <http://www.vmip.org/>. Accessed Dezember, 2012.
- Warfield, S. K., A. Guimond, A. Roche, A. Bharatha, A. Tei, F. Talos, J. Rexilius, J. Ruiz-Alzola, C.-F. Westin, S. Haker, S. Angenent, A. Tannenbaum, F. Jolesz, and R. Kikinis (2002). *Advanced Nonrigid Registration Algorithms for Image Fusion* (2 ed.), Chapter 24, pp. 661–690. Academic Press.
- Warfield, S. K., K. H. Zou, and W. M. W. III (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Medical Imaging* 23(7), 903–921.
- Wasserman, R., R. Acharya, C. Sibata, and K. H. Shin (1996). A Patient-Specific In Vivo Tumor Model. *Math Biosci.* 136(2), 111–140.
- Winer-Muram, H. T., S. G. Jennings, C. A. Meyer, Y. Liang, A. M. Aisen, R. D. Tarver, and R. C. M. and (2003). Effect of Varying CT Section Width on Volumetric Measurement of Lung Tumors and Application of Compensatory Equations. *Radiology* 229, 184–194.
- Wood, J., S. Green, and W. Shapiro (1988). The prognostic importance of tumor size in malignant gliomas: a computed tomographic scan study by the Brain Tumor Cooperative Group. *Journal of Clinical Oncology* 6, 338–343.
- Wu, H., G. C. Sharp, B. Salzberg, D. Kaeli, H. Shirato, and S. B. Jiang (2004). A finite state model for respiratory motion analysis in image guided radiation therapy. *Physics in Medicine and Biology* 49, 5357–5372.
- Yoshimura, K., H. Kato, M. Kuroda, A. Yoshida, K. Hanamoto, A. Tanaka, M. Tsunoda, S. Kanazawa, K. Shibuya, S. Kawasaki, and Y. Hiraki (2003). Development of a Tissue-Equivalent MRI Phantom Using Carrageenan Gel. *Magnetic Resonance in Medicine* 50(5), 1011–1017.
- Zhang, C., P. R. Bakic, and A. D. Maidment (2008). Development of an Anthropomorphic Breast Software Phantom Based on Region Growing Algorithm. In *SPIE Medical Imaging*, Volume 6918, pp. 69180V–1–69180V–10.

- Zhang, H., J. E. Fritts, and S. A. Goldman (2008). Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding* 110, 260–280.
- Zienkewicz, O. C. and R. L. Taylor (1987). *The Finite Element Method*. McGraw Hill Book Co.
- Zijdenbos, A. P., B. M. Dawant, R. A. Margolin, and A. C. Palmer (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Medical Imaging* 13(4), 716–724.
- Zubal, I. G., C. R. Harrell, E. O. Smith, Z. R. G. Gindi, and P. B. Hoffer (1994). Computerized three-dimensional segmented human anatomy. *Medical Physics* 21(2), 299–302.