JACOBS
UNIVERSITY

Fraunhofer
MEVIS

Lesion segmentation and tracking for CT-based chemotherapy monitoring

Jan Hendrik Moltz

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

Dissertation Committee:

Prof. Dr. Horst Karl Hahn
Jacobs University Bremen · Fraunhofer MEVIS
Prof. Dr. Andreas Nüchter
Jacobs University Bremen · Julius-Maximilians-Universität Würzburg
Prof. dr. Bram van Ginneken
Radboud Universiteit Nijmegen · Fraunhofer MEVIS

School of Engineering and Science
Jacobs University Bremen

Date of Defense: November 5, 2013

Summary

This PhD thesis makes contributions in the field of medical image analysis for assisting CT-based staging and follow-up examinations of patients under chemotherapy.

In the first part, an algorithm for semi-automatic segmentation of liver metastases in CT images is presented. The user interaction is a stroke across the lesion. Special care was taken of keeping the runtime within a clinically acceptable limit of less than 1 s on average. The method is based on histogram analysis, region growing, and morphological postprocessing. It is able to deal with inhomogeneous density distributions and prevents leakage through the liver boundary. A comprehensive evaluation of accuracy, reproducibility, and efficiency was performed on 371 test lesions with manual segmentations. The method produces results of similar quality as other state-of-the-art methods but is significantly faster.

In order to accelerate follow-up examinations in the clinic, I implemented a framework for automatic lesion tracking. For a segmented baseline lesion, it identifies the corresponding lesion in the follow-up image, automatically initializes the segmentation, and performs a plausibility check. The method is general, but optimized for lung nodules, liver metastases and lymph nodes. So far, no other framework exists that automatizes follow-up examinations to this degree. The second part of the thesis starts with a problem analysis, examining the change of 994 follow-up lesions under chemotherapy. A simulation of the behavior of different similarity measures on a lesion phantom motivates the subsequent presentation of a template matching algorithm tailored to this problem. The stages of the method are validated from a technical point of view on 207 independent cases before reporting a user study that evaluated possible benefits of the method for the clinical workflow.

For validating segmentation algorithms, manual delineations are often used, but their high variability makes it difficult to achieve reliable statements. The third part of my thesis collects some ideas how to quantify this problem and to overcome it in practice. Liver tumor segmentation in CT is used as a consistent example. First, I present a generalization of the MICCAI Grand Challenge score that takes the variability of multiple reference segmentations in account for each case. Second, an analysis of the variability in manual delineations of ten experts is performed, using a novel methodology for measuring the variability within a set of segmentations. The part is concluded by a concept and a validation study for a tool that allows experts to generate probabilistic reference segmentations.

Contents

1 Introduction 7
- 1.1 About this thesis 7
- 1.2 Clinical motivation 7
- 1.3 The images 8
- 1.4 Software assistant 9
- 1.5 Measures for comparing segmentations 10
- 1.6 Challenges in commercially oriented research 11

I Segmentation of liver lesions 13

2 Introduction 15
- 2.1 Introduction 15
- 2.2 Appearance of liver lesions 16
- 2.3 Related work 17
- 2.4 Data 20

3 Algorithm 21
- 3.1 The basic algorithm: "smart opening" 21
- 3.2 Histogram analysis and threshold selection 22
- 3.3 Special ROI configurations 26
- 3.4 Two-step segmentation of inhomogeneous lesions 30
- 3.5 Segmentation of peripheral liver metastases 32

4 Evaluation 37
- 4.1 Technical evaluation 37
- 4.2 Evaluation with multiple reference segmentations 45
- 4.3 Clinical evaluation 46
- 4.4 Discussion 49

II Automatic lesion tracking 51

5 Introduction 53
- 5.1 Introduction 53
- 5.2 Related work 54

6 Data and problem analysis 57
- 6.1 Data 57

6.2 Statistical analysis of change 59
6.3 Measuring similarity 62
6.4 Simulation of change 63
6.5 Discussion 66

7 Algorithm and technical evaluation 69
7.1 Overview 69
7.2 Global registration 70
7.3 Candidate detection and template matching 71
7.4 Stroke propagation and segmentation initialization 83
7.5 Plausibility check 88

8 Workflow-centered evaluation 97
8.1 Goals 97
8.2 Materials and methods 97
8.3 Results 99
8.4 Discussion 107

III Uncertainty-aware validation of segmentation algorithms 109

9 Validation using multiple reference segmentations 111
9.1 Introduction 111
9.2 Related work: A critical review 113
9.3 A score for uncertainty-aware validation 117
9.4 Experiments 118
9.5 Results 118
9.6 Discussion 122

10 Variability of manual segmentations 125
10.1 Introduction 125
10.2 Related work 125
10.3 Data 126
10.4 Methodology 126
10.5 Results 128
10.6 Discussion 132

11 A tool for creating probabilistic expert segmentations 135
11.1 Introduction 135
11.2 Related work 135
11.3 Workflow 136
11.4 Evaluation 137
11.5 Discussion 139

12 Conclusion 143

Chapter 1

Introduction

1.1 About this thesis

This PhD thesis is based on my work as a research scientist at Fraunhofer MEVIS between 2006 and 2013. In agreement with the principal orientation of the institute, it is largely application-centered and a result of close cooperation with clinical and industrial partners. Specifically, my work was part of a joint project with Siemens Healthcare on the one hand and radiologists from the university hospitals in Berlin, Kiel, Mainz, Marburg, München, and Münster on the other hand. The project aims at providing software assistance for CT-based staging and follow-up examinations of patients under chemotherapy. This includes the development of quantitative image analysis techniques such as the semi-automatic segmentation and volumetry of lesions, but also the design of workflow support for detecting lesions and tracking their changes over time. The entities in focus are lung nodules, liver metastases and enlarged lymph nodes. Of the three main parts of this thesis, the first two present results that have an immediate clinical application and a clear commercial perspective. This is complemented by a more theoretical part that illuminates validation methodology for the methods developed earlier.

1.2 Clinical motivation

Chemotherapy is a treatment for cancer patients that is associated with severe side effects and high costs. Therefore it is important to estimate the success of a therapy as soon as possible. Typically, CT examinations are performed in intervals of three to six months. Reading these follow-up examinations is one of the major tasks of a radiologist. The most important criterion for standard therapies is the change in size. Without software support, it is only feasible to measure size in terms of the diameter. This procedure has been standardized for clinical studies by an international consortium in 2000 (Therasse et al. 2000). According to the latest version of *RECIST* (Response Evaluation Criteria in Solid Tumors) from 2009 (Eisenhauer et al. 2009), the largest diameter in an arbitrary but consistent orientation is measured. The only exception are lymph nodes, where the diameter perpendicular to the largest diameter is used.

Being a one-dimensional measurement, the diameter can only give a coarse indication of the actual volume of a lesion. RECIST defines 20 % growth or 30 % shrinkage of the diameter as significant change. This corresponds to 73 % or 66 % volume change, respectively, if spherical shape and uniform growth are assumed. These assumptions, however, are not true in general and may lead to wrong conclusions in some cases. Furthermore, manual measurements are always associated with inaccuracies and inconsistencies, as has been investigated in several studies. A famous example is an experiment by Erasmus et al. (2003) where five radiologists examined the

same lung nodules. In almost 30 % of the patients the difference between the results of different readers exceeded the RECIST threshold for progressive disease. The measurements were repeated after a week, and even within the same reader this number was as high as 10 %. This is not surprising if we consider that for a lesion with a typical diameter of 20 mm and an in-plane resolution of 1 mm, the growth required by RECIST corresponds to just four voxels. It can easily be imagined that the typical error of a manual measurement on an image with noise and partial volume effects is in the same order of magnitude.

The use of semi-automatic volumetry aims to handle both problems at the same time. While manual contouring in all slices is much too time-consuming in practice, a computer is able to perform a three-dimensional measurement in a few seconds. Also, an algorithm can give results with lower variability, even if it depends on some user initialization. Of course, precision and efficiency of an algorithm have to be examined carefully and compared to the current standard procedures in the clinic; in addition to accuracy, which is necessary to make computer-based methods usable in the first place.

Some recent treatment options do not aim at reducing the size of a tumor, but still a semi-automatic delineation can be useful for therapy monitoring. If the mean density or the necrosis fraction of a lesion is the parameter of interest, it can easily be extracted. If, as in ablation procedures, an apparent growth by a particular safety margin is expected, this can be visually verified once the extents of the lesion have been determined.

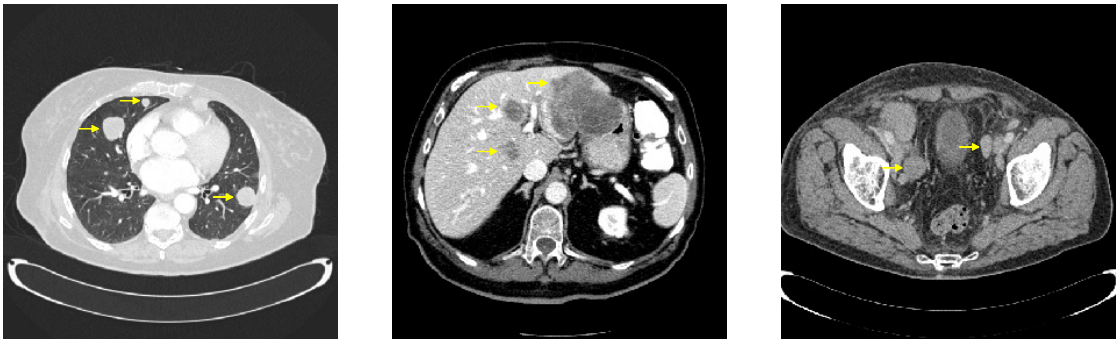
However, providing accurate and reproducible measurements is not the only purpose of medical image analysis in the context of chemotherapy monitoring. Workflow support and thus increasing efficiency and avoiding errors is gaining importance. Among our clinical partners, many radiologists complain about their workload and appreciate software that automatizes tedious procedures and allows them to focus on tasks that actually require expert knowledge.

1.3 The images

The images used in this thesis are acquired by *computed tomography* (CT). CT scans are three-dimensional and usually consist of a stack of axial slices which show cross-sections of the human body. The resolution is about 0.5 to 1 mm on each slice and 1 to 5 mm orthogonal to the slices. A slice typically consists of 512×512 voxels, whereas the number of slices depends on the imaged body region and the resolution.

The image values are standardized and have a range of –1000 to 3072 *Hounsfield units* (HU). Low values correspond to low attenuation and are visualized with darker gray values. Therefore, regions filled with air, such as the background and the lungs, are dark, and the brightest structures are typically bones. Since CT is based on X-ray attenuation, the values have a physical meaning and correspond directly to the tissue density. This allows using fixed thresholds for simple segmentation tasks such as lung segmentation or for checking the plausibility of segmentation results of particular tissues.

Structures can be enhanced by applying *contrast agents* that have a higher attentuation than soft tissue. When they are injected into a vein, blood vessels get higher HU values. Depending on the time between injection and image acquisition, arteries or veins are enhanced.



(a) Lung nodules (b) Liver metastases (c) Lymph nodes in the pelvis

Figure 1.1: Example CT slices showing the tumor entities relevant for this thesis.

Figure 1.1 shows slices of CT scans to illustrate the appearance of tumors and their surroundings. Lung nodules are bright spots in the lungs, typically with a high contrast to the lung parenchyma. Liver metastases can be bright or dark spots in the liver, depending on the contrast agent state and type (see Section 2.2 for details). Lymph nodes are only visible in CT when they are pathologically enlarged. They have a spherical shape and appear at particular positions in the body, often along the main vessels.

1.4 Software assistant

The *Oncology Prototype Software* is a software assistant that was developed at Fraunhofer MEVIS to support volumetric follow-up examinations. The methods described in this thesis were integrated into the software, which was then delivered to the clinical partners for evaluation. In order to illuminate the context in which the new methods are used, the workflow of the software assistant is summarized in this section. An earlier version has been described in more detail by Bornemann et al. (2007). Since it is essentially a prototype for evaluating algorithms, it runs on standalone computers and the data have to be imported manually.

The software offers semi-automatic segmentation methods for lung nodules, liver metastases and enlarged lymph nodes (Moltz et al. 2009b). The user selects a segmentation method, draws a stroke across the lesion and gets an initial segmentation typically after less than 2 s, depending on the size and complexity of the lesion. The segmentation algorithm for liver lesions will be described in detail in Part I. Various tools are available for efficient manual refinement of segmentation results. When the user draws a partial contour to add or remove parts of the segmentation in a single slice, this correction is automatically propagated to a set of adjacent slices. The propagation can be chosen to be either image-based (Heckel et al. 2009) or purely geometrical, based on a method described by Heckel et al. (2011). In either case, the user has the complete control over the final segmentation result.

For comparing lesions in baseline and follow-up images, both datasets are loaded and the results of the baseline examination are displayed. An optional synchronization of the viewers aligns

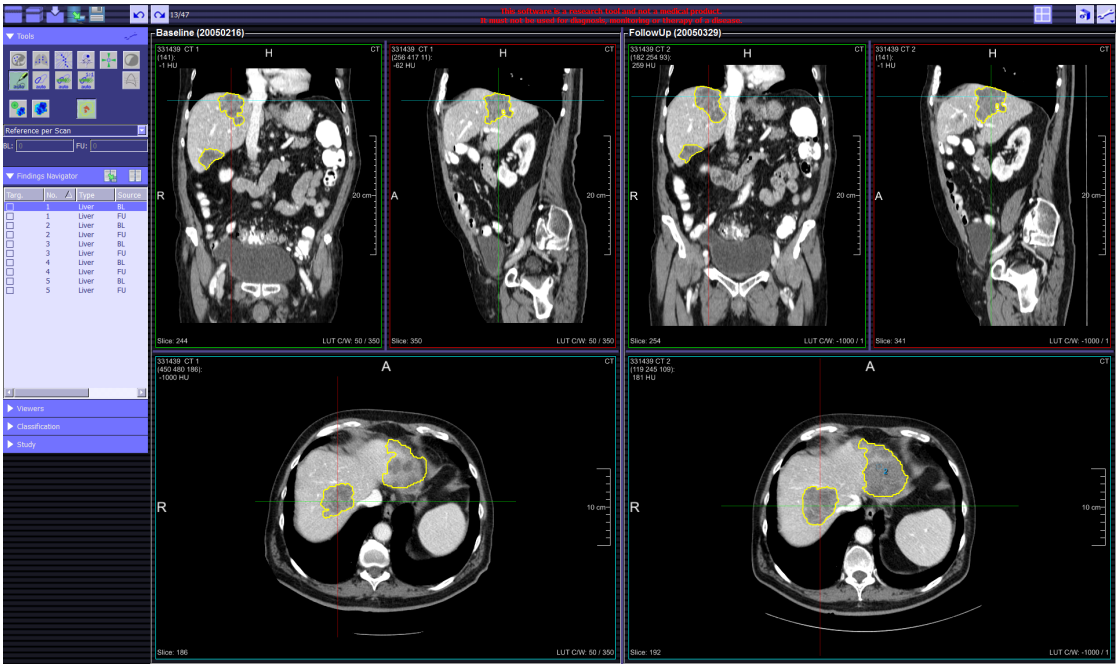


Figure 1.2: Screenshot of the Oncology Prototype Software in follow-up mode.

anatomically corresponding slices automatically and helps to find the segmented lesions in the follow-up image and to detect changes.

An automatic lesion tracking, presented in Part II, can be run as a preprocessing step before a radiologist starts reading a case. It identifies the target lesions in the follow-up image and computes the segmentations. The radiologist then checks whether the correct lesion was found and refines the segmentation if necessary. If a wrong lesion was segmented, the user can discard the result and initialize a new segmentation. The algorithm contains a mechanism that discards implausible results automatically in order to account for lesions that vanish under therapy and for difficult cases with a large number of lesions or strong anatomical changes. In such cases, no precomputed results are available.

Figure 1.2 shows a screenshot of the Oncology Prototype Software. It contains two viewing areas with three orthogonal viewers each, displaying the baseline and follow-up images. In this example, a pair of CT images of a patient with liver metastases has been loaded. The segmentation results are overlaid by yellow outlines. The menu area on the lefthand side contains a list of the findings and several buttons to select segmentation algorithms, start tracking for individual lesions, display a report and start various other functionalities.

1.5 Measures for comparing segmentations

Since at several points in this thesis segmentation results are compared, an overview of common measures is given here. Let A be an algorithmic segmentation and M a manual reference segmentation.

Volume-based measures interpret the masks as voxel sets. The *volume overlap* or *Jaccard coefficient*

$$\frac{|A \cap M|}{|A \cup M|} \tag{1.1}$$

and the *Dice coefficient*

$$\frac{|A \cap M|}{\frac{1}{2}(|A| + |M|)} \tag{1.2}$$

are equivalent measures of agreement. If volumetry is the focus, the *relative volume error*

$$\frac{||A| - |M||}{|M|} \tag{1.3}$$

is a suitable metric.

Another possibility is to interpret the segmentations as sampled surfaces and compute distances between them. From the definition of a point-to-mask distance

$$d(\boldsymbol{a}, M) = \min_{\boldsymbol{m} \in M} ||\boldsymbol{m} - \boldsymbol{a}||, \tag{1.4}$$

we can derive the *mean surface distance*

$$\operatorname*{mean}_{\boldsymbol{a} \in A} d(\boldsymbol{a}, M) \tag{1.5}$$

and the *maximum surface distance* or *Hausdorff distance*

$$\max_{\boldsymbol{a} \in A} d(\boldsymbol{a}, M). \tag{1.6}$$

1.6 Challenges in commercially oriented research

Large parts of this thesis were written in a product-oriented environment. The algorithms were developed specifically with a commercial exploitation in mind and in close cooperation with an industrial partner. This resulted in a slightly different prioritization than is common in academic research, which had an effect on some of the design decisions.

Efficiency is a paramount criterion. From the start, algorithms were developed in an optimized environment, using *MeVisLab* (Ritter et al. 2011) and C++ implementations of most image processing functionalities, and the choice of methods was limited by computation time restrictions. Therefore, the thesis explores how simple and fast methods can be used to achieve accurate results.

Another important criterion is plausibility of the results. Users typically expect good results for cases they consider easy. If this is not fulfilled, they might lose their trust in the software and stop using it although it might work well on difficult cases. The more complex an algorithm is internally, the harder it gets for users to predict its behavior. Optimally, users should be able to get a good sense for the strengths and weaknesses of an algorithm. This requirement is not necessarily equivalent with the highest quality in the sense of accuracy evaluation and is therefore hard to verify formally.

A more technical challenge results from the fact that the segmentation algorithm presented in this thesis was integrated into a commercial software at an early stage. For all further developments, the requirement was essentially to leave all good results unchanged and at the same time increase the number of cases having good results. Hardly being possible at all, this means that initial decisions could not easily be reverted. It also emphasizes the importance of regression testing. Determining whether a result has become "better" or "worse" requires a sophisticated validation framework and data. This was the motivation for the deeper investigation of algorithm validation techniques in Part III.

Part I

Segmentation of liver lesions

Contributions An algorithm for semi-automatic segmentation of liver metastases in CT images with a clinically acceptable runtime of less than 1 s on average.
Special consideration of inhomogeneous and peripheral lesions.
A comprehensive evaluation of accuracy, reproducibility and efficiency on 371 test lesions with manual segmentations.

Acknowledgments Some of the work described in this part was done in close cooperation with my colleague Lars Bornemann, and a clear separation of authorship is not always possible. Valuable feedback was provided in several workshops at Fraunhofer MEVIS by our clinical partners Hans-Christian Bauknecht, Hendrik Bolte, Michael Fabel, Markus Hittinger, Andreas Kießling, Stephan Meier, Elena Peitgen, and Michael Püsken. The manual segmentations for the evaluation were created by Christiane Engel, Michaela Jesse, Ulrike Kayser, and Susanne Zentis.
This work was funded in part by Siemens AG, Healthcare Sector, Imaging & IT Division, Computed Tomography, Forchheim, Germany.

Publications A preliminary description of the algorithm has been published in a special issue on *Digital Image Processing Techniques for Oncology* of the *IEEE Journal of Selected Topics in Signal Processing* (Moltz et al. 2009b). The method participated and was presented in the *3D Liver Tumor Segmentation Challenge* at *Medical Image Computing and Computer Assisted Intervention 2008* in New York. An oral presentation at the *European Congress of Radiology 2010* in Vienna was distinguished as a *Best Scientific Paper Presentation*. Parts of the algorithm and the evaluation with manual segmentations have not been published previously.

Chapter 2

Introduction

2.1 Introduction

Segmentation is the main requirement for automated lesion volumetry. Generally, segmentation methods can be subdivided into automatic, semi-automatic and interactive ones. While an *automatic* method only needs the image itself as an input, the other methods also require some indication on where the object to be segmented is located. A *semi-automatic* method will then do the remaining computation autonomously while *interactive* methods typically consist of a loop of user input and processing.

While a higher degree of automation generally seems to be a better choice in terms of user workload and reproducibility, there is practically no segmentation task that can be solved fully automatically in all conceivable cases. Furthermore, if an algorithm first has to detect an object of interest, the computation time is inherently higher, if only because it has to process the entire image and not just a *region of interest* (ROI) that has been derived from the user input. As a compromise, methods with minimal user input are often used in practice. A single click can define the location of an object, but not its extents. Possible interactions that indicate size are a line through the object or a rough outline in one slice which can be free-hand or a parameterized shape like an ellipse. In the context of lesion segmentation for chemotherapy monitoring, drawing the largest diameter is a natural initialization since it corresponds to the current standard measurement. The only difference is that it does not have to be as exact. Both a deviation from the largest diameter of the object and a deviation from its boundary in the range a few voxels should be tolerated by the algorithm. Such a coarse approximation of the RECIST diameter will be called a *stroke*.

This analogy also allows an estimate of the admissible computation time of a segmentation algorithm. As a general rule, clinicians would not use semi-automatic volumetry if it takes significantly longer than a manual diameter measurement. The visual inspection of a lesion and the careful adjustment of the largest diameter takes a few seconds, so this is the time an algorithm may spend since the stroke can be drawn faster. However, since the segmentation result also has to be verified and possibly corrected by the user, a 3 s limit for the processing time was decided that must be kept in 90 % of the cases. This is important since virtually all state-of-the-art methods are quite far away from fulfilling this clinical acceptance criterion.

When I started working in the project, a mature algorithm for lung nodule segmentation (Kuhnigk et al. 2006) and a preliminary version of a segmentation algorithm for liver and brain lesions (Bornemann et al. 2007) were already available. They formed the basis of the developments described in this part.

2.2 Appearance of liver lesions

The research in this part focuses on the segmentation of *liver metastases* because among hepatic lesions they are most relevant for chemotherapy monitoring. Primary liver tumors like *hepatocellular carcinoma* (HCC) are much less frequent in most parts of the world and are typically treated surgically or by interventions such as radiofrequency ablation. Still, from an algorithmic point of view, the actual diagnosis is less important than the appearance of a lesion. Since some HCCs look much like metastases, the algorithm might still be able to segment them. In general, however, HCCs pose problems that would require a different algorithmic approach. In the following, I assume that a lesion is either homogeneous and clearly darker or brighter than the liver parenchyma, or that it is composed of a core and a rim, each of which is homogeneous in itself.

In native CT scans, most liver lesions are hardly visible, because their density does not differ much from that of the healthy parenchyma. The usage of contrast agent, however, allows not only the detection of lesions, but results in different enhancement patterns which carry further information about the lesion type.

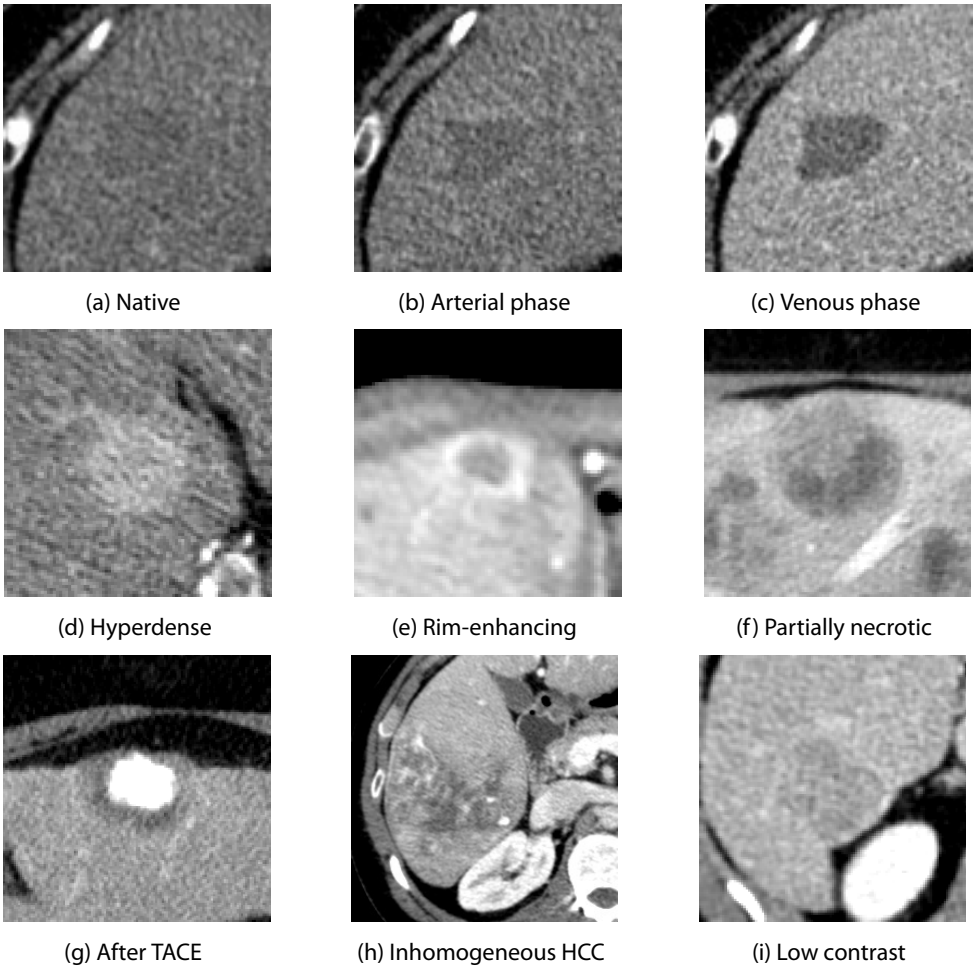
Liver metastases are mostly *hypovascularized*. Since they do not take up contrast agent themselves, they are best visible when the parenchyma is enhanced, which is the case in the *venous phase*. They will then appear *hypodense*, i.e., darker than their surroundings. Figures 2.1a to 2.1c show such a metastasis in three different contrast phases.

HCCs and some metastases are *hypervascularized*. These lesions enhance in the *arterial phase*, earlier than the parenchyma, and appear *hyperdense* (Figure 2.1d). In the venous phase, the parenchyma enhances as well and the lesion is no longer discernible.

Rim-enhancing lesions are a hybrid type, composed of a hypovascularized core surrounded by a hypervascularized rim (Figure 2.1e). Lesions may also have an irregular density distribution. For instance, they can have areas of *necrosis* which are darker than the rest of the lesion (Figure 2.1f), whereas brighter spots are most often caused by *calcification* or *transarterial chemoembolization* (TACE). The latter is a treatment where a drug is injected to block the blood supply of the lesions. The deposits of the drug appear as strongly hyperdense spots in the lesion (Figure 2.1g). Large lesions, mostly HCCs, can have inhomogeneous vascularization, resulting in irregular patterns of hypodense and hyperdense areas (Figure 2.1h). This last class will not be targeted by the developed method.

Depending on the imaging parameters, but also on the general condition of the liver, the contrast-to-noise ratio can be quite low (Figure 2.1i). This makes liver lesions more difficult to segment than, for example, lung nodules. This is even a problem for manual segmentation, because the boundaries are often diffuse. For threshold-based segmentation algorithms, this means that the thresholds have to be chosen carefully and that further pre- and postprocessing will be necessary.

Another important issue when using thresholding is the possible vicinity of a lesion to a structure of similar density. Several anatomical structures around the liver can look similar as a lesion, most importantly the intercostal musculature (a band of muscles spanned by the ribs) or the parenchyma of other abdominal organs such as the kidneys or the stomach. Lesions adjacent to one of these structures often have no or very low contrast to them and can visually be delineated only by extrapolating the liver shape.



(a) Native (b) Arterial phase (c) Venous phase

(d) Hyperdense (e) Rim-enhancing (f) Partially necrotic

(g) After TACE (h) Inhomogeneous HCC (i) Low contrast

Figure 2.1: Examples of liver lesions in CT.

2.3 Related work

Regarding lesion segmentation in CT, lung nodules were the first entity for which algorithms were developed and established in the clinic. The segmentation of liver metastases, in contrast, had not been an area of intensive research when I started my work in 2006. Since about 2008, however, especially elicited by the Liver Tumor Segmentation Challenge at MICCAI 2008 (Deng and Du 2008), the number of publications has increased substantially.

In order to structure the available literature, papers can first be classified by their level of automation. Completely automatic methods that include segmentation of the liver and detection of tumors will not be considered in the following review. Since the task is much more complex than segmenting a given tumor, these methods are computationally more expensive and mostly less accurate with respect to the final segmentation quality.

On the level of semi-automatic methods, more or less all modern segmentation approaches have been applied to liver tumors, often in different variations.

2.3.1 Region growing

A method as simple as region growing can be successfully applied for liver tumor segmentation if it is combined with adequate threshold selection criteria and some shape constraints or morphological postprocessing. Zhao et al. (2006) incorporate knowledge about the expected local and global shape directly into the region growing process. This is meant to avoid irregularly shaped segmentation masks and leakage at thin connections to isodense structures. Region growing is performed separately on each slice. A similar approach was used by Wong et al. (2008) in the Challenge. However, they only apply a size constraint to the segmentation.

The in-house method (Bornemann et al. 2007) my work is based upon uses 3D region growing, but it is purely threshold-based and the morphological refinement is done afterwards. In fact, it is an extension of our lung nodule segmentation (Kuhnigk et al. 2006). The core of this method, the "smart opening" procedure, is described in Section 3.1.

2.3.2 Level sets

Two different applications of the level sets concepts were described by Cai et al. (2007) and Smeets et al. (2010). Cai's basic idea is to consider a shell whose medial line is the current segmentation boundary. Under the assumption that ideally the histogram of the shell should have two peaks of similar height, the optimization of the shell location is driven by a level set speed function. Smeets, who won the Challenge with his method, uses level sets in a more classical sense with the speed image based on a fuzzy pixel classification. The level set is initialized by applying a spiral-scanning technique.

2.3.3 Watershed transform

Another classical approach, the watershed transform, was tested in different variants by Ray et al. (2008). They use an iterative version on a gradient vector field rather than on the gradient image and apply this technique in 2D and 3D. The user can influence the result by setting markers.

2.3.4 Graph cuts and other graph-based methods

In the last couple of years, graph cuts and other graph-based methods seemed to be the most popular solution to the liver tumor segmentation problem. Stawiawski et al. (2008) combine it with a watershed transform and work on the resulting region adjacency graph rather than the voxel adjancency graph. Szilágyi et al. (2009) apply graph cuts on voxels, but their boundary prior relies not only on intensities but also on local phase information. Drechsler et al. (2011) apply a multi-resolution graph cuts scheme that reduces complexity without demanding the user to specify a region of interest. A comparison of different graph-based methods including graph cut and random walker was done by Su et al. (2011) and no significant differences in accuracy were found for five of the six methods considered. Another application of the random walker algorithm was proposed by Jolly and Grady (2008). Their method is suitable for different kinds of lesions including liver metastases. The seeds for the random walker are computed automatically from 2D presegmentations based on fuzzy connectedness.

A particularly interesting method was presented by Li and Jolly (2008). They search for a path in a graph that optimizes particular boundary, regional and elasticity constraints. Notably, the algorithm is able to detect multiple surfaces simultaneously, making it possible to segment tumors with necroses and calcifications in one pass. This is so far the only method that explicitly addressed such cases.

2.3.5 Machine learning approaches

Some methods based on different machine learning approaches should be mentioned. Li et al. (2006) train a classifier to detect the boundary positions on 1d intensity profiles. Unfortunately, the paper does not state clearly how the training samples are generated. In the method by Zhou et al. (2008), which ranked second in the Challenge, a support vector machine is trained to classify voxels as tumor or background by manually selected samples. The classification works slice by slice and propagates its results from the center to the margins. Finally, Taieb et al. (2008) iteratively apply a smoothed Bayesian classifier based on a multi-class intensity model and refine the result by geodesic active contours.

2.3.6 Statistical approaches

The algorithm of Häme and Pollari (2012) is based on non-parametric intensity distribution estimation and a hidden Markov measure field model with a spherical shape prior. It was designed to achieve reliable results in cases with low contrast-to-noise ratio. It can, however, not automatically separate attached isodense structures from the lesion. Therefore, a post-processing step removes "handle"-shaped regions at the boundary. The method was extended to use multiple images from different contrast phases and perform a combined optimization. While this increases robustness, it requires a computationally expensive registration.

2.3.7 Discussion

Given this diversity of approaches that have been applied to liver tumor segmentation, it is hard to tell whether one is more adequate than the other. In the Challenge at MICCAI 2008, methods as different as level sets, voxel classification and region growing showed almost equal accuracy. Methods that did not participate in the challenge can only be compared analytically since the reported results were obtained on different data sets. Also, computation times are often marked as "non-optimized" or not reported at all. None of the cited publications states a runtime even close to the 3 s limit that is required for my development. The challenge winners reported times of 20 s to 2 min (Smeets et al. 2010) and 7 to 30 min (Zhou et al. 2008).

Only few of the available methods deal explicitly with the possible problems stated in Section 2.2: inhomogeneous lesions and lesions with contact to similar structures. Algorithms that incorporate shape constraints may be able to prevent leakage outside the liver in some cases, but probably not for large lesions or when the connection to an isodense structure is extensive. "Easy" solutions such as computing a liver segmentation beforehand or allowing the user to set background markers in such structures were not allowed in my project.

As already mentioned, only Li and Jolly (2008) have dealt explicitly with severely inhomogeneous lesions that feature necroses or calcifications. They do not show any examples of rim-enhancing

Dataset	Hospitals	Scanners	Patients	Lesions
Siemens	5 (DE, US)	Siemens	15	72
MICCAI Challenge	undisclosed	undisclosed	4	10
MeVis Distant Services	27 (CH, CN, DE, JP, SE, US)	GE, Philips, Siemens, Toshiba	152	371

Table 2.1: Overview of available data sets with manual segmentations.

Parameter	Min	Median	Max
Slice thickness (mm)	0.5	1.0	5.0
Reconstruction increment (mm)	0.5	1.0	5.0
Tube voltage (kV)	120	120	140
Tube current (mAs)	100	293	849
Number of frames	46	190	505
Pixel spacing (mm)	0.53	0.71	0.98
Reconstruction kernel	Siemens: B10s–B60f Philips: B Toshiba: FC03–FC13 GE: Soft–Standard		

Table 2.2: Overview of acquisition parameters of the available data.

metastases. Since their method is completely different than the one on top of which I am building my extensions, I cannot apply their methodology.

2.4 Data

An overview of the available data sets for development and evaluation is given in Table 2.1. In total, there are 453 lesions from 171 patients with manual segmentations drawn by experts (radiologists or radiology technicians). These are from a variety of clinics in Europe, North America and Asia and scanners by four different manufacturers (Table 2.1). They include test data provided by Siemens, the training data from the MICCAI Liver Tumor Segmentation Challenge, and data submitted to MeVis Distant Services for liver surgery planning. The imaging parameters are summarized in Table 2.2.

The first two data sets as well as a large amount of other data without reference segmentations were used for development and parameter optimization. The third data set was used strictly for evaluation purposes. Together, the data form a representative collection of liver lesions and the reference segmentations are completely independent of the algorithm.

Chapter 3

Algorithm

3.1 The basic algorithm: "smart opening"

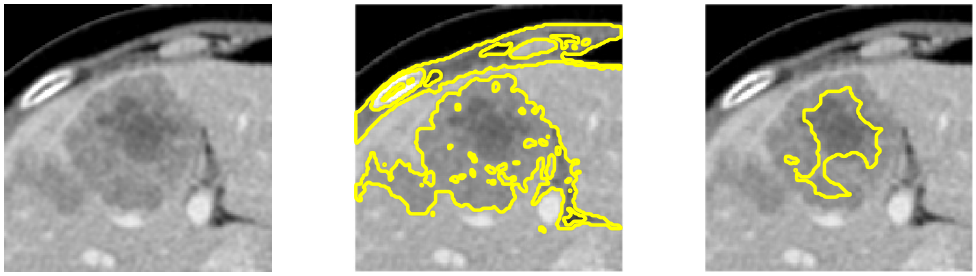
The *smart opening* algorithm was developed by Kuhnigk et al. (2006) for segmenting solid lung nodules. It tackles a problem that is relevant for other tumor types as well: the contact to vessels or other thin, elongated structures that feature a similar density in CT scans. It is obvious that a mere threshold-based method is not sufficient for segmentation in such a situation. Since the lesions are mostly homogeneous it can, however, be used as a first step to obtain a superset of voxels that may be part of the lesion. This can be implemented efficiently as a 3D region growing starting from the center of the ROI. The thresholds can be fixed for lung nodules or determined adaptively from an analysis of the density distribution in the ROI.

For a typical lung nodule, the region growing result contains the complete lesion and additionally parts of the attached vasculature. A morphological opening operation is an obvious choice to remove the vessels, but the challenge lies in determining the optimal erosion strength. Since tumors and their supplying vessels can differ in size significantly, an erosion with a fixed-size kernel cannot be used since this would either maintain thick vessels that should be removed or erode the lesion too strongly so that details of its boundary would be lost.

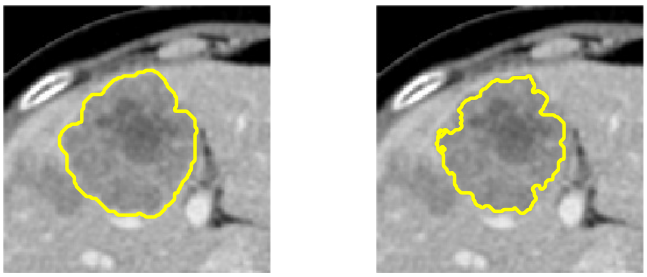
The idea of smart opening is to choose the erosion strength adaptively. To facilitate the computations, erosion is implemented by thresholding on a distance map that contains the distance of each mask voxel to the closest background voxel. The underlying assumption is that all vessels are connected to the boundary of the ROI and that the diameter of a vessel decreases monotonically in its course. In order to disconnect the mask, all paths from the ROI center to the boundary are considered. The maximum of all minimum path diameters is the cut-off value that removes all vessels.

The second step of the opening operation is a dilation that reconstructs the lesion in its original size without regrowing the vessels. The dilation is again implemented by thresholding on a distance map, but this time it shows the distances of all background voxels to the eroded mask. The threshold is chosen slightly higher than the erosion threshold so that in a final refinement step all boundary details can be reproduced by intersection with the region growing result.

It should be noted that this procedure alone is not suitable if more extensive connections to structures of similar density are present. In the liver, this problem occurs in some lesions at the organ boundary if they have contact to isodense structures outside the liver, most importantly the intercostal musculature. This requires a special handling, either before or after smart opening.



(a) Original (b) Region growing (c) Erosion



(d) Slight over-dilation (e) Boundary refinement

Figure 3.1: Example application of smart opening on a liver metastasis. Note how the connections to three adjacent structures (muscle, biliary duct and another lesion) are removed.

3.2 Histogram analysis and threshold selection

Although initially developed for lung nodules, smart opening is also a suitable basis for segmenting liver metastases. While vessels are only a problem for hyperdense lesions, it can handle non-extensive contact to the biliary duct, other lesions or neighboring organs as well. Also, due to the low contrast-to-noise ratio, smart opening helps to create results with compact shapes, which cannot be generated by thresholding alone. Figure 3.1 shows an example where smart opening is applied to a liver metastasis.

For lung nodules, a fixed threshold range was used for the initial region growing. For liver metastases, due to the high diversity in their appearance, the thresholds have to be determined adaptively. This will be done based on an analysis of the density distribution in the ROI. Information given by the user is reflected in the size and center of the ROI and in the stroke that is assumed to be an approximation of the maximum diameter of the lesion. The density distribution under the stroke provides information about the "relative density" of the lesion compared to the parenchyma and can be used to detect inhomogeneous metastases.

Bornemann et al. (2007) described a generic procedure for computing thresholds that has been applied to liver and brain lesions. It is based on two estimates: ℓ, the typical lesion value, is the value at the stroke center after Gaussian smoothing, and p, the typical parenchyma value, is the highest peak of the ROI histogram. Then, three cases are considered, using a threshold δ that describes the minimum expected contrast between lesion and parenchyma.

If $\ell < p - \delta$, the lesion is hypodense and the threshold range is set to $[-\infty, \frac{1}{2}(\ell + p)]$. Analogously, if $\ell > p + \delta$, the lesion is hyperdense and the threshold range is set to $[\frac{1}{2}(\ell + p), \infty]$. If, however,

$|\ell - p| < \delta$, it is assumed that p is not a suitable value for separation and a fixed-width interval around ℓ is used instead. This interval may be narrowed down iteratively if it includes too many voxels.

When this procedure is applied to a particular lesion type with known properties in CT imaging, the open "outer" thresholds are not optimal. For instance, if a hypodense lesion lies adjacent to the lung, the latter is included in the threshold range. Similarly, a hyperdense lesion can have contact to a rib or a contrast-enhanced vessel. Therefore, as a first change, the thresholds were set to 10 HU and 180 HU, respectively. These values were extracted by analyzing the development data. Fixed thresholds are preferable where possible because they reduce the dependency of the user interaction.

The chosen thresholds span the typical range of lesions, but there might be cases where the stroke clearly covers even darker or brighter regions. So far, however, only the stroke center is taken into account. Since liver lesions can have low contrast to the parenchyma and have an inhomogeneous density distribution, the values under the stroke can give additional information or allow more robust statistics. Therefore, another histogram is computed for the stroke voxels where the stroke is dilated with a $3 \times 3 \times 3$ kernel without elongating it in order to get a larger amount of representative lesion values. In order to make the threshold range cover unusually dark or bright lesions, it is extended to the 10 % or 90 % quantiles of the stroke histogram. This choice of quantiles accounts for noise and strokes that possibly extend into the background. Quantiles are also more reproducible than the actual extrema.

At this point, the threshold range is

$$\left[\min(10\,\mathrm{HU}, \tilde{\ell}_{0.1}),\ \frac{1}{2}(\ell + p)\right] \tag{3.1}$$

for hypodense lesions and

$$\left[\frac{1}{2}(\ell + p),\ \max(180\,\mathrm{HU}, \tilde{\ell}_{0.9})\right] \tag{3.2}$$

for hyperdense lesions, where $\tilde{\ell}_\bullet$ denotes quantiles of the stroke histogram. This is already sufficient for a large amount of liver lesions which can be adequately described by a single typical lesion and background value, i.e., for cases with homogeneous density within and outside the lesion and clear contrast.

The general procedure from Bornemann et al. (2007) invokes a special handling if the difference between ℓ and p is small. This handling is heuristic and does not take into account that there can be different reasons which require different reactions. Of course, there can be cases where the contrast is actually lower than the value of δ that has been determined by parameter optimization. Most of the time, however, it has to be assumed that either ℓ or p are not representative of the lesion or background, respectively, or that a single value is not sufficient to describe a distribution that is more complex due to inhomogeneities.

Therefore, a more detailed analysis of the *ROI configuration* is introduced. Such a configuration is described by a list of lesion and background values. The threshold range is then determined depending on the number and relation of these values. Figure 3.2 shows an overview of nine ROI configurations for hypodense lesions. By inverting the gray value relations, the corresponding configurations for hyperdense lesions can be generated. Together, this set is sufficient to describe

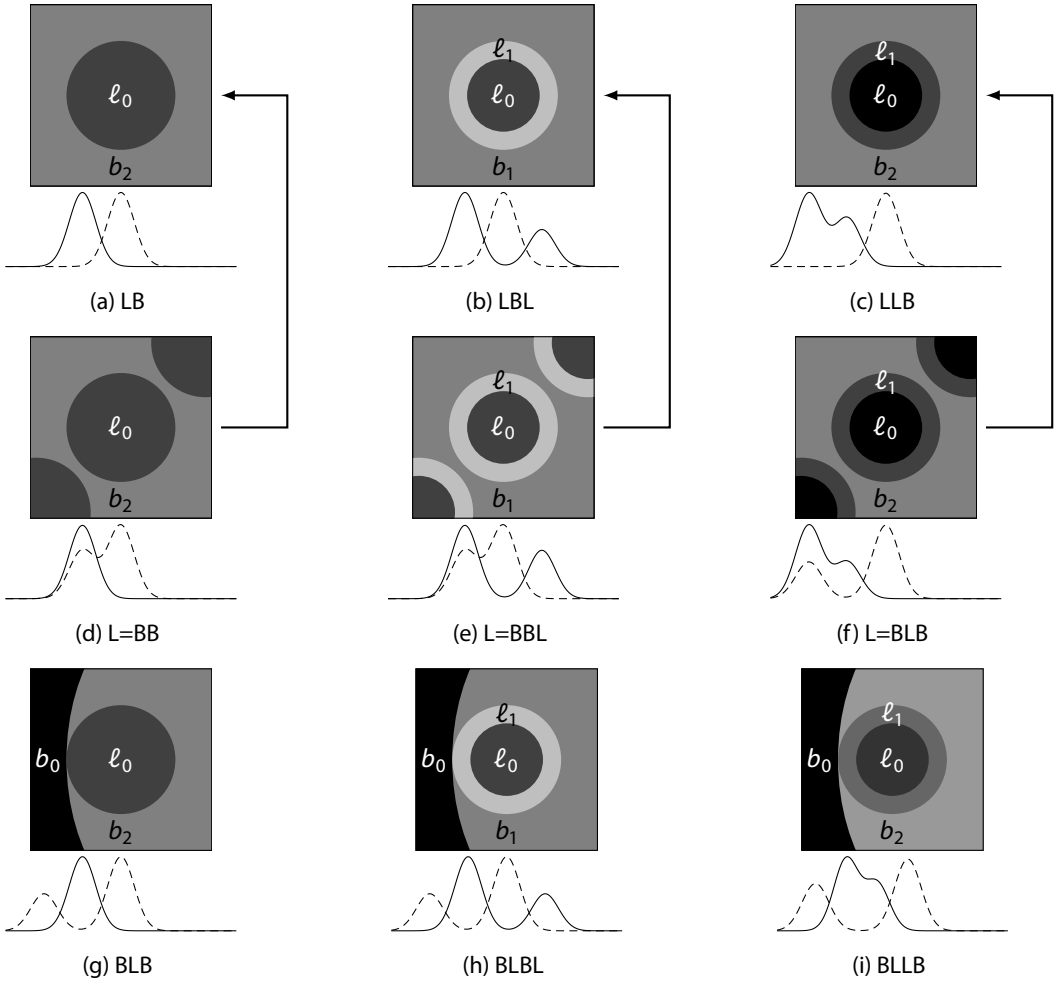


Figure 3.2: Models for possible ROI configurations of hypodense lesions with one or two lesion values and one or two background values. Stroke and ROI histograms are solid and dotted, respectively. (a) Simple hypodense lesion. (b) Hypodense lesion with enhanced rim. (c) Hypodense lesion with necrotic core. (d) Multiple hypodense lesions. (e) Multiple rim-enhanced lesions. (f) Multiple necrotic lesions. (g) Hypodense lesion with dark adjacent structure. (h) Rim-enhanced lesion with dark adjacent structure. (i) Necrotic lesion with dark adjacent structure.

almost all cases encountered in practice. Note that the model images in the figure only serve as illustrations; the actual analysis relies purely on the histograms.

In fact, only the peaks of the two histograms are considered. Each peak is a candidate for a typical lesion or background value, respectively. In practice, one or two values of each type are sufficient, although it is possible to construct cases with three or even more background values.

A single lesion value is used for homogeneous lesions where the density distribution is approximately normal (Figure 3.2a). Here and in the remainder of Figure 3.2, a shorthand notation is used to describe the configurations. For a simple hypodense lesion, *LB* denotes that there is one

lesion and one background value and that the lesion value is lower. A simple hyperdense lesion would be *BL*.

A pair of lesion values occurs if a lesion consists of two separate parts, a *core* and a *rim*, as shown in the middle and right columns of Figure 3.2. These are most often hypodense lesions with a hyperdense rim (*LBL*, Figure 3.2b) or with a necrotic, i.e. even more hypodense, core (*LLB*, Figure 3.2c). Lesions that do not follow this core-rim model may not be segmented properly by the algorithm, but they are rare among liver metastases. Obviously, it cannot be deduced from the histogram which of the two peaks belongs to the core and which one to the rim. Since segmentation has to start in the core, it is assumed that the value closer to the one at the stroke center represents the core.

A single background value is sufficient if the liver parenchyma is the only dominant structure in the ROI. However, there are cases where other structures create additional peaks in the ROI histogram, which might even be higher. Essentially, two situations can be distinguished. If a lesion is very large compared to the liver or if there are multiple lesions in close vicinity, the typical lesion value can also be represented by a peak of the ROI histogram. This means that there is a pair of a lesion value and a background value which are very close (denoted by *L*=*B*). Such a background value must not be used for threshold determination because it does not represent a structure that should be separated from the lesion. Therefore, background values are discarded if they do not keep a safety distance of δ to all lesion values. Figures 3.2d to 3.2f show the possible constellations and how they are reduced to cases with a single background value.

Unfortunately, a similar situation can occur when the stroke is drawn too long or when the lesion contains voxels with parenchyma density, often caused by partial volume effects. Then, the lesion value would have to be discarded, whereas the background value is correct. This occurs most often for very small lesions and short strokes, where the dilation will include too many background and partial volume voxels and the histogram may have spurious peaks. Therefore, the stroke histogram is only used if the stroke is longer than a threshold which was empirically optimized and set to 10 mm. Otherwise, the value at the stroke center is used as the typical lesion value.

If the lesion is close to the liver boundary, structures outside the liver can be represented in the ROI histogram as well and this can be helpful to prevent the segmentation from leaking out of the liver. This gives rise to the *BLB, BLBL* and *BLLB* types (Figures 3.2g to 3.2i). In these cases, the fixed lower threshold is replaced by the mean of the lower background value and the (lower) lesion value.

In theory, it is possible to have even more peaks in the histograms, but not all of them are relevant. Therefore, the following five values are extracted:

- ℓ_0: the leftmost peak of the stroke histogram,
- ℓ_1: the rightmost peak of the stroke histogram,
- b_0: the rightmost peak of the ROI histogram below $\ell_0 - \delta$, if any,
- b_1: the leftmost peak of the ROI histogram between $\ell_0 + \delta$ and $\ell_1 - \delta$, if any,
- b_2: the leftmost peak of the ROI histogram above $\ell_1 + \delta$, if any.

```
function AnalyzeStroke
    if stroke available and longer than 10 mm then
        Compute histogram of the ROI masked with the dilated stroke
        Smooth histogram using linear diffusion
        Compute peak positions from smoothed histogram
        ℓ0 ← leftmost peak
        ℓ1 ← rightmost peak
        s ← value at seed point (smoothed)
    else
        ℓ0 ← ℓ1 ← value at seed point (smoothed)
    return (ℓ0, ℓ1)
```

Algorithm 3.1: Stroke analysis

Note that the two ℓ values may be equal and that only one of the b values actually needs to be set. If no b can be found according to these rules, the highest peak is used regardless of its distance to the ℓ values. The histogram analysis is summarized in Algorithms 3.1 and 3.2.

The thresholds are now computed using these five values. If the lesion has both hypodense and hyperdense components, i.e. if b_1 is available, the value at the seed point is used to determine which ℓ value represents the core. Any available b values are used to separate the lesion from the background. Otherwise, the algorithm uses fixed thresholds and stroke quantiles as before. This is summarized in Algorithm 3.3.

Some examples of different ROI configurations in real images are shown for illustration in Figure 3.3.

3.3 Special ROI configurations

So far, the analysis of ROI configuration relies on the histogram peaks and removes peaks only if two of them are very close together and probably represent the same structure. In practice, of course, there are cases which do not fit to any of the models in Figure 3.2 or where so much noise is present that the histograms look different. Then, it can be reasonable to discard peaks or add additional values which do not correspond to a peak. The rules described in this section are driven by problems observed in practice. Although they annoy users when they occur, they are too rare to have a significant effect in the evaluation and therefore parameters could not be optimized formally. Furthermore, the extensions make stronger use of the stroke quantiles and may thus reduce reproducibility. Therefore, the following steps are highlighted in Algorithm 3.2 and may optionally be disabled.

Three types of problems that occurred several times in the data base are illustrated in Figure 3.4. In the first type (Figure 3.4a), a structure outside the liver creates a peak in the ROI histogram between lesion and parenchyma. It should not be used because then the threshold range for the lesion could get too small. A rule was added not to use a peak between $\tilde{\ell}_{0.1}$ and $\tilde{\ell}_{0.9}$ unless one of two conditions is fulfilled: either it is the maximum peak, i.e., it is probably the parenchyma and

function AnalyzeROI(ℓ_0, ℓ_1, δ)
  Compute histogram of the ROI
  Smooth histogram using linear diffusion
  Compute peak positions from smoothed histogram
  $b_{\max} \leftarrow$ position of highest peak
  $b_0 \leftarrow$ NaN; $b_1 \leftarrow$ NaN; $b_2 \leftarrow$ NaN
  for all peaks b, sorted by position, **do**
    if $b < \ell_0 - \delta \ \wedge (b = b_{\max} \vee b < \tilde{\ell}_{0.1})$ **then**
      $b_0 \leftarrow b$ ▷ b_0 is the rightmost peak below $\ell_0 - \delta$
    else if $b \in [\ell_0 + \delta, \ell_1 - \delta]$ **then**
      $b_1 \leftarrow b$ ▷ b_1 is the rightmost peak in $[\ell_0 + \delta, \ell_1 - \delta]$
    else if $b > \ell_1 + \delta \ \wedge (b = b_{\max} \vee b > \tilde{\ell}_{0.9})$ **then**
      $b_2 \leftarrow b$ ▷ b_2 is the leftmost peak below $\ell_0 - \delta$
      break
  if none of b_0, b_1, b_2 assigned **then**
    if $b_{\max} < \ell_0$ **then**
      $b_0 \leftarrow b_{\max}$
    else if $b_{\max} \in [\ell_0, \ell_1]$ **then**
      $b_1 \leftarrow b_{\max}$
    else if $b_{\max} > \ell_2$ **then**
      $b_2 \leftarrow b_{\max}$
  if $\ell_0 = \ell_1$ **then**
    $h \leftarrow \tilde{\ell}^{-1}(b)$
    if $b_0 \neq$ NaN $\wedge\ b_0 > \tilde{\ell}_{0.2}$ **then**
      $\ell_0 \leftarrow \tilde{\ell}_{\frac{h}{2}}$
      $b_1 \leftarrow b_0$; $b_0 \leftarrow$ NaN
    else if $b_2 \neq$ NaN $\wedge\ b_2 < \tilde{\ell}_{0.8}$ **then**
      $\ell_1 \leftarrow \tilde{\ell}_{\frac{h+1}{2}}$
      $b_1 \leftarrow b_2$; $b_2 \leftarrow$ NaN
  if $b1 =$ NaN $\wedge\ \ell_0 \neq \ell_1$ **then**
    if $b_0 \neq$ NaN $\wedge\ \ell_0 < b_0 + \delta \wedge |\ell_0 - s| < |\ell_1 - s|$ **then**
      $\ell_0 = \ell_1$
    if $b_2 \neq$ NaN $\wedge\ \ell_1 > b_2 - \delta \wedge |\ell_1 - s| < |\ell_0 - s|$ **then**
      $\ell_1 = \ell_0$
  return (b_0, b_1, b_2)

Algorithm 3.2: ROI analysis. The gray overlays indicate the handling for special ROI configurations from Section 3.3.

function ComputeThresholds($b_0, \ell_0, b_1, \ell_1, b_2, s$)
  if b_1 = NaN **then** ▷ homogeneous lesion, no matter whether it is hypodense or hyperdense
    if b_0 = NaN **then**
      $t_- \leftarrow \min(10\,\text{HU}, \tilde{\ell}_{0.1})$
    else
      $t_- \leftarrow \frac{1}{2}(\ell_0 + b_0)$
    if b_2 = NaN **then**
      $t_+ \leftarrow \max(180\,\text{HU}, \tilde{\ell}_{0.9})$
    else
      $t_+ \leftarrow \frac{1}{2}(\ell_1 + b_2)$
  else ▷ inhomogeneous lesion, s decides whether segmentation starts with ℓ_0 or ℓ_1
    if $s < b_1$ **then** ▷ start with hypodense part
      if b_0 = NaN **then**
        $t_- \leftarrow \min(10\,\text{HU}, \tilde{\ell}_{0.1})$
      else
        $t_- \leftarrow \frac{1}{2}(\ell_0 + b_0)$
      $t_+ \leftarrow \frac{1}{2}(\ell_0 + b_1)$
    else ▷ start with hyperdense part
      $t_- \leftarrow \frac{1}{2}(\ell_1 + b_1)$
      if b_2 = NaN **then**
        $t_+ \leftarrow \max(180\,\text{HU}, \tilde{\ell}_{0.9})$
      else
        $t_+ \leftarrow \frac{1}{2}(\ell_1 + b_2)$
  return (t_-, t_+)

Algorithm 3.3: Threshold determination

the contrast-to-noise ratio is just low; or it lies between $\ell_0 + \delta$ and $\ell_1 - \delta$, i.e., the lesion is probably inhomogeneous.

Figure 3.4b illustrates the opposite situation where a lesion value should be discarded. For small rim-enhancing lesions, the core is sometimes still slightly brighter than the parenchyma due to partial volume effects. If the lesion value contributed by the core were used, the threshold range would cover much of the parenchyma. Therefore, in such a case, the lesion value should be removed. Unfortunately, the histogram configuration is quite similar to the one in Figure 3.2c if gray value relations are inverted, so in general it will not be a good idea to discard the lesion value that is closer to the background value. The main difference here is that this lesion value belongs to the core. Therefore removing it will not be a problem because the core can be closed during the smart opening operation.

As a third problem, inhomogeneous lesions do not always have two peaks in the stroke histogram. Especially if there is a broad partial volume zone between the core and the rim, the stroke histogram can be very skewed but unimodal (Figure 3.4c). However, it is characteristic of these cases that a b value is "in the middle" of the stroke histogram. Formally, the ratio of stroke voxels less than b

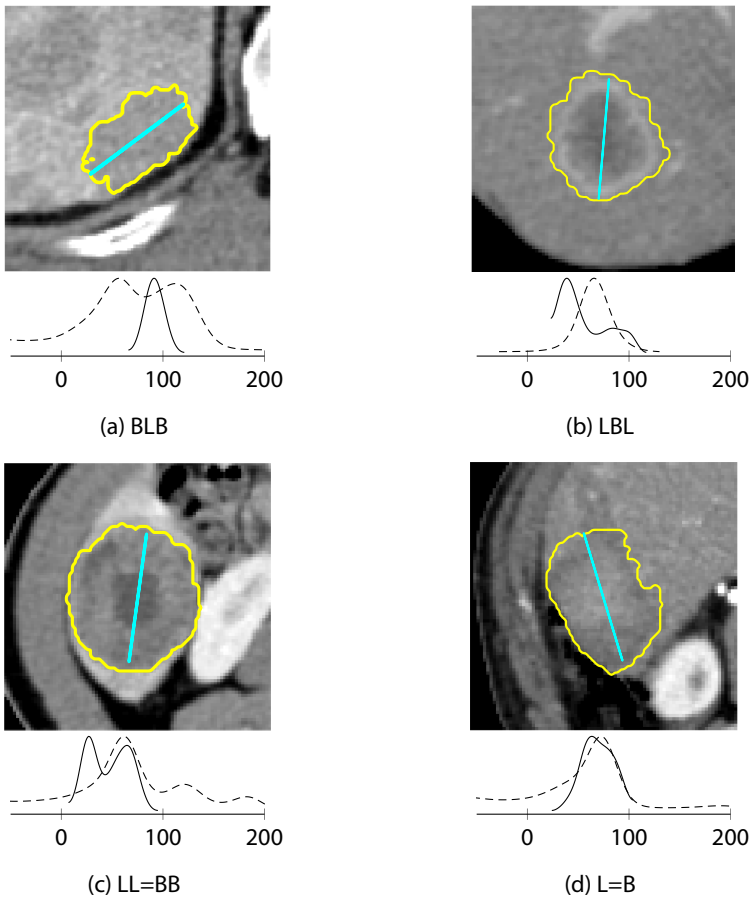


Figure 3.3: Examples of different ROI configurations, showing the central axial slices of the ROIs, the stroke histograms (solid) and the ROI histograms (dashed).

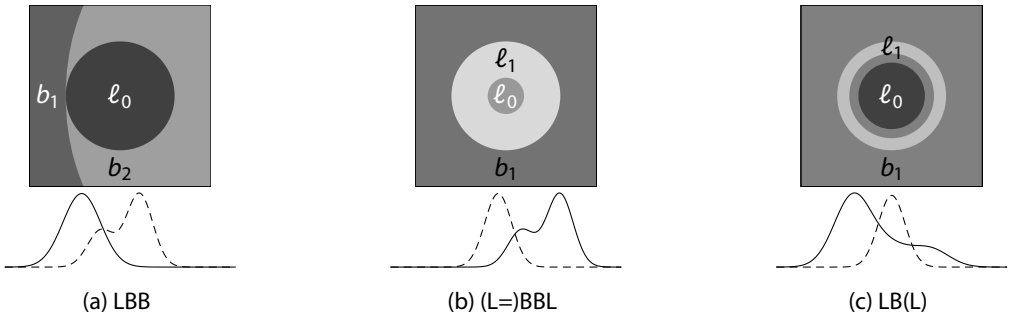


Figure 3.4: Models of some ROI configurations for which a special handling was implemented. (a) Hypodense lesion with medium dark adjacent structure. (b) Hyperdense lesion with small necrosis and partial volume effect. (c) Hypodense lesion with small enhanced rim and partial volume effect.

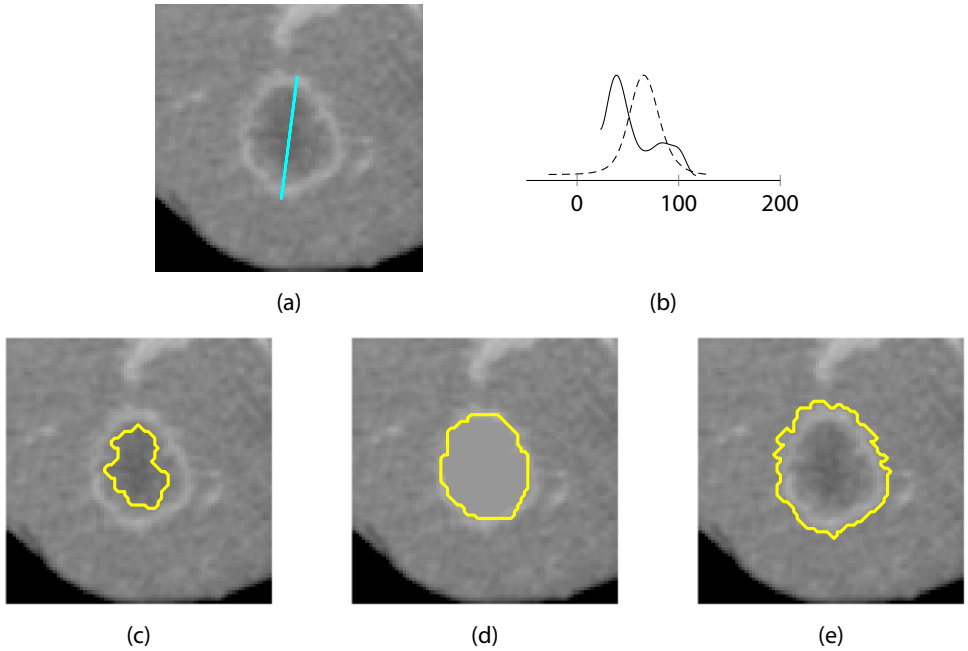


Figure 3.5: Step-by-step illustration of the segmentation algorithm for liver metastases with hyperdense rims. (a) Original ROI with stroke. (b) Histograms of the stroke (solid) and the ROI (dashed). (c) Segmentation of the hypodense core of the lesion. (d) ROI after filling the dilated core (in yellow) with a typical value of the hyperdense rim. (e) Final segmentation result.

is computed, giving the amount of hypodense voxels. It can expressed as $h = \tilde{\ell}^{-1}(b)$, using the CDF of the stroke as the inverse of the quantile function $\tilde{\ell}$. If h differs clearly from both 0 and 1, the lesion has both hypodense and hyperdense compartments and should be handled as if it had two lesion values. In practice, the threshold range for h was set $[0.2, 0.8]$. The additional lesion value is then a quantile of ℓ at the average of h and either 0 or 1, depending on whether h is less or greater than 0.5. After adding a lesion value, the background values have to be updated so that the relation $b_0 \leq \ell_0 \leq b_1 \leq \ell_1 \leq b_2$ is still fulfilled for all b values that are set.

3.4 Two-step segmentation of inhomogeneous lesions

For inhomogeneous lesions, only the core has been segmented so far. In order to add the rim, a second segmentation step is added. It is triggered by the presence of b_1, a background value between the two lesions values. The procedure is illustrated in Figure 3.5.

The basic idea is to fill the core with the typical value of the rim so that a "virtual" hyperdense lesion is created. This value is either ℓ_0 or ℓ_1, depending on which lesion value was used in the first step. Due to the partial volume effect, there is a narrow zone of voxels between the originally hypodense and hyperdense parts of the lesion that have a density similar to the parenchyma. Therefore the core mask is slightly dilated to bridge this zone, because otherwise the segmentation might not be able to reach the rim. The dilation strength depends on the longest diameter d_{core} of

function ComputeThresholds2($b_0, \ell_0, b_1, \ell_1, b_2$)
  if $b_1 \neq$ NaN **then** ▷ inhomogeneous lesion
    if $s < b_1$ **then** ▷ add hyperdense rim
      Fill mask with ℓ_1
      $t_- \leftarrow \frac{1}{2}(\ell_1 + b_1)$
      if b_2 = NaN **then**
        $t_+ \leftarrow \max(180\,\text{HU}, \tilde{\ell}_{0.9}, \ell_1)$
      else
        $t_+ \leftarrow \frac{1}{2}(\ell_1 + b_2)$
      $t_+ \leftarrow \frac{1}{2}(\ell_0 + b_1)$
    else ▷ add hypodense rim
      Fill mask with ℓ_2
      if b_0 = NaN **then**
        $t_- \leftarrow \min(10\,\text{HU}, \tilde{\ell}_{0.1}, \ell_0)$
      else
        $t_- \leftarrow \frac{1}{2}(\ell_2 + b_2)$
      $t_+ \leftarrow \frac{1}{2}(\ell_2 + b_1)$
  return (t_-, t_+)

Algorithm 3.4: Threshold determination for second segmentation step.

the core mask and the length of the stroke, which is the expected lesion diameter d_{lesion}. The width of the rim is $\frac{1}{2}(d_{\text{lesion}} - d_{\text{core}})$. This value, truncated to whole voxels, is the size of the dilation kernel, so that about half of the rim will be covered. Note that these voxels are not necessarily included in the final segmentation, because smart opening might still remove them. The only purpose of this step is to have them included in the threshold range. This is important because rims may have varying width or be incomplete. Also, in order to make sure that only the density range of the partial volume zone is affected, the dilation is only allowed to add voxels between ℓ_0 and ℓ_1. The thresholds for the second smart opening are computed in a similar way as before (Algorithm 3.4).

The fact that the special handling for inhomogeneous lesions is triggered by an analysis of the stroke histogram allows the user to decide whether the hyperdense part should be segmented without requiring any additional interaction. If the stroke is drawn across the hypodense part only, the postprocessing step is omitted. I found that often radiological expertise is needed to make this decision and that an automatic detection is not satisfying. This is also important because follow-up measurements have to be consistent and this cannot be guaranteed by a fully automatic solution.

It should be noted that this procedure is robust in cases where a rim was detected but is not actually present. In this case, the dilation strength will be zero, and due to the sophisticated threshold selection, the second segmentation causes no significant changes to the mask. Therefore the thresholds for the fraction of hypodense voxels h are not critical.

As mentioned earlier, there are other forms of inhomogeneous liver metastases that do not have such a regular core-rim structure. These cases are not handled correctly by the algorithm and demand a completely different approach that is outside the scope of this thesis.

3.5 Segmentation of peripheral liver metastases

The segmentation of lesions with contact to structures outside the liver constitutes a challenge because there is often no visual contrast. The algorithm presented so far has a tendency to either create a too small, roundish segmentation or to leak into the adjacent structure. Although the results are not necessarily bad in terms of volume overlap, they are visually unpleasing and sometimes obviously incorrect.

In many cases, this problem could be solved by computing a liver segmentation beforehand and incorporating the liver mask into the lesion segmentation. A global view on the liver and a shape model, as used by many modern liver segmentation algorithms, could help here. Unfortunately, this approach would take too much time. Also, there might still be cases where the liver segmentation fails and excludes a lesion from the liver mask.

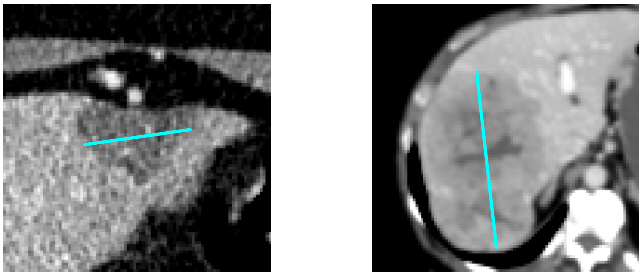
Therefore, I developed a method that estimates the liver boundary locally in the ROI. Similar to what Kuhnigk et al. (2006) do for juxtapleural lung nodules, the method makes use of the fact that the liver is at least locally convex in most parts. Now, lesions in the periphery of the liver can essentially be divided into three classes:

No leakage For lesions that lie adjacent to the lung, the heart or some other clearly contrasted structure, there is no risk of leakage. These lesions do not require a special handling, but of course the results should not be significantly deteriorated by a special handling for other cases. This refers to accuracy as well as efficiency (Figure 3.6a).

Within convex hull of liver If a tumor is not too large and not situated in an area where the liver contour has a very high curvature, it is possible to reconstruct the liver shape by computing a convex hull of the parenchyma contained in the ROI. It is crucial to make sure that the convex hull covers the lesion completely. Otherwise the segmentation will cut right through a lesion. This is worse than leakage because it is harder to understand for a user, especially in places where the lesion is clearly delineated (Figure 3.7a).

Beyond convex hull of liver For large lesions, the convex hull approach does not work, because it is not able to reconstruct the curvature of the liver boundary. Looking at an example (Figure 3.6b), it becomes clear that it is very hard or impossible to estimate the liver boundary without using a global shape model. For these cases, however, slight leakage is less of a problem because it does not have a large impact on the measured volume. They are also perceived as "difficult" cases by most users and segmentation problems are more likely to be tolerated.

These considerations motivate the following approach. First, a lesion is assigned to one of the three classes. Second, if the convex hull approach is necessary and feasible, it is applied; otherwise, no special handling is performed. The two steps will be described in reverse order, because it is easier to understand the classifier once the method with its strengths and weaknesses has been introduced.



(a) (b)

Figure 3.6: (a) Peripheral lesion with no risk of leakage. (b) Peripheral lesion that protrudes from the convex hull of the liver parenchyma.

Similar to what was said in Section 3.3, peripheral liver lesions are not frequent enough to perform a formal parameter optimization. Therefore the method was mostly designed by visual inspection of results on cases where no reference segmentation was available.

3.5.1 Local liver boundary estimation

The liver boundary estimation starts with a coarse segmentation of the liver parenchyma. If a single b value was set during histogram analysis, this is supposed to be a typical parenchyma value. Otherwise a heuristics is used: If b_1 is given, it is most probably the parenchyma. If both b_0 and b_2 are given and ℓ_0 is less than the ROI median, b_2 is used, otherwise b_0. Let the chosen value be p.

An initial parenchyma mask is created by thresholding in a range of 20 HU around p (Figure 3.7b). Since this mask can include other structures outside the liver, it is morphologically opened, the largest connected component is chosen, and it is closed again (Figure 3.7c). Additionally, to exclude the ribs, which often have values in the parenchyma range around a very bright center, a bone mask is subtracted from the parenchyma mask. The bone mask is computed by thresholding above 200 HU and dilating it.

Now the convex hull of the parenchyma is constructed. Since doing this in 3D is very expensive, the 2D convex hulls on all axial, sagittal, and coronal slices are computed instead, and the union of all voxels contained in any of these convex hulls is used as an approximation (Figure 3.7d). This is a subset of the actual 3D convex hull and thus not necessarily convex, but it proved to be sufficient for the purpose at hand. The result of this procedure is used as a mask for smart opening (Figure 3.7e).

The procedure is summarized in the non-overlaid part of Algorithm 3.5.

3.5.2 Classification

The classifier that decides whether the boundary estimation should be used consists of several criteria that are evaluated at different points during the computation. If any of these criteria is met, the computation is canceled and no liver mask is used. This approach was chosen to avoid unnecessary computations.

In summary, these criteria formalize the following observations:

function EstimateLiverBoundary(t_-, t_+, p)
  if stroke length > 60 mm **then**
    return
  Compute lesion mask L by region growing from the stroke center with thresholds $[t_-, t_+]$
  if result does not touch ROI boundary **then**
    return ▷ No leakage suspected
  Compute Euclidean distance transform D of L
  $t_{\text{max}} \leftarrow$ maximum value of D under dilated stroke
  $t_{\text{con}} \leftarrow$ connection threshold from position of t_{max} to ROI boundary on D
  if $\frac{t_{\text{con}}}{t_{\text{max}}} < 0.25$ **then**
    return ▷ Estimate of necessary erosion strength
  Compute bone mask B by thresholding with $[\max(200, p + 20), \infty]$
  Dilate B with a $5 \times 5 \times 5$ kernel
  if $\frac{1}{2}(t_- + t_+) < p$ **then**
    Compute parenchyma mask P by thresholding with $[\max(p - 20, t_-), p + 20]$
  else
    Compute parenchyma mask P by thresholding with $[p - 20, \min(p + 20, t_+)]$
  Dilate P with a $3 \times 3 \times 3$ kernel, select the largest connected component,
    and erode with a $3 \times 3 \times 3$ kernel
  Compute convex hull C of $P - B$
  Compute an ellipsoid approximation E of L
  if (fraction of rays used for ellipsoid approximation < 0.5) ∨
    (maximum curvature of C in E > 15) ∨
    (coverage of E by C < 0.6) ∨
    (stroke center not covered by C) **then**
    return
  return C

Algorithm 3.5: Liver boundary estimation. The gray overlays indicate the classification.

- A special handling is not necessary if the lesion is not connected to the liver boundary. This is tested by a preliminary region growing and an estimation of the erosion strength that will be used by smart opening. Appyling the liver boundary estimation in such a case usually would not change the result, but increase computation time.
- For large lesions, the liver would mostly be underestimated and the lesion would not be segmented completely. Also, the increase in computation time is most noticeable here.
- If the approximated boundary of a lesion has too little contact to the liver parenchyma, it probably protrudes from the liver and will not be included completely in the convex hull.
- If the convex hull does not cover the approximated lesion sufficiently or has a high curvature, it should not be used.

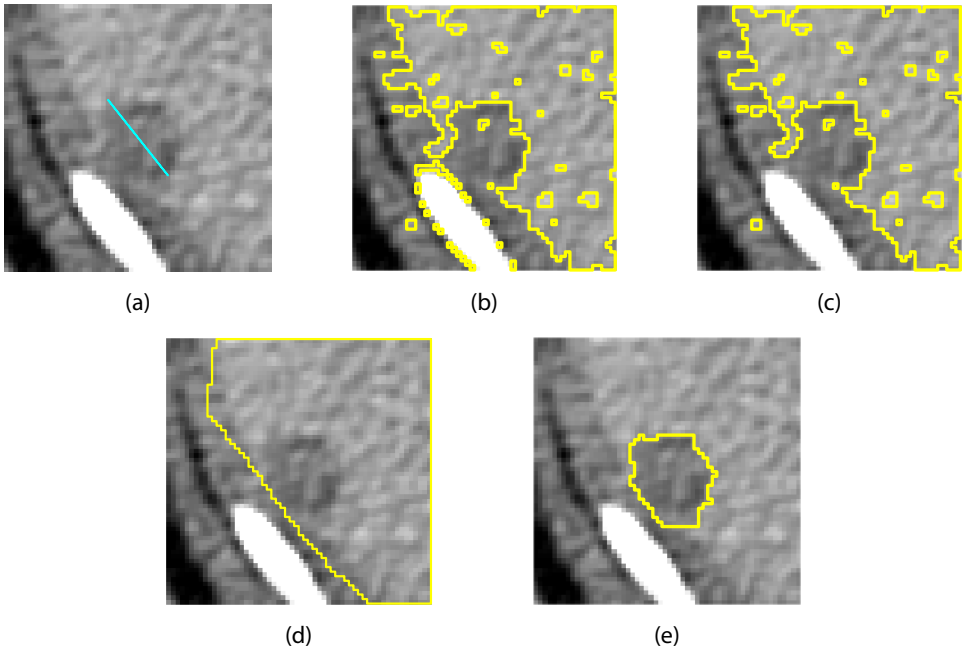


Figure 3.7: Step-by-step illustration of the segmentation algorithm for peripheral liver metastases. (a) Original ROI with stroke. (b) Coarse parenchyma mask obtained by thresholding, including some rib voxels. (c) Largest connected component of opened parenchyma mask. (d) Approximate convex hull of the parenchyma. (e) Final segmentation result.

Note that some of these criteria can be evaluated before actually computing the liver boundary estimation, whereas others check the plausibility of the result once it is available.

The decision tree that was implemented for classification is shown in the overlaid parts of Algorithm 3.5. The classifier was trained by visually comparing results with and without liver boundary estimation on a special data set of peripheral lesions. This approach was favored over an automatic training because for these data no manual segmentations were available and because the visual impression was supposed to be an important factor.

Chapter 4

Evaluation

4.1 Technical evaluation

4.1.1 Parameter optimization

The algorithm described in the previous chapter has several parameters. Some of them were formally optimized on the development data with respect to manual segmentations, maximizing the median volume overlap. The optimization was performed on the development data only (82 lesions), so the results reported for the test data are independent. An overview of the experiments is shown in Table 4.1.

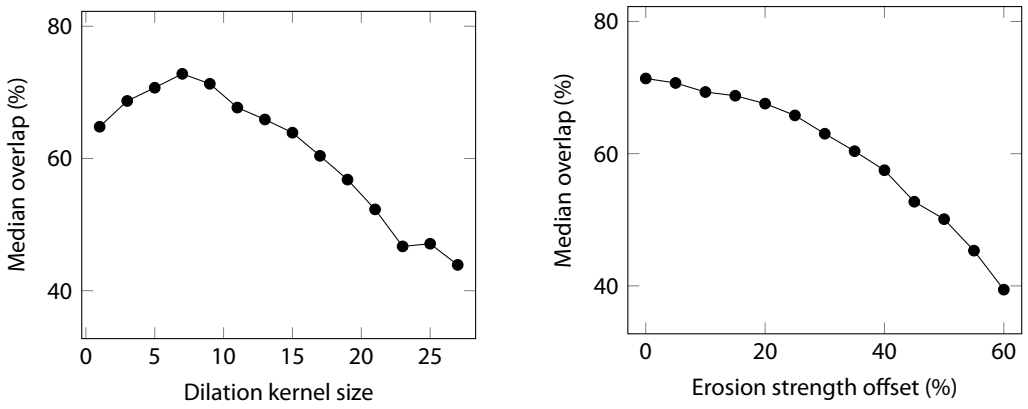
It turned out that the algorithm is not very sensitive to changes of any of these parameters. The optimal achieved overlap on the development data is 71.2 % and for some parameters it hardly dropped below 70 % when values were varied within a reasonable range. Two plots are shown in Figure 4.1 for illustration. The first example (Figure 4.1a) shows how the dilation of the stroke for the histogram computation affects the results. If the dilation kernel is too small, there may not be enough values for doing statistics, and if it is too large, it will also cover background values.

For smart opening, the optimal parametrization differs from what Kuhnigk et al. (2006) reported for lung nodules. For example, an erosion strength offset is not necessary according to my tests (Figure 4.1b), whereas 25 % were proposed in the lungs.

While an experimental optimization makes sense for the parameters of smart opening and some internal parameters of the threshold selection, other parameters were set manually for several

Parameter	Optimal value	Tested range
Threshold selection:		
ROI smoothing kernel size	5	$1, 3, \ldots, 15$
Stroke histogram smoothing iterations	10, 15	$0, 5, \ldots, 30$
ROI histogram smoothing iterations	10	$0, 5, \ldots, 30$
Stroke dilation kernel size	7	$1, 3, \ldots, 15$
Stroke extension quantile (%)	10	$0, 5, \ldots, 20$
Smart opening (see Kuhnigk et al. (2006)):		
Preliminary erosion strength (%)	30	$0, 5, \ldots, 50$
Minimum erosion strength ε (%)	15	$0, 5, \ldots, 40$
Erosion strength offset μ (%)	0	$0, 5, \ldots, 40$

Table 4.1: Parameter optimization overview.



(a) Stroke dilation

(b) Erosion strength offset

Figure 4.1: Two exemplary parameter optimization plots.

	Development	Test
Volume overlap (%)	71.2	62.8
Volume error (%)	14.1	19.8
Diameter error (%)	6.1	8.3
Average surface distance (mm)	1.11	0.83
Hausdorff distance (mm)	6.15	4.47
MICCAI score	78.8	79.4

Table 4.2: Medians of various metrics for accuracy evaluation on the development and test data sets.

reasons. The parameters of the special parts of the algorithm for inhomogeneous and peripheral lesions could not be optimized formally because not enough cases with manual segmentations were available for training. Instead, they were determined by visual inspection of the results. Some other parameters are used to balance accuracy and reproducibility and their current values are the result of long-term observations and user feedback.

4.1.2 Accuracy

For evaluating the accuracy of the algorithm, the test data were segmented and compared to the reference segmentation using several metrics. The strokes were generated automatically by computing the end points of the longest axial diameter of the reference segmentation passing through its center of gravity. The results are summarized in Table 4.2, giving the median values for both the development and test data. It can be seen that the performance is roughly the same for both data sets. Volume-related metrics are slightly higher for the development data, while the test data performed better in terms of distance-based metrics. The median MICCAI score, which averages both kinds of metrics, is almost equal.

A more detailed discussion about the MICCAI score, including definitions of the incorporated metrics, will be given in Section 9.2. Here, it is sufficient to know that a score of 90 is meant

to correspond to the quality of a manual segmentation and that the best result achieved by a semi-automatic method at the Liver Tumor Segmentation Challenge 2008 was 69.4. This is a mean score and should thus be related to a mean of 74.7 achieved by the proposed algorithm on the test data, although results on different data cannot be directly compared. A previous version of the algorithm participated in the Challenge and got a score of 69.1.

For a deeper analysis, box plots for all six metrics are shown in Figure 4.2. The whiskers show the 10 % and 90 % quantiles, so the quality achieved for the best 90 % cases can easily be seen in the plots. All values outside this range are additionally shown so that the worst-case behavior of the algorithm is also documented.

In four cases (of totally 453 in both development and test data), the algorithm was not able to compute a result. They are shown in Figure 4.3 along with the strokes. In all cases, the problem is the low contrast-to-noise ratio, either due to strong noise or low contrast. This causes the thresholding result to be so fragmented that all voxels are removed by the erosion in smart opening.

Figure 4.4 illustrates some bad segmentation results with a volume overlap below 25 %. The problems are mainly caused by unsuitable thresholds. The first lesion is too inhomogeneous and its inner density is too close to that of the parenchyma, although it has a clear boundary (Figure 4.4a). In the second case, the liver parenchyma value is estimated incorrectly because it covers only a small portion of the ROI and the contrast to the lesion is relatively low (Figure 4.4b). The two other lesions are very small and the images are noisy, which can lead to a segmentation that is either too small (Figure 4.4c) or too large (Figure 4.4d).

Figure 4.5 is a collage of successful segmentations. It includes cases with inhomogeneities, enhancing rims or cores, contact to the liver boundary or structures of similar density, low contrast and very large lesions.

Some cases that are improved by the special ROI configuration handling can be seen in Figure 4.6. In the first example, a second lesion value is added, since the highest background value is clearly within the range of the stroke voxels. The complete lesion can then be segmented in two steps (Figures 4.6a to 4.6c). In the second case, the lower background value is removed because it is in the stroke range, thus avoiding a too narrow threshold range (Figures 4.6d to 4.6f).

Figure 4.7 illustrates the benefit of the local liver boundary estimation. Both cases have contact to the intercostal muscles, which have similar density as the lesion. In Figure 4.7a, smart opening tries to avoid leakage to the ROI boundary and therefore sets the erosion strength so high that the final result is too small. In Figure 4.7c, on the other hand, the segmentation leaks into the muscles because smart opening is not able to make a separation. Both problems are solved by applying an approximate liver mask (Figures 4.7b and 4.7d).

4.1.3 Reproducibility

The reproducibility of the algorithm was tested by comparing the volumes of the segmentation results after initialization with various strokes. The strokes were generated automatically by the following procedure. First, a random point on the unit sphere is generated, using the method by Shao and Badler (1996). The vector from the center of gravity of the reference segmentation to this random point defines the orientation of the stroke, which is then cropped at the boundaries of the segmentation. For each lesion, ten strokes are generated and the segmentations are computed. The reproducibility is measured in terms of the coefficient of variation (COV) of the volume.

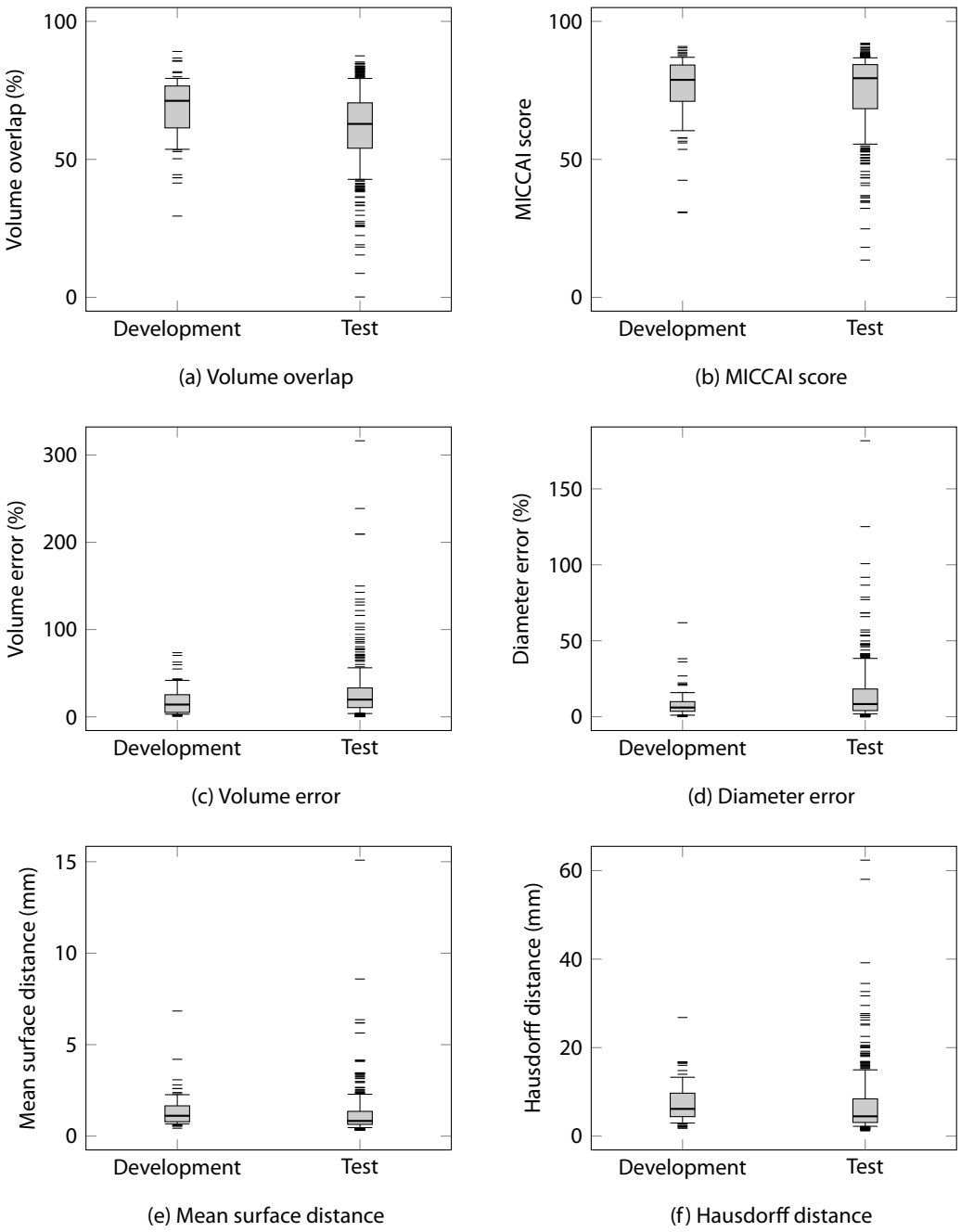
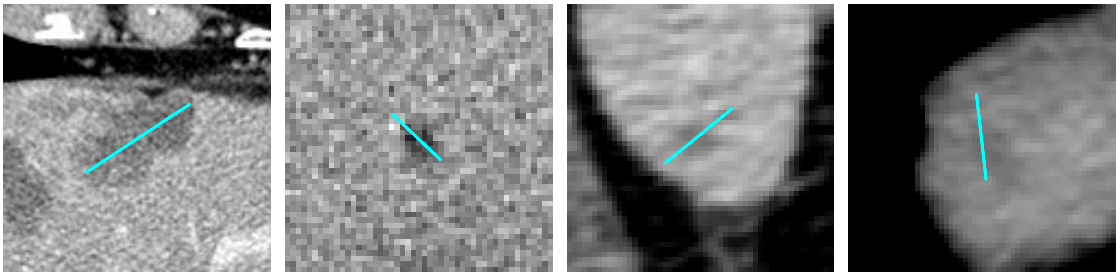
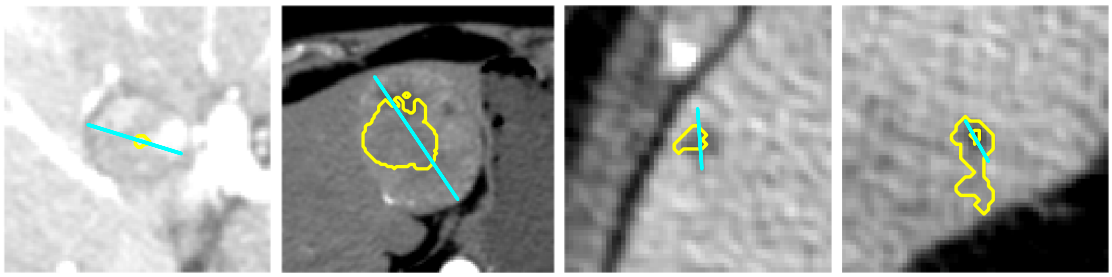


Figure 4.2: Box plots of six metrics for accuracy evaluation. The whiskers show the 10 % and 90 % quantiles, respectively.



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

Figure 4.3: The four cases where the segmentation algorithm did not return a result.



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

Figure 4.4: Examples of bad segmentation results (volume overlap < 25 %).

Figure 4.8a shows a box plot of the results. The median COV is 9 % on the test data. It is less than 37.3 % in 90 % of the cases.

Additionally, it was tested how the reproducibility is affected when strokes are slightly too short or too long or do not pass through the center of the lesion. For this, two stroke offset factors and three center offset factors were randomly generated and multiplied with the diameter of the reference segmentation. Both offset factors were drawn from a uniform distribution with varying limits from ±5 % up to ±20 %. The stroke offsets are used to lengthen or shorten the stroke at both endpoints independently, so the stroke length may be changed by twice the offset factor. The center offsets are added to the center point, separately for the x, y and z coordinates. Again, the COV of the volume for ten strokes is computed.

The results of this experiment are summarized in Figure 4.8b. It can be seen that a stroke that does not pass through the center of gravity is typically not a problem. If the stroke, however, is too long or too short, this has a stronger effect. A deviation of up to 5 % of the stroke length is not a problem, and a median COV of 20 % for an offset factor of 15 % is still tolerable. Above this, however, the COV rises quickly, indicating that segmentation fails for some strokes in an increasing number of cases.

In addition to these quantitative results, Figure 4.9 shows two exemplary cases, each with an offset factor of 0 and 20 %. A typical case with an average COV can be seen in Figure 4.9a. In the same case, stroke variations cause leakage and one segmentation failure (Figure 4.9b). The second example illustrates a case that is quite sensitive to stroke variations due its internal inhomogeneity

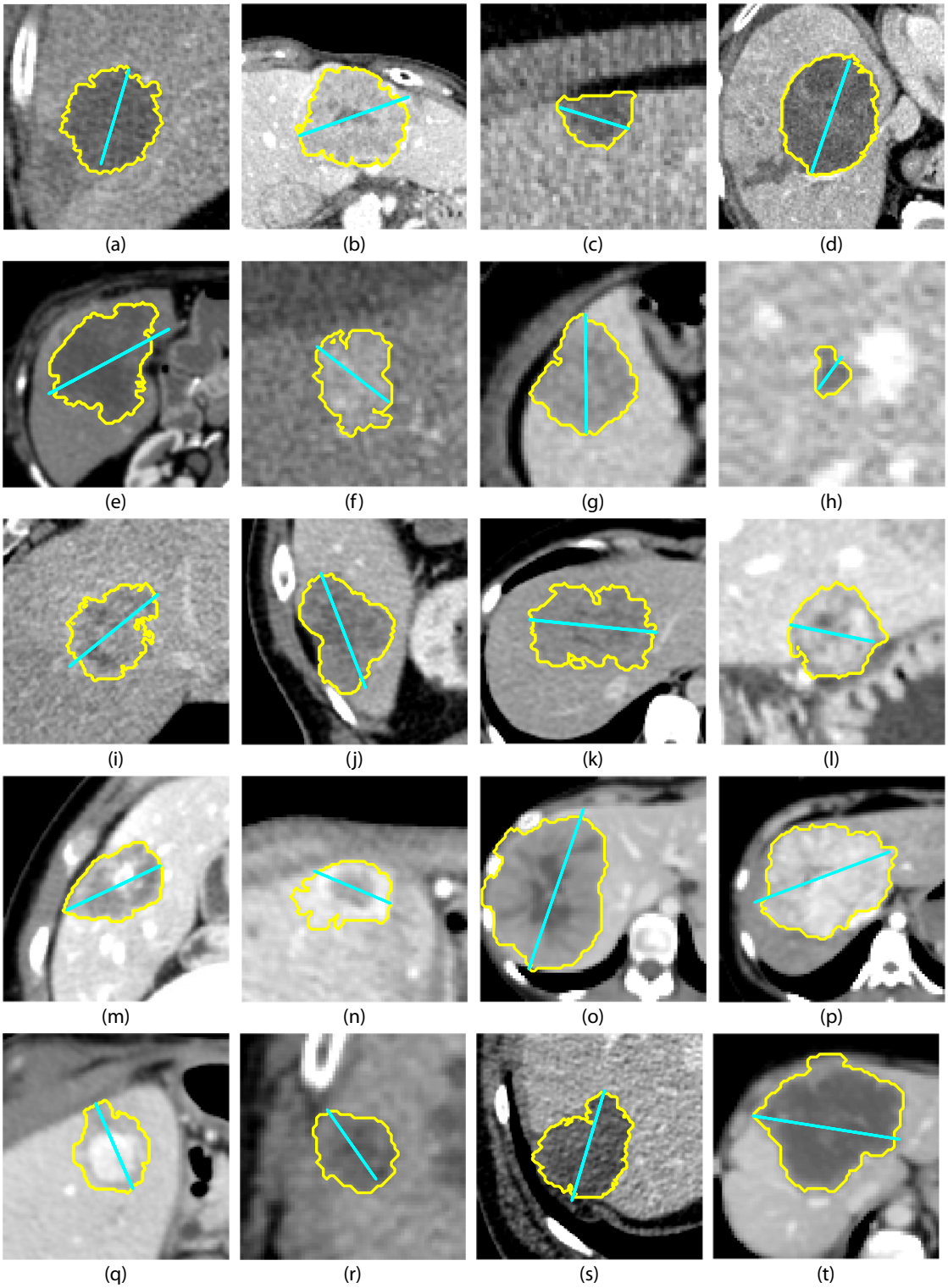


Figure 4.5: Examples of successful segmentation.

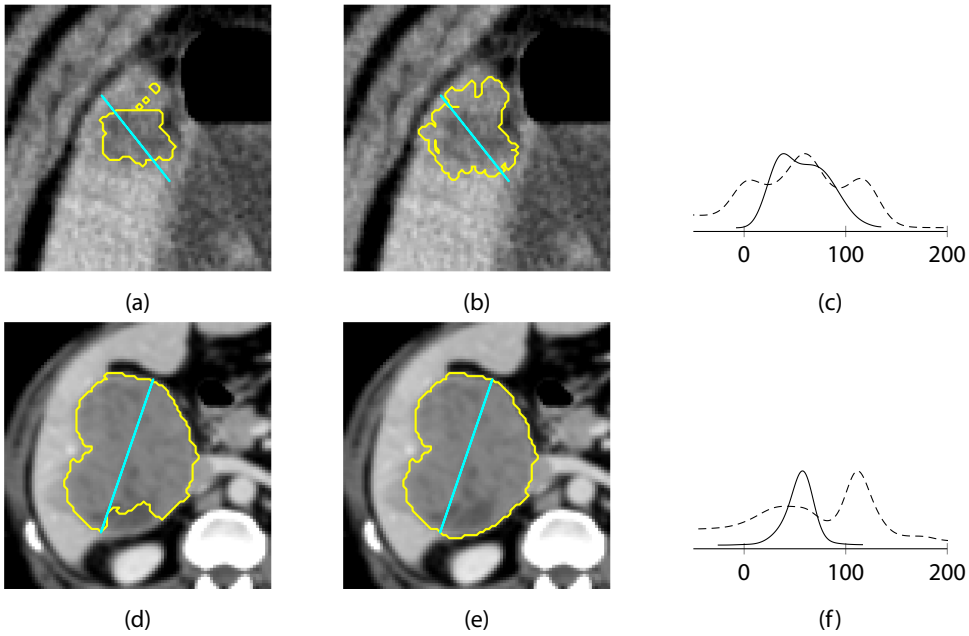


Figure 4.6: Examples of improved segmentation by the special ROI configuration handling.

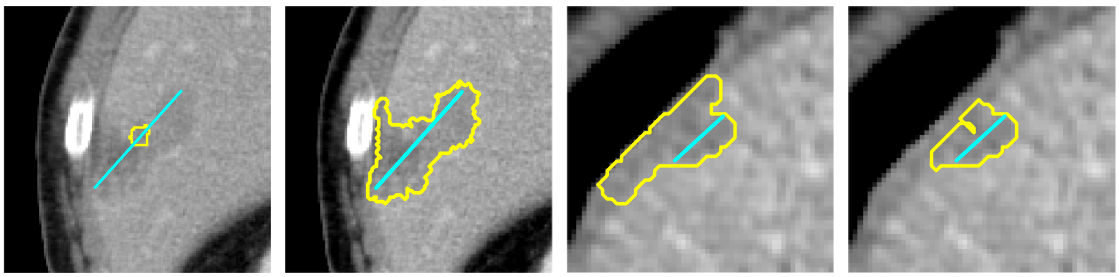
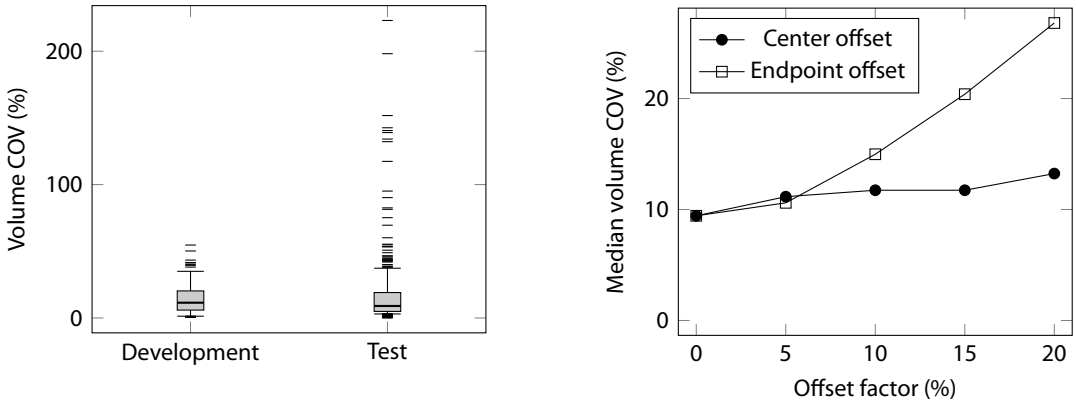


Figure 4.7: Examples of improved segmentation by the local liver boundary estimation.

even if there is no offset. As shown in Figure 4.9c, the dark core is always segmented, but the slightly brighter areas need to be covered by the stroke. If the stroke is too long, however, this can result in extensive leakage (Figure 4.9d).

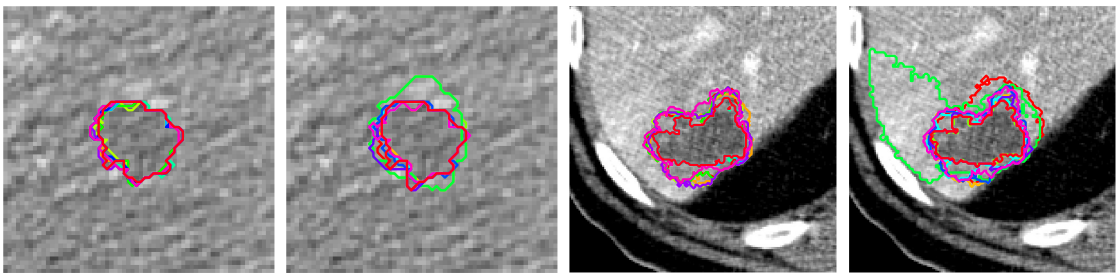
4.1.4 Efficiency

As discussed earlier, efficiency is a paramount criterion for clinical acceptance and the algorithm was designed such as to fulfill these requirements. The computation times for segmenting the development and test data are shown as a box plot in Figure 4.10. They were measured on a state-of-the-art PC with a 1.73 GHz QuadCore processor and 16 GB RAM. The median is 0.75 s, the 90 % quantile 2.9 s. The maximum is 13.6 s, but as the boxplot shows, this is exceptional. Cases with relatively long computation time were typically large or required two segmentation steps.



(a) (b)

Figure 4.8: Evaluation of reproducibility. (a) Box plot of volume COV for stroke variation without offset. (b) Median volume COV for stroke variation with varying offset factors.



(a) COV = 8.9 % (b) COV = 52.2 % (c) COV = 38.1 % (d) COV = 83.0 %

Figure 4.9: Example results of the reproducibility test. (a), (c) No offset. (b), (d) Stroke offset 20 %.

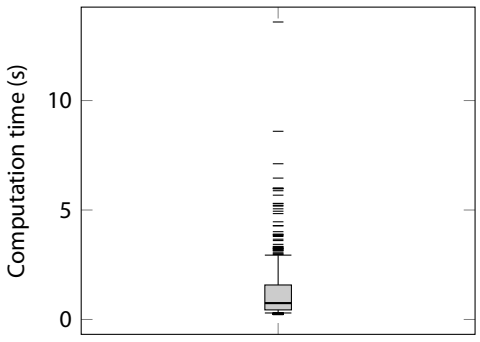


Figure 4.10: Box plot for computation time (development and test data).

	Algorithm vs. reference	Reference vs. reference
Volume overlap (%)	64.3	67.7
Volume error (%)	24.5	20.9
Average surface distance (mm)	0.97	0.80
Hausdorff distance (mm)	3.84	3.23

Table 4.3: Pairwise comparison of algorithm results and reference segmentations. The values show the averages of all pairs and the medians over all 50 cases.

4.2 Evaluation with multiple reference segmentations

So far, the algorithm was evaluated by comparing results to a single reference segmentation on 453 cases. From this data, 50 cases were selected randomly for each of which two additional reference segmentations were created by experienced radiology technicians. The lesions originate from 38 CT scans from several hospitals and CT scanners that were acquired for liver surgery planning. They can be considered as a representative collection of segmentable liver tumors. Very small, very inhomogeneous, and not clearly delimitable lesions had been excluded in advance.

As an initial experiment, I compared the algorithmic result to each of the three reference segmentations in terms of volume overlap and Hausdorff distance. The results for all individual cases are shown in Figure 4.11. It can be seen that the results may differ considerably depending on which reference is used. The median difference between the best and worst value per lesion is 8 percentage points for the overlap and 0.84 mm for the Hausdorff distance, with a maximum difference of 37 percentage points and 14.14 mm, respectively. These extreme cases are displayed in Figure 4.12, along with the two cases where the difference was lowest. Note how in Figure 4.12b the segmentations are actually quite different, although the Hausdorff distances are almost equal.

Two further important observations can be made here. First, the variability itself of the validation results is different across cases. Although there is always *some* degree of variability, it seems to depend on characteristics of the individual cases. These can be *anatomical factors* such as size, position or the tumor entity and *imaging-related factors* like noise or resolution. Second, the choice of the reference segmentation does not have a great effect on the *overall* quality assessment. The average volume overlap is between 60.3 % and 62.1 % and the average Hausdorff distance between 6.0 mm and 6.5 mm.

So the benefit of using multiple reference segmentations can be twofold. It makes the results for *individual* cases more reliable, for example when actual problem cases of an algorithm should be detected, and it can help to interpret the results. It is hard to tell whether a volume overlap of 60 % is good or not, but knowing the overlap between two manual segmentations makes this clearer. So, the inter-reference variability can be used to calibrate a measure of algorithm quality. This was done in Table 4.3, which shows that the results for the algorithm are actually quite close to the inter-reference variability.

These results were the initial motivation for Part III of this thesis, which investigates the problems of validating segmentation algorithms in more depth.

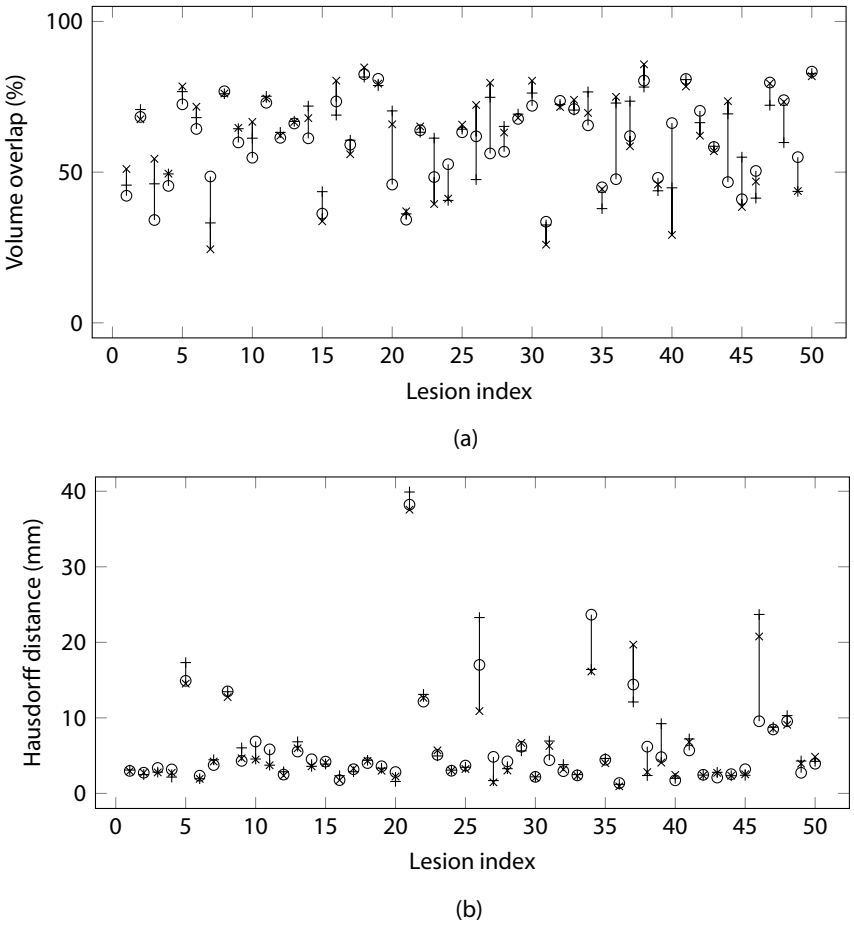


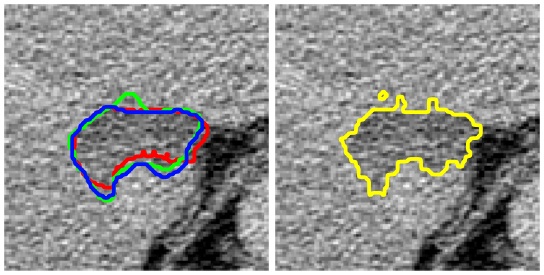
Figure 4.11: Volume overlap and Hausdorff distance of the 50 liver tumors in the study, using each of the three references as the "ground truth". It can be seen that results may differ significantly depending on which reference is used.

4.3 Clinical evaluation

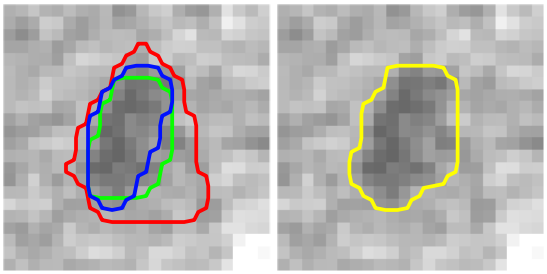
The algorithm presented above has been evaluated by our clinical partners in several studies. An overview of the related publications is given in Table 4.4, showing which manual and semi-automatic measurements were compared. In some cases, a particular imaging parameter was varied and the robustness of the results under different imaging conditions was analyzed. The agreement between two measurements is often determined in terms of the *concordance correlation coefficient* (CCC). It is computed from the means, variances and covariances of two samples x and y as

$$\frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}. \tag{4.1}$$

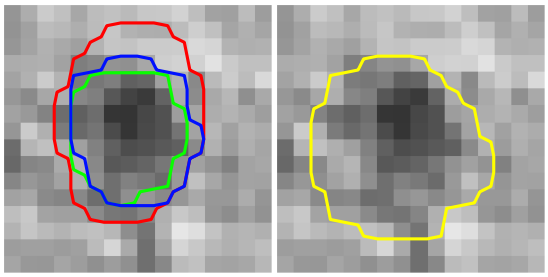
This metric can compare the volumes of two segmentation results, but not their actual shapes. Its range is $[-1, 1]$, where 1 indicates perfect agreement.



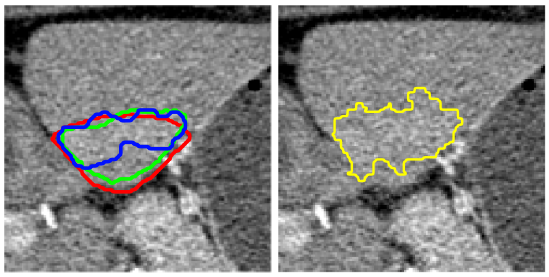
(a) Volume overlap: 66.1 %, 66.7 %, 66.8 %



(b) Hausdorff distance: 2.5 mm, 2.4 mm, 2.4 mm



(c) Volume overlap: 66.3 %, 29.2 %, 44.8 %



(d) Hausdorff distance: 9.6 mm, 20.8 mm, 23.7 mm

Figure 4.12: Cases where validation results differed (a), (b) least and (c), (d) most depending on the chosen reference segmentation. Left: three reference segmentations, right: algorithmic segmentation.

Publication	# lesions	Manual RECIST	Manual WHO	Manual volume	Semi-automatic RECIST	Semi-automatic WHO	Semi-automatic volume	Remarks
Heußel et al. (2007)	2 · 198				•	•	•	
Keil et al. (2008)	165	•	•		•	•		
Keil et al. (2009)	14				•		•	phantom
Puesken et al. (2009)	94	•			•		•	varying contrast phase
Keil et al. (2010b)	2 · 50			•			•	pre and post RFA
Keil et al. (2010a)	79				•		•	varying dose
Puesken et al. (2011)	106				•	•	•	varying slice thickness, 2 readers
Wulff et al. (2012)	2 · 77	•					•	3 readers

Table 4.4: Summary of clinical publications using the presented liver tumor segmentation algorithm. If data at two timepoints were used, this is indicated by "2 · *x*".

The accuracy of the algorithm was examined using a hardware phantom by Keil et al. (2009). The phantom contained 14 lesions in various sizes, densities, and orientations and was scanned with different tube currents, reconstruction kernels, and slice thicknesses. Under a standard protocol (165 mAs_{eff}, Siemens B30f, 3 mm), the mean volume error was 6.93 %. For almost all tested imaging settings, the CCC for the volume was 1.00.

In two studies, the agreement between manual and semi-automatic results was determined. Keil et al. (2008) measured CCCs of 0.94 and 0.95 for RECIST and WHO measurements, respectively. For volume measurements in lesions pre and post RFA, Keil et al. (2010b) determined a CCC of 0.99. The lesions were actually segmented manually in this study and the concordance seems to be remarkable. Unfortunately, no other metrics were computed to compare the segmentation results directly.

The effect of imaging parameters was also examined on clinical data. Puesken et al. (2009) compared measurements in arterial and portal venous phase for hypovascularized, hypervascularized and liquid lesions. They found a high agreement between manual and semi-automatic RECIST diameters in most groups (CCC 0.75 to 0.99), the only exception being hypervascularized lesions in the portal venous phase where a significant underestimation was observed (CCC 0.61). This result can in part be explained by the fact that segmentation results were not edited and at that time the algorithm still often missed the contrast-enhanced rim of a lesion.

Results on normal-dose and simulated low-dose scans were compared by Keil et al. (2010a). They did not find a significant difference in either RECIST diameter or volume for any low dose compared to normal dose (160 mAs_{eff}), except for volume at the lowest dose (40 mAs_{eff}). Using the normal dose measurements as actual reference values, however, stronger results might have been possible in this study.

Puesken et al. (2011) varied the slice thickness and accepted between 73 % (5 mm) and 85 % (1.5 mm) of the segmentations without editing. But even after editing, volumetric results on 5 mm showed limits of agreement of $[-58\,\%, 73\,\%]$, compared to 1.5 mm slices, which is used as the reference. These values are so close to the RECIST tresholds converted to volume $[-65.7\,\%, 72.8\,\%]$ that the authors recommend a maximum slice thickness of 3 mm for semi-automatic measurements. The interobserver difference, however, did not increase significantly at higher slice thicknesses.

Another study focusing on reproducibility of measurements was published by Wulff et al. (2012). Here, variation coefficients (VC) of the *diameter changes* measured by three readers are given. The median VC is 12.8 % for manual RECIST and 8.2 % for semi-automatic effective diameters. In this study, semi-automatic measurements were rated as "good" on a three-point scale for 84 % of the lesions. For only 75 %, however, the initial result was accepted without manual editing. This is interesting in two respects. First, the authors call it a "remarkable finding" that their satisfaction with liver lesion segmentation was almost as high as with lung nodule segmentation and attribute this to recent algorithm improvements. Second, the fact that some "good" segmentations were still edited raises the question whether necessity of correction is really an adequate quality measure. In a study with lymph nodes, Weßling et al. (2012) made a similar observation and suspect an "overuse bias" due to the "advancing refinement of correction tools." This means that users may tend to edit segmentation results although it is not strictly necessary in terms of volumetry.

One of the earliest studies concerning the impact of volumetry on response evaluation was performed by Heußel et al. (2007). They do not comment on the segmentation quality, but state that 13 % of the patients were classified differently by volume than by RECIST or WHO. They

conclude that diameter and area are not always adequate approximations of the volume, which should therefore be the parameter of choice.

In summary, the clinical studies showed that the algorithm gives satisfying results in 75 to 85 % of the cases, is robust under variations in dose, kernel and slice thickness (up to 3 mm) and has a better reproducibility than manual measurements.

4.4 Discussion

The goal of this part was to develop an algorithm for liver lesion segmentation that is fast enough for clinical use and provides accurate and reproducible results on cases that are relevant in practice. In order to keep the constraints in computation time, a relatively simple but powerful algorithmic concept was used, based on region growing and morphological processing combined with a flexible threshold selection method.

With a median runtime of 0.75 s and a maximum of 13.6 s, the algorithm is much faster than reported in any other publication. At the same time, it is among the most accurate methods, as shown in the Liver Tumor Segmentation Challenge in 2008. I evaluated it on a representative set of 371 lesions and achieved a median volume overlap of 62.5 % and a median Hausdorff distance of 4.5 mm, equivalent to a MICCAI score of 79.5. This score cannot be compared to the results of the Challenge because different data were used, but it proves that the algorithm is able to produce results on a large variety of liver lesions which do not lag far behind manual segmentations. This was confirmed by evaluating a subset of 50 lesions with three reference segmentations.

Unlike most other publications, I also evaluated the dependency on the initialization. For ten optimal strokes that are actual lesion diameters, the median COV of the volume is less than 10 %, which is lower than the COV of ten manual segmentations (see Chapter 10). Even if strokes are allowed to be lengthened or shortened by up to 15 % at both ends, the median COV rises hardly above 20 %. These results indicate that a good compromise was achieved between incorporating information from the user input and not depending too much on it.

Although the special procedures for inhomogeneous and peripheral liver metastases did not make a measurable impact on the results due to their low frequency, some examples were shown to illustrate their benefit.

Part II

Automatic lesion tracking

Contributions A statistical analysis of change in a data base of 994 follow-up lesions under chemotherapy.
A simulation of the behavior of similarity measures on a lesion phantom and a discussion of invariance, robustness and specificity in the context of lesion tracking.
An algorithm for automatic lesion tracking, including segmentation initialization and a plausbility check.
A technical evaluation of the algorithmic performance on 209 independent lesions.
A study that evaluates possible benefits of the method for the clinical workflow.

Acknowledgments The work described in this part was done almost completely by myself, the only exception being the stroke propagation that was initially developed by Michael Schwier. The data for development and evaluation were collected and annotated by Hendrik Bolte, Michael Fabel, Ole Fischer, Rasmus Fortkamp, Barbara Frisch, Katrin Haag, Andreas Heverhagen, Andreas Kießling, Stephan Meier, Elena Peitgen, Nanette Peitgen, Michael Püsken, and Asmus Wulff. The workflow-centered evaluation was performed in a workshop at Fraunhofer MEVIS with Melvin D'Anastasi, Andreas Kießling, Daniel Pinto dos Santos and Christoph Schülke, all of whom also contributed to a paper about the results. Lars Bornemann, Stefan Braunewell, Frank Heckel, Benjamin Geisler, and Lei Wang were involved in the organization of the workshop.
This work was funded in part by Siemens AG, Healthcare Sector, Imaging & IT Division, Computed Tomography, Forchheim, Germany.

Publications A preliminary version of the algorithm has been presented at the *IEEE International Symposium on Biomedical Imaging 2009* in Boston (Moltz et al. 2009a) and the *European Congress of Radiology 2011* in Vienna. The workflow study has been published in *European Radiology* (Moltz et al. 2012). The problem analysis and most parts of the algorithm have not been published previously.

Chapter 5

Introduction

5.1 Introduction

Reading follow-up computed tomography (CT) examinations of patients undergoing chemotherapy constitutes a major part of the workload of a radiologist. In a follow-up situation, the user has to identify the target lesions in the follow-up image, initialize the segmentation, wait for the result, and refine it if necessary. This can be a time-consuming process, because it implies visual comparison of two 3D datasets and the relevant findings can be spread over various body regions.

On the other hand, this task has a high potential for automation. In many cases, finding lesions that have already been marked in a previous CT examination does not require a radiologist's knowledge. It often suffices to be able to detect similar image regions, which can also be done automatically by an algorithm. This enables starting the segmentation algorithms without any user interaction.

The goal of this part is to develop a comprehensive framework for *automatic lesion tracking* (ALT), which can run as a preprocessing step before a radiologists starts reading a case. Given segmented target lesions in baseline, it identifies these lesions in follow-up and precomputes the segmentations. The radiologist then checks whether the correct lesions were found and corrects the segmentation result if necessary. The benefit in clinical routine may be a reduction in both reading times and inter-reader variability.

Lesion tracking would be a trivial problem if a pair of images of the same patient always had exactly the same field of view or the same world coordinate system. In general, however, this is not the case, therefore establishing correspondences between the two images requires an analysis of the image content.

Two important design decisions were made with clinical acceptance in mind. First, the developed method should be as general as possible. ALT should be available for all lesion types that can be segmented in CT scans. However, it was also decided to optimize the algorithm for the most relevant tumor entities: lung nodules, liver metastases and lymph nodes. Second, the method should rather output an error message than a wrong result. This may apply to two classes of cases. First, an algorithm may have problems if there is a large number of lesions or if the anatomy changes markedly between the examinations, e.g. due to surgery. Second, lesions may simply become invisible in CT during therapy. The algorithm should be able to detect these situations and suppress implausible results.

5.2 Related work

Automatic lesion tracking is a relatively new field compared to lesion segmentation, and when I started working on the problem in 2007, no related work was available. This has changed in the meantime. In this context, however, I am not referring to "lesion pairing" methods that establish correspondences between two given sets of lesions from two images. This is a conceptually different approach, that requires highly sensitive lesion detection methods, which currently are available only for lung nodules.

The literature on lesion tracking can be categorized along two criteria. The first one is the lesion type in focus. The majority of papers deal with lung nodules, but interest in lymph node tracking is growing. Up to now, no other group has presented a generally applicable lesion tracking framework, although there is potential for generalization in some of the methods.

Secondly, there are different methodological approaches to the tracking problem. All of them, however, make the basic assumption that the lesions and/or their surroundings look similar in baseline and follow-up. The approaches are different in the way similarity is measured and, more importantly, the way the most similar image position is found.

In order to get an initial estimate of the lesion position in follow-up, a global rigid transformation between the two images is often used. Since we have to deal with elastic deformations of tumors and organs over time, this is obviously not sufficient for exactly transforming a baseline lesion into the corresponding follow-up lesion. But defining a coarse search region can help to speed up the computation and avoid implausible results.

A common strategy is to combine a global registration with template matching in a local search region. This has been done by Shi et al. (2007) and Wiemker et al. (2008) for lung nodules, by Opfer et al. (2008) for general lesions in PET/CT, and it is also the basis of my algorithm. They all use cross-correlation as a similarity measure, but different policies are used to further restrict the search region. Shi et al. (2007) developed a classifier that detects spherical structures based on the eigenvalues of the Hessian as lung nodule candidates. They also restrict the search region depending on the distance to the lung surface, which requires a lung segmentation. Wiemker et al. (2008) use a search scheme that follows increasing similarity, while Opfer et al. (2008) use information from the PET data and follow increasing standard update values (SUV), which indicate the presence of a tumor.

Both Shi and Wiemker developed their own lung registration techniques, the former aligning the centerlines of segmented ribs, the latter using lung volume percentiles.

Template matching transforms only a single point inside the lesion to the follow-up image and requires an additional step to obtain a segmentation of the lesion. A local non-linear registration, on the other hand, transforms the image region surrounding the tumor and can therefore directly give at least an initial estimate of the desired segmentation.

This idea was implemented by Yan et al. (2007) in a lymph node tracking application. They apply a registration based on optical flow, which computes a displacement field that minimizes the sum of squared differences under a smoothness constraint. The deformed segmentation mask is then used to set internal and external markers, which are the input to the segmentation algorithm. Also for lymph nodes, Xu et al. (2011) and Yu et al. (2012) model their deformation using B-splines on a set of control points. Although all papers use the term "free-form deformation", the referred methods are actually quite different. Xu transforms not only the lymph node segmentation to the

follow-up image, but also a number of neighboring structures which are determined in the baseline image using mean shift clustering. They are used as a restriction for the subsequent refinement of the lymph node segmentation.

A similar strategy has already been adopted for lung nodules by Sun et al. (2007). They do, however, not use a single non-linear transformation, but an individual rigid transformation for the nodule and each surrounding structure, which is optimized under a global consistency constraint. The initial alignment of the two images is based on relative coordinates in the segmented lungs.

Another registration-based approach for lung nodules was presented by Sofka and Stewart (2010). They skip the global alignment and replace it by a matching between feature points based on SIFT-like descriptors. The position of a lung nodule in follow-up is then determined by the established correspondences of the surrounding feature points. This method is so far the only one that contains a plausibility test and tries to correct itself when a possible error is detected. This is decided by a support vector machine that is fed with different measures from the forward and backward transforms.

For evaluating lesion tracking methods, two measures are common. Registration-based methods typically report the distance between the transformed lesion center and a manually marked reference point. However, lesions can vary substantially in size, and the distance from the center does not tell us whether the detected point is inside the lesion. For subsequent segmentation, a seed point that is slightly off-center is typically not a problem. Therefore, the hit rate is a more meaningful criterion for this application. Interestingly, reported hit rates are almost always 90 % or higher for the different methods and lesion types, but readers should always scrutinize the study setup. In several papers, the number of *different* lesions used for testing was less than 30, or the "follow-up" image pairs were acquired within a few minutes. At least for lung nodules, however, hit rates around 90 % have been confirmed by several clinical studies using commercial tools (Beyer et al. 2004; Beigelman-Aubry et al. 2007; Lee et al. 2007; Tao et al. 2009).

Chapter 6

Data and problem analysis

Virtually all existing methods are based on what may be called the *similarity assumption*. This means that a lesion in follow-up is not only identified by its position in the body, but also by its appearance as compared to the baseline lesion. Although this assumption is intuitively reasonable, it has never been verified on a large data base. Several authors have evaluated the performance of their specific algorithms, but it is also instructive to analyze the data on a more abstract level. This can provide a better understanding of the task and potential problems that a method will have to face. In the context of this thesis, it also serves as a specification of the data used for development and testing. So far, there is no publication that provides a statistical analysis of how lesions change their appearance under chemotherapy from an image analysis point of view. Furthermore, I have access to more than 1200 annotated lesion pairs, which is far more than what previous papers are based on. However, this should be regarded as a purely technical investigation that does not claim any medical relevance.

6.1 Data

The basis of my analysis is an extensive collection of follow-up data from patients undergoing chemotherapy. In total, it comprises 1268 lesions. Table 6.1 provides an overview of the data sources and the number of lesions per type. The data were collected retrospectively from five German university hospitals. Scans were acquired according to local protocols with imaging parameters as listed in Table 6.2. The data from Kiel were not available during development and were only used as independent test data in the final evaluation. They are not included in the subsequent statistics and in the parameter tests.

The data are relatively heterogeneous, because it was initially collected in the context of various studies that evaluated different aspects of lesion volumetry. This way, however, it constitutes a representative collection of follow-up images that may appear in clinical practice. In particular, it covers CT scanners from three different manufacturers, images with reconstruction increments between 0.5 and 3 mm, various reconstruction kernels (mostly standard kernels for lungs and soft tissue) and contrast agent protocols. In some cases, data from a patient population with a particular diagnosis were assembled: Most of the lymph nodes are in patients with either malignant melanoma or lymphoma. In many other cases, the exact diagnosis was not disclosed.

An important parameter is the time interval between the two scans. Its distribution is shown as a histogram in Figure 6.1a. In the great majority of cases, it is approximately three months, but there are also some six-months and very few nine-months follow-up pairs.

In all the cases, radiologists have segmented the lesions and established the correspondences between baseline and follow-up. The identification of the corresponding follow-up lesion is visually

Lesion type	Hospital	Scanner	Patients	Lesions		Diagnosis
				with FU	no FU	
Lung	Freiburg	Siemens	32	174	7	various
	Marburg	Siemens	11	46	2	lung cancer
	Kiel	Siemens	19	65		various
Liver	Freiburg	Siemens	10	49		various
	Mainz	Phil./Siem.	47	130	12	various
	Marburg	Siemens	26	79	44	various
	Kiel	Siemens	23	71		various
Lymph node	Freiburg	Siemens	12	27		various
	DKFZ	Toshiba	42	200		melanoma
	Münster	Siemens	92	289		lymphoma
	Kiel	Siemens	30	73		various
			320	1203	65	

Table 6.1: Overview of the available data. Remarks: The Freiburg data contain 44 and the Kiel data 59 patients, some of whom have lesions of two types. In the Mainz data, both Philips and Siemens scanners were used, sometimes even for the same patient.

Parameter	Min	Median	Max
Slice thickness (mm)	1.0	1.5	3.0
Reconstruction increment (mm)	0.5	0.8	3.0
Tube voltage (kV)	120	120	120
Tube current (mAs)	128	300	500
Number of frames	63	401	743
Pixel spacing (mm)	0.25	0.70	0.98
Reconstruction kernel	Siemens: B20s–B60f Philips: B–C Toshiba: FC12		
Time between scans (d)	15	75	478

Table 6.2: Overview of acquisition parameters of the available data.

unambiguous in most of the cases. Although there may be uncertainty in a few cases, I consider all given correspondences to be correct. The segmentations were created by an algorithm and corrected interactively if necessary. Not all of them would be suitable as reference segmentations, but they are a sufficient basis for evaluating a lesion tracking method and for analyzing the change of appearance over time.

The lesions cover a wide range of sizes with a minimum of 0.002 ml and a maximum of 772 ml. The distribution of lesion volumes is shown in Figure 6.1b. It can be seen that the great majority of lesions are in the order of a few milliliters, the median volume being 0.35 ml (lung), 3.68 ml

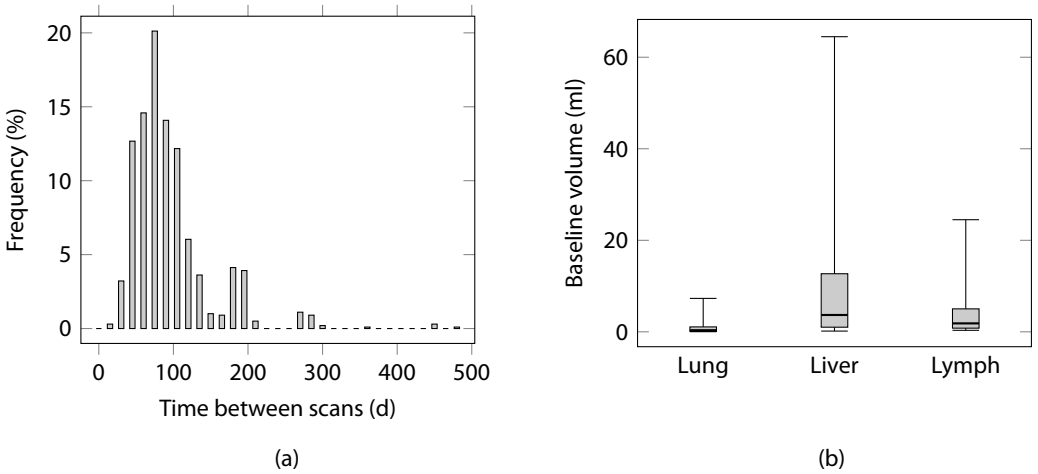


Figure 6.1: (a) Histogram of time between scans in available data. (b) Box plot of baseline lesion volumes in the data base.

(liver), and 1.73 ml (lymph node), respectively. There are, however, also some lesions which are considerably larger. Although they are exceptional cases, they are useful for testing. For instance, it is important to check how the computation time of a method behaves for large lesions.

In addition to the data mentioned so far, I collected 65 lesions that are visible in only one of the two scans. I detected and segmented these lesions myself in the same data base as the rest of the lesions. It is probably not a representative collection and does not reflect the frequency of vanishing lesions in clinical practice, but for testing it is important to have at least a small set of lesions where the method is supposed to return an error message rather than a result. Most of these lesions are in the liver, and unfortunately no lymph nodes are available.

6.2 Statistical analysis of change

The subsequent analysis will focus on two aspects of change: size and density. Changes of shape, texture or other more complex features will not be considered. The main purpose of this investigation is to get an idea of how much dissimilarity between baseline and follow-up a lesion tracking method may have to face.

6.2.1 Change in size

The change in size is measured by the relative volume difference

$$d_v = \frac{v_1 - v_0}{v_1 + v_0}. \tag{6.1}$$

This measure effectively normalizes the difference to the mean of the volumes of the two timepoints, leading to a symmetric measure for growth and shrinkage, which is advantageous for the analysis.

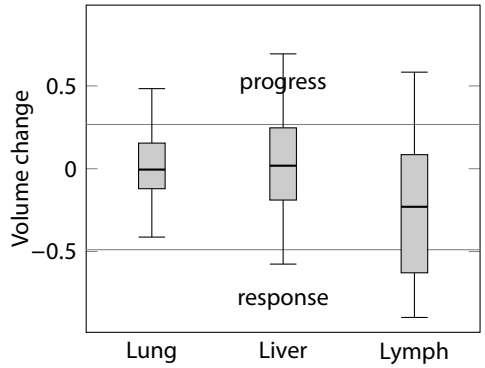


Figure 6.2: Box plots of volume change for the three lesion types. The whiskers show the 5 % and 95 % quantiles, respectively.

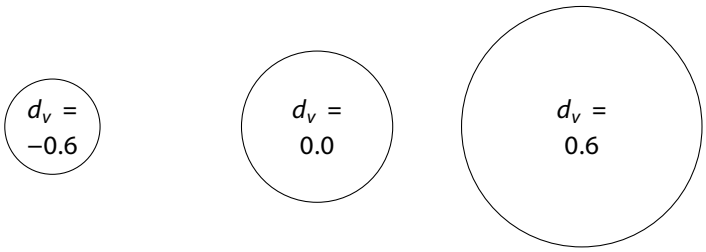


Figure 6.3: Illustration of volume change of $|d_v| = 0.6$ compared to the baseline volume (center). Each image shows a cross section through the center of a sphere with the respective volume.

As an example, we have $d_v = \pm\frac{1}{3}$ for a lesion that has doubled or halved its volume, respectively. Note that $d_v = -1$ if a lesion vanishes completely and $d_v \to 1$ if it becomes infinitely large.

A box plot of the volume change of the lesions in the data base is shown in Figure 6.2. Lung and liver lesions follow approximately a normal distribution with zero mean. Lymph nodes, in contrast, have a tendency to become smaller with a median d_v of −0.23, which is equivalent to a shrinkage by one third. According to RECIST, 19.4 % of all lesions classify as response and 17.4 % as progress, while 63.2 % are stable. This means that in most of the cases we have only moderate changes, which is already a partial confirmation of the similarity assumption. Still, volume change of a factor of 4 or more ($d_v \geq 0.6$) occurs with a frequency of more than 10 %. Figure 6.3 illustrates what this amount of change looks like in a 2D cross-section.

The box plot also shows a difference in overall volume variability between the lesion types. Since d_v is signed, this is reflected in dispersion measures such as the inter-quartile range. We observe increasing change from lung nodules (0.27) over liver metastases (0.44) to lymph nodes (0.71). Of course, these are specific properties of the data base that cannot be generalized. It is even a bit surprising that lung nodules are the most stable lesion type, since they are smaller on average and therefore more prone to inaccurate measurements which may feign stronger volume change. For lymph nodes, on the other hand, it is plausible to observe more change, because they are generally more variable in size. This is not only caused by cancer and its treatment, but also by their function

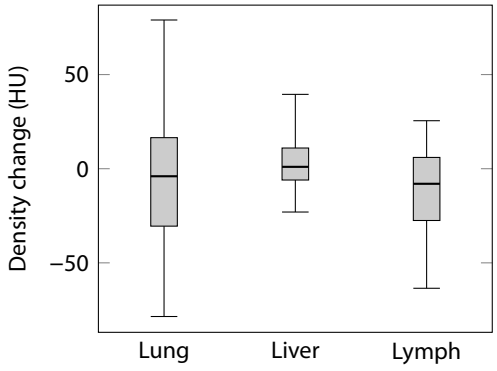


Figure 6.4: Box plots of density change for the three lesion types. The whiskers show the 5 % and 95 % quantiles, respectively.

in the immune system. The trend towards shrinkage is easily explained by the fact that lymph nodes are only relevant target lesions if they are enlarged, but often shrink again by themselves.

6.2.2 Change in density and contrast

The second aspect that I analyzed is the change in absolute lesion density and the change in contrast between a lesion and its surroundings. Image values are normalized in CT and a reasonable follow-up protocol requires the acquisition parameters, in particular the contrast agent state, to be similar in baseline and follow-up. Therefore in a first approximation, I assume that the appearance of the healthy tissue is the same in the two images. The density of tumors, however, may be affected by some therapies and also by progressing disease. A straightforward measure for absolute density change is the difference of the median densities under the lesion masks M_0 and M_1

$$d_d = (\widetilde{M_1})_{0.5} - (\widetilde{M_0})_{0.5}. \tag{6.2}$$

The box plots of d_d for the different lesion types are shown in Figure 6.4. It does not come as a surprise that the distributions are approximately normal with zero mean, i.e., on average lesions do not change their density. However, the variances seem to be different for the individual lesion types. As for the volume, the absolute amount of change can again be quantified by computing the inter-quartile ranges (IQR). We see that there is little change for liver metastases (17 HU), while lung nodules show a much higher variability (46 HU) and lymph nodes are in between (33.5 HU).

The variability of the lesion density alone does not give a complete picture of the similarity between baseline and follow-up. It needs to be related to the contrast between a lesion and the surrounding healthy tissue. In order to get estimates from the data base, I compared the median lesion density with the median density of a region around the lesion. This makes sense for lesions in the lungs and liver because most of them are surrounded by homogeneous parenchyma. For lung nodules, the median contrast in the data base is 739 HU (IQR 165 HU), which makes the observed density changes negligible in comparison. For liver metastases, on the other hand, we have a median contrast of −45 HU (IQR 29 HU). In relation to this, the typical density change

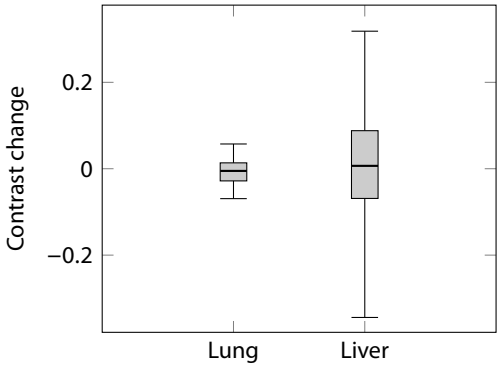


Figure 6.5: Box plots of contrast change for the two lesion types. The whiskers show the 5 % and 95 % quantiles, respectively.

may already pose problems for a similarity-based approach. This is especially true for lesions near the liver boundary, which have structures other than liver parenchyma in their neighborhood.

More quantitatively, a measure for relative contrast change can be defined in analogy to d_v. Let

$$c_i = (\widetilde{M_i})_{0.5} - (\widetilde{S_i - M_i})_{0.5} \qquad \text{and} \qquad d_c = \frac{c_1 - c_0}{c_1 + c_0}. \tag{6.3}$$

Again, the range is between −1 and 1 with the interpretation that $d_c = -1$ when there is no visible contrast left in follow-up and $d_c \to 1$ for a maximally increased contrast.

Box plots of the contrast change for lung and liver lesions are shown in Figure 6.5. It is confirmed that the density variations in lung nodules are negligible compared to the high contrast to the lung parenchyma and will not pose a problem. Liver lesions, on the other hand, may double or halve their contrast to the healthy tissue in some cases. This is equivalent to $d_c = \pm\frac{1}{3}$, which are the 5 % and 95 % quantiles in the data base.

For lymph nodes, a similar estimation is not possible because they can be adjoined by multiple structures of different densities. But it is known that some of these structures, like muscles or blood vessels, may have very low contrast to the lymph node, which makes the task inherently more challenging than for lung and liver lesions.

6.3 Measuring similarity

The next step of the problem analysis is to investigate which degree of change is compatible with the similarity assumption. So far, I have used the term *similarity* without specifying how it is actually measured. Before discussing different ways to do this, some important terms shall be clarified. A similarity measure is *invariant* against (a particular kind of) change if it gives approximately the same values under changing conditions. It is *robust* against change if the position of its optimum is approximately the same under changing conditions. In the context of lesion tracking, robustness may be interpreted in a broader sense as the optimum being located inside the lesion, but not necessarily in its center. Invariance and robustness are closely related, but theoretically independent.

In the following formulas, assume that the template block is cubic with $b \times b \times b$ voxels. Here, b is odd and the central voxel is the origin of the local coordinate system. The sums iterate over all positions within a block, i.e., over all $\boldsymbol{v} \in \mathbb{Z}^3$ with $\|\boldsymbol{v}\|_\infty < \frac{b}{2}$. Furthermore, let $\bar{I}(\boldsymbol{x})$ denote the mean of a block in image I centered at $\boldsymbol{x}$, i.e.,

$$\bar{I}(\boldsymbol{x}) = \frac{1}{b^3} \sum_{\boldsymbol{v}} I(\boldsymbol{x} + \boldsymbol{v}). \tag{6.4}$$

Similarity measures typically used for template matching can be divided into two groups. The first group is based on the assumption that corresponding pixels have similar image values and therefore compares values directly. An example is the *sum of squared differences*

$$\mathrm{SSD}(\boldsymbol{x}) = \sum_{\boldsymbol{v}} \left(\left(I_0(\boldsymbol{x}_0 + \boldsymbol{v}) - I_1(\boldsymbol{x} + \boldsymbol{v}) \right) \right)^2 \tag{6.5}$$

and several variants that use a median instead of a sum or absolute instead of squared differences. Since these measures are based on direct comparisons, it is obvious that they cannot be invariant against changes in size or density. However, they may still be robust against moderate changes. For example, when a lesion grows, only a small number of voxels in a block actually change their values, so it is likely that the lesion center still has the best similarity.

The second group of similarity measures is based on the correlation between the values of corresponding voxels. In theory, this makes them perfectly invariant against uniform intensity changes. In practice, this means that when the contrast between a lesion and the background changes, the similarity will remain very high. The most popular measure with this property is *zero-mean normalized cross-correlation*

$$\mathrm{ZNCC}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{v}} \left(I_0(\boldsymbol{x}_0 + \boldsymbol{v}) - \bar{I}_0(\boldsymbol{x}_0) \right) \cdot \left(I_1(\boldsymbol{x} + \boldsymbol{v}) - \bar{I}_1(\boldsymbol{x}) \right)}{\sum_{\boldsymbol{v}} \left(I_0(\boldsymbol{x}_0 + \boldsymbol{v}) - \bar{I}_0(\boldsymbol{x}_0) \right)^2 \cdot \sum_{\boldsymbol{v}} \left(I_1(\boldsymbol{x} + \boldsymbol{v}) - \bar{I}_1(\boldsymbol{x}) \right)^2}. \tag{6.6}$$

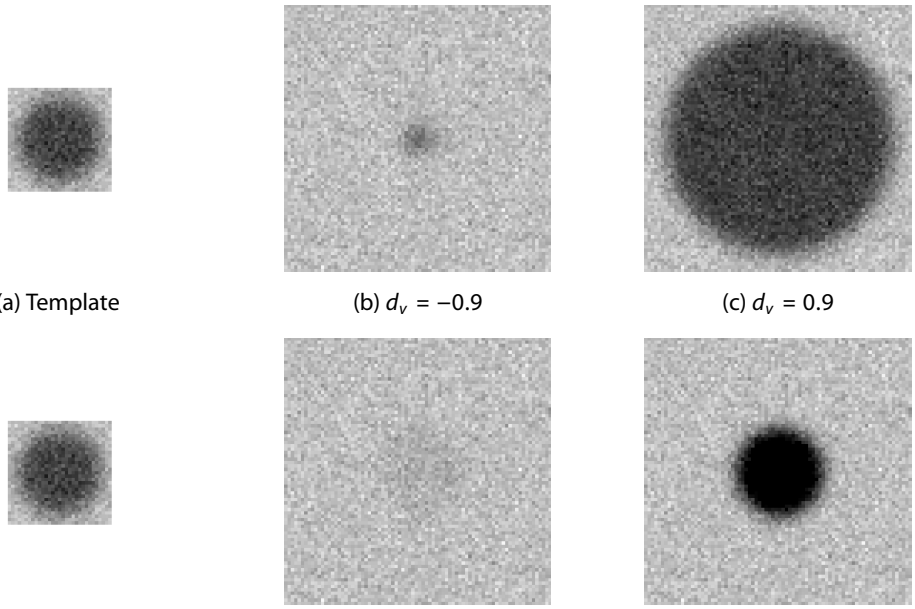
When the subtraction of the mean is omitted in each term, we get *normalized cross-correlation*. The measures of this group are not invariant against volume change either, because the assumption of correlation is violated for the voxels that are added to or removed from the lesion.

6.4 Simulation of change

In order to better understand the behavior of the two classes of similarity measures under change, I made some experiments with a simple lesion phantom that can grow or shrink and increase or decrease its density.

6.4.1 Lesion phantom and experiment setup

The lesion phantom is an abstraction of the "median liver lesion" in the data base. It is a sphere with a radius of 9 mm (volume approximately 3 ml) in baseline and has a density of 52 HU, surrounded by a background of 92 HU. In order to approximate some properties of CT imaging, the sphere is blurred by convolution with a Gaussian kernel ($\sigma = 2$ mm), and zero-mean Gaussian noise is



(a) Template (b) $d_v = -0.9$ (c) $d_v = 0.9$

(d) Contrast = 40 HU (e) Contrast = 5 HU ($d_c = -0.78$) (f) Contrast = 80 HU ($d_c = 0.33$)

Figure 6.6: Illustration of the lesion change simulation with the purely geometric phantom. Upper row: Volume change. Lower row: Density change.

added ($\sigma = 5$ HU). For computational reasons, I chose an image resolution of $1.5 \times 1.5 \times 1.5$ mm^3, which is slightly lower than usual. The baseline phantom is displayed in Figure 6.6a.

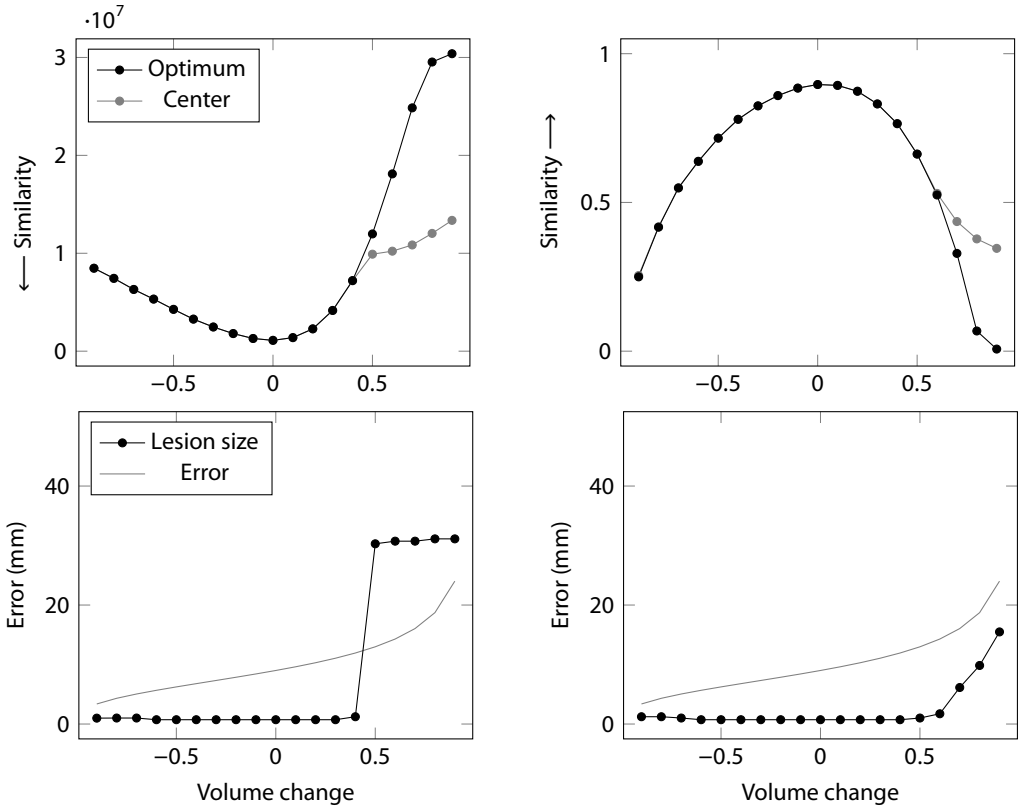
In the first experiment, I vary d_v from −0.9 to 0.9 in steps of 0.1 (Figures 6.6b and 6.6c). This is equivalent to follow-up volumes between 0.16 ml and 57 ml. In the second experiment, the contrast is varied from 5 HU to 80 HU in steps of 5 HU. This corresponds to contrast changes d_c between −0.78 and 0.33. (Figures 6.6e and 6.6f). Since the initial contrast between lesion and background is 40 HU, the lesion becomes hardly visible for the highest values of d_c, whereas for negative values the contrast is enhanced.

In both experiments, a template matching is applied using SSD and ZNCC. For each setting I record

- the similarity at the actual lesion center,
- the optimum similarity in the search region,
- the distance of the position of optimum similarity to the lesion center; by comparing it to the radius of the lesion, it can be checked whether the optimum is inside the lesion.

6.4.2 Similarity under change in size

The results of the volume change experiments are shown in Figure 6.7. We see a characteristic behavior of the two similarity measures under volume change. SSD is not invariant against



(a) SSD (b) ZNCC

Figure 6.7: Results for volume change (d_v) in the lesion phantom.

shrinkage, but reacts much more strongly to growth, whereas ZNCC has a roughly symmetric invariance for shrinkage and growth.

Regarding robustness, we can make the important observation that SSD is not robust for $d_v \geq 0.5$. At this point, the similarity measure "flips" and has better values in the background than in the lesion, because it is not able to match the additional lesion voxels any more. This means that even in theory, SSD is not suitable in tasks where moderate growth can occur. It is, however, robust for shrinking lesions.

In spite of its invariance, ZNCC turns out to be robust (in a broader sense) even for strong changes in size. For $d_v \geq 0.6$, the optimum gradually moves away from the lesion center, but it stays inside the lesion.

6.4.3 Similarity under change in density and contrast

The results of the density change experiments are shown in Figure 6.8. As predicted in theory, it is not invariant for the density-based measure SSD and shows a symmetric behavior for increasing and decreasing contrast. For ZNCC, it is perfectly invariant as long as the contrast is higher than

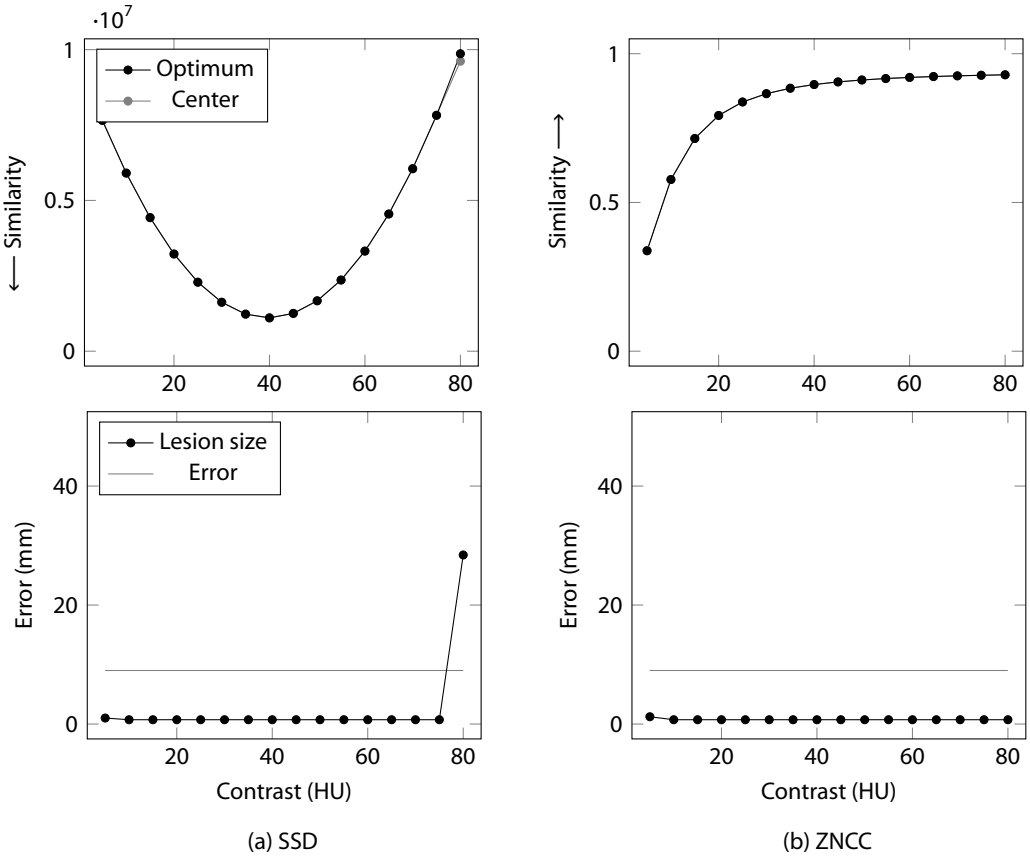


Figure 6.8: Results for varying contrast in the lesion phantom.

the noise level. Then, similarity drops quickly. Still, both measures are very robust (in a strict sense) and always locate the center of the lesion accurately.

An observation that may come as a surprise is that ZNCC does not assume its optimum when the lesion has the same density as in the template, but that it still increases slightly when the contrast gets higher. Although this does not affect robustness in the phantom experiment, it might cause problems in real data, because it means that high-contrast structures have a strong attraction for correlation-based measures, no matter whether or not the absolute image values are similar.

On the other hand, it is also striking that even when the contrast is as low as 5 HU (see Figure 6.6e), both measures are still able to robustly locate the lesion.

6.5 Discussion

Before discussing the results of the experiments, let us summarize the main limitations of these simulations. While the lesion phantom has been modeled to resemble the lesions found in the data base, the background was assumed to be homogeneous. In real images, however, the background contains a lot of structural information that can help a matching procedure. The vessels in the

lungs or liver, the boundaries of these organs, and for lymph nodes all kinds of adjacent structures can serve as landmarks. But on the other hand, the background may also contain structures that can be confounded with the lesion, such as other lesions or, in the case of lymph nodes, vessels or muscles that often have a similar density. In this sense, the phantom is a model for cases that would be deemed easy by users. They would expect the method to work for a single lesion on a relatively homogeneous background, even if it shows a strong change in size or density.

In spite of these limitations, there are some things to learn for the design of a lesion tracking method. Standard difference-based measures such as SSD have problems with growth and strongly increased contrast. The latter is especially critical since it is an unintuitive behavior if a lesion that sticks out more than before cannot be tracked.

ZNCC is robust to density changes by construction and it also proved robust to volume changes in my experiments. However, the simulation also revealed an unfavorable property of correlation-based measures: They have a tendency to be attracted by objects with a high contrast. Since the density values are not directly compared, this can also lead to unintuitive results from a user perspective.

This last aspect is a symptom of a general problem. Invariance and robustness seem to be desirable properties of a similarity measure when the objects of interest can change their appearance. However, they inherently reduce *specificity*, i.e., if a similarity measure tolerates a large amount of change in the "right" object, it is also more likely to assign "wrong" objects a high similarity. This may lead to errors which can be hard to understand for users. They would primarily expect the method to work reliably on "easy" cases where the lesion remains visually similar.

It was shown that the majority of the lesions have a more or less stable apperance over time, as compared to the changes in size and contrast that similarity measures can tolerate. This justifies an approach based on similarity. The optimal trade-off between robustness to change and specificity, however, cannot be deduced by theoretical reasoning or simulations. Only experiments on real data can capture all possible scenarios and reveal the best strategy.

Chapter 7

Algorithm and technical evaluation

7.1 Overview

The interface of the algorithm is specified as follows:

Input A baseline image I_0 and a follow-up image I_1 of the same patient. A list of mask images $M_{0,i}$ for the target lesions in I_0 and additional information about the lesion type $t_i \in \{\text{lung, liver, lymph, other}\}$.

Output For each target lesion, either the mask image $M_{1,i}$ or the message not found.

The basic structure of the method is represented as a flowchart in Figure 7.1. As motivated by the previous section, the core is a *template matching* that looks out for the image region which is most similar to the baseline lesion. Since an exhaustive search over the entire follow-up image would take several minutes, two steps are added to limit the number of voxels to be considered. First, a *global registration* of the two images is performed, so that the body region of interest can be found and the search space can be restricted. Then, *candidate voxels* within the search region are identified. When looking for a particular type of lesion, many voxels can be discarded just by checking their gray values or other local features. The candidates are then handled by the more expensive template matching. In the best case, the candidate detection may not only speed up the algorithm, but also increase the success rate. It is also advantageous when using a correlation-based similarity measure, which may be attracted to high-contrast structures outside the expected intensity range.

Besides the actual matching, the method comprises an automatic initialization of the *segmentation* in follow-up. Since template matching computes only a single point in the lesion and segmentation algorithms often require more information, this is a necessary step for automating the complete process. Finally, a *plausibility check* was implemented. For now, it just discards results that are probably wrong, for example if a lesion has vanished under therapy, but also if a lesion or its surroundings have changed so much that template matching fails. As a possible extension, this check might also trigger a reparametrization of the algorithm, which could use a different similarity measure or weaken some assumptions of the candidate detection.

The following sections will discuss the individual components of the method in more detail. The evaluation of each component is included in the respective section, whereas a discussion of the framework as a whole will be given in Chapter 8.

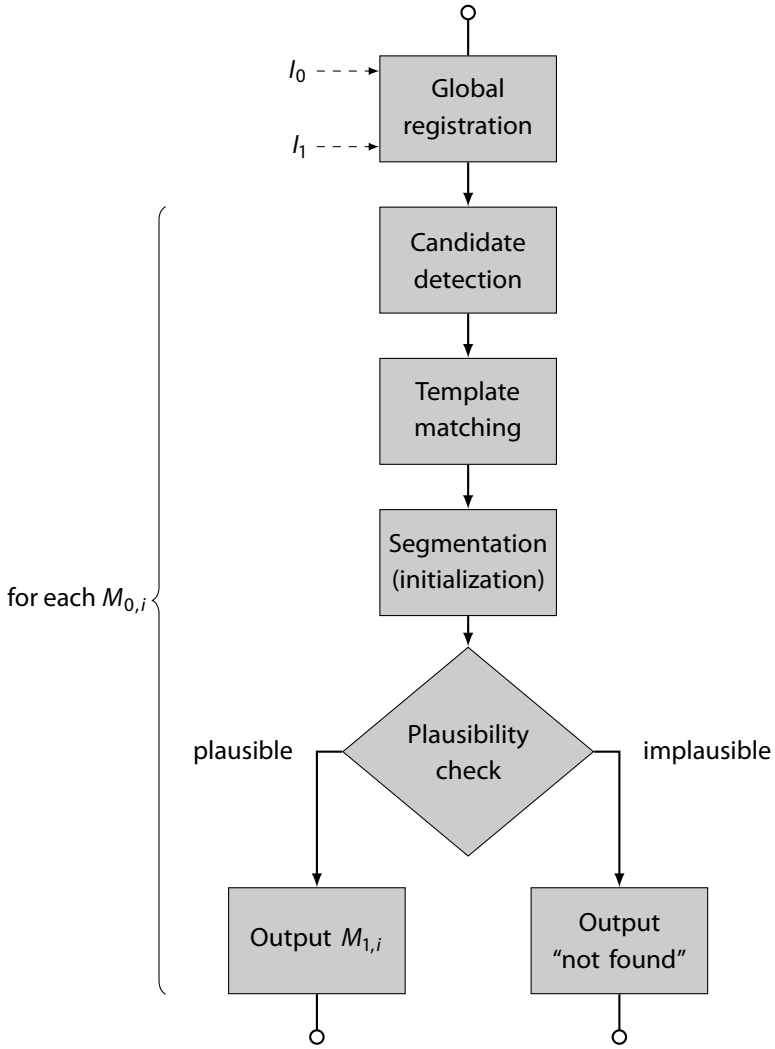


Figure 7.1: Flowchart of the proposed method.

7.2 Global registration

The first step is a global registration of the two images, which allows the computation of an initial seed point for each lesion. I use a general-purpose rigid registration that coarsely aligns two CT images.

The seed point is computed by applying the rigid transformation to the center of gravity of the lesion mask. Then a search region around each seed point is extracted. The optimal size of the search region is governed by two contrary effects: On the one hand, the risk of missing a lesion that is too far away from the seed point should be minimized, but on the other hand a large search region can contain other lesions or similar structures that might distract the algorithm. Of course, the former aspect depends on the accuracy of the registration.

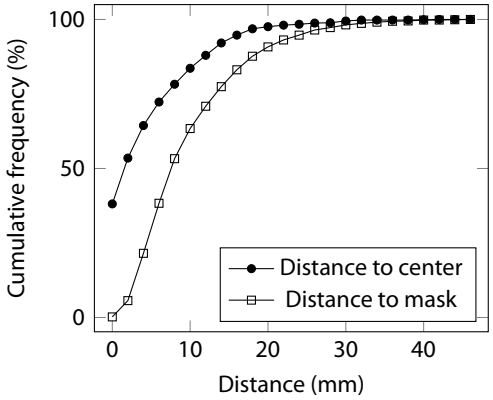


Figure 7.2: Quality of global registration. Cumulative histograms of the distance between the seed point and the lesion center or the lesion mask on the test data base consisting of 994 lesions.

I made experiments with the registration and computed the distance from the seed point to the lesion center as well as to the lesion mask. It seems to be a reasonable requirement that the lesion center be included in the search region. For large lesions, however, it is usually sufficient if the search region covers a part of the lesion. Note that the search region contains all voxels which are used as block *centers* for template matching. The union of all actual blocks will be larger.

Figure 7.2 shows cumulative histograms of the distances on the development data base. The maximum norm was used as a distance measure in order to account for the cubic shape of the search region, which is used for computational simplicity. I found that the lesion center is *always* within a distance of 47 mm from the seed point, whereas a distance of 42 mm is sufficient to cover at least a part of the lesion *in all cases*. In practice, however, such a large search region is not optimal. It will not only slow down the computation considerably, but also increase the risk of mismatches since in *almost all of the cases* a smaller search region would suffice. For example, with a distance of 20 mm, we still cover the lesion center in 89.5 % and the lesion mask in 97.3 % of the cases. In the evaluation, a search region size of 20 mm will be used.

Investigating the influence of the chosen registration method to the overall performance was outside the scope of my work. It can be expected that a more sophisticated registration allows a smaller search region and thereby reduces the risk of errors.

7.3 Candidate detection and template matching

7.3.1 Template matching

In its next phase, the algorithm tries to find the lesion candidate that is most similar to the lesion that was segmented in the baseline scan. This section shows how a template matching approach was optimized for this purpose.

Reference block

The size of the reference block from the baseline image is derived from the lesion size, so that that the block will always contain the complete lesion and a particular fraction of background voxels. This can be seen as an adaptation of scale, which is necessary since lesions can vary in size considerably. If the block size were fixed, two unfavorable situations could arise: For large lesions, a block would contain no background at all and template matching would pick a more or less random position. For small lesions, on the other hand, the block would contain so much background that the lesion itself could hardly have an impact on the matching result.

In my implementation, the reference block is a cube around the center of gravity of the lesion. Its size is derived from the longest edge of the bounding box of the lesion mask, multiplied by a constant factor. The factor was experimentally optimized and set to 1.2. The results of the optimization tests for this and other parameters can be found in Section 7.3.3.

Performing template matching in the original image resolution is computationally prohibitive, even more so since voxels should be isotropic. I decided to employ a resampling strategy that creates reference blocks with a fixed number of voxels, regardless of their physical size. This means that the resolution is adapted to the scale of the object of interest. A large lesion can still be clearly identified in a coarse resolution, whereas for small lesions all the information from the original image is needed. Experiments showed that $15 \times 15 \times 15$ voxels offer a good compromise between accuracy and efficiency.

Search region

As motivated in Section 7.2, the search region is a cube with an edge length of 40 mm. Since template matching requires approximately the same resolution for both images and ideally isotropic voxels, the resolution of the reference block is also applied to the search region. In both cases, trilinear interpolation is used.

Similarity measure

Zero-mean normalized cross-correlation is used. This choice has already been discussed in the context of the phantom experiments (Section 6.4) and was backed up by experiments.

Search strategy

An exhaustive search is conducted. Non-exhaustive search strategies based on increasing similarity are not applicable because the search region contains too many surrounding structures which might lead the search into a wrong direction.

7.3.2 Candidate detection

Even with the adaptive resampling strategy, template matching is still a computationally expensive procedure. In a typical search region, the lesion is surrounded by healthy lung or liver tissue or by other structures that clearly do not resemble a tumor. Therefore, a majority of the voxels in the search region can be discarded by criteria much simpler than template matching.

Since the idea of a candidate detection step is to speed up template matching, the detection itself must be very fast. On the other hand, the sensitivity has to be very high, i.e., it should be very unlikely that the correct lesion is discarded completely. However, it is not necessary to preserve *all* lesion voxels as candidates.

There is an analogy between this step and algorithms for computer-aided detection (CAD) of lesions for initial staging. But aside from the fact that in the present scenario sensitivity is much more important than specificity, there are two further differences. First, the baseline image provides some information to predict the appearance of a lesion in follow-up. As the statistical analysis in Section 6.2.2 showed, there is little change in size and density for most cases. The second difference to a typical CAD scenario is the lack of a lung or liver segmentation, which means that there may be spurious candidates outside the organ in question.

A straightforward choice is to use intensity features which can be implemented by simple thresholding. I propose to use specific thresholds for lung nodules and liver metastases, motivated by the results of the statistical analysis and inspired by the respective segmentation algorithms developed at Fraunhofer MEVIS and based on on the density range in baseline. The procedure for lymph nodes, however, is generic and can be used for other lesion types as well.

Lung nodule candidates

The statistical analysis revealed that lung nodules have a relatively high variability in density. Given the strong contrast to the lung parenchyma, this is not a problem for template matching, but in order to make sure that all nodules are included as candidates, I use a large threshold range comprising all voxels over −400 HU. This is the same threshold range that is applied in our lung nodule segmentation method (Kuhnigk et al. 2006).

The downside of this large range is that it includes most structures outside the lungs and therefore does not speed up the template matching as much as desired. As a solution, I first segment the lungs with an *upper* threshold of −400 HU. To be sure to include the nodule of interest, a rolling ball closing is applied with a kernel size of twice the diameter of the baseline lesion. Then, candidates are detected only under this coarse lung mask.

Liver lesion candidates

Unlike lung nodules, liver lesions often have a relatively low contrast to the healthy liver tissue and they cannot be captured with fixed threshold values. In order to ensure a good separation of lesions and parenchyma, a simple histogram analysis is performed. Estimates for the typical values of these two structures are obtained in a similar way as in my liver lesion segmentation method (see Part I).

Let ℓ_-, ℓ and ℓ_+ be the 10 %, 50 % and 90 % quantiles of the baseline mask, and let p_0 and p_1 be the modes in a ROI around the baseline lesion or the follow-up seed point, respectively. The threshold for the lesion-parenchyma separation is set to $p_1 - \frac{1}{2}(\ell - p_0)$, assuming a similar contrast in baseline and follow-up. The threshold for separating structures outside the liver is $\min(\ell_-, 10\,\text{HU})$ or $\max(\ell_+, 180\,\text{HU})$ for hypodense and hyperdense lesions, respectively, to make sure that the typical range of liver lesions in CT is covered.

Generic candidates

For lymph nodes, neither fixed thresholds nor a homogeneous background structure can be used. The same is true for a "generic" mode that should be usable for arbitrary lesion types. The only available information is the gray value distribution in the baseline image and the fact that it is mostly constant.

In these cases, the range between the 10 % and 90 % quantiles of the gray values in the baseline lesion is used. This allows a robust estimation that is not too susceptible for noise.

7.3.3 Evaluation

Template matching parameter tests

In order to ensure the optimal parametrization of the method, I conducted tests on the 994 lesions of the development data base. The parameters in question are the block size factor, the size of the template block after resampling, the size of the search region, and the similarity measure. I varied these parameters while leaving the rest of the algorithm unchanged. Since there is often a trade-off between quality and computation time, the results are presented in plots inspired by ROC curves, which are used to select the optimal working point (Figures 7.3a to 7.3d).

The plots reveal that the method is very robust to changes of parameters. The best overall matching rate that could be achieved was 80.0 % at a computation time of approximately 1 s. Although a wide range of values and several similarity measures were tested, the matching rate dropped hardly below 75 %. The computation time varied in a range of about 0.2 s and never exceeded 1.3 s.

Looking at the individual plots, a clear optimum was apparent for the block size factor (Figure 7.3a). This is plausible because it represents a trade-off between tumor and background being present in the template block. For the resampled block size, better results were observed for higher resolutions, but above $15 \times 15 \times 15$ voxels the improvement in accuracy becomes marginal in relation to the increasing computation time (Figure 7.3b). According to Figure 7.3c, the optimal search radius is 20 mm.

In Figure 7.3d, six different similarity measures are compared: sum of squared differences (SSD), sum of absolute differences (SAD), median of squared differences (MSD), median of absolute differences (MAD), normalized cross-correlation (NCC), and zero-mean normalized cross-correlation (ZNCC). ZNCC clearly has the highest success rate (80.0 %), whereas all other tested measures showed very similar results around 77 %. ZNCC is slightly slower than SSD, SAD and NCC, but the difference is in the order of magnitude of 0.1 s. Measures involving median computation take significantly more time but are still clearly below 1 s. These results are in agreement with the theoretical considerations of Section 6.4, although the superiority of correlation-based measures might be less pronounced than expected.

Candidate detection parameter tests

The first parameter to test in candidate detection is the quantile argument that is used for determining the thresholds for lymph nodes. The result is shown in Figure 7.4a. Note that the difference in matching rate between the tested values is less than 1 percentage point.

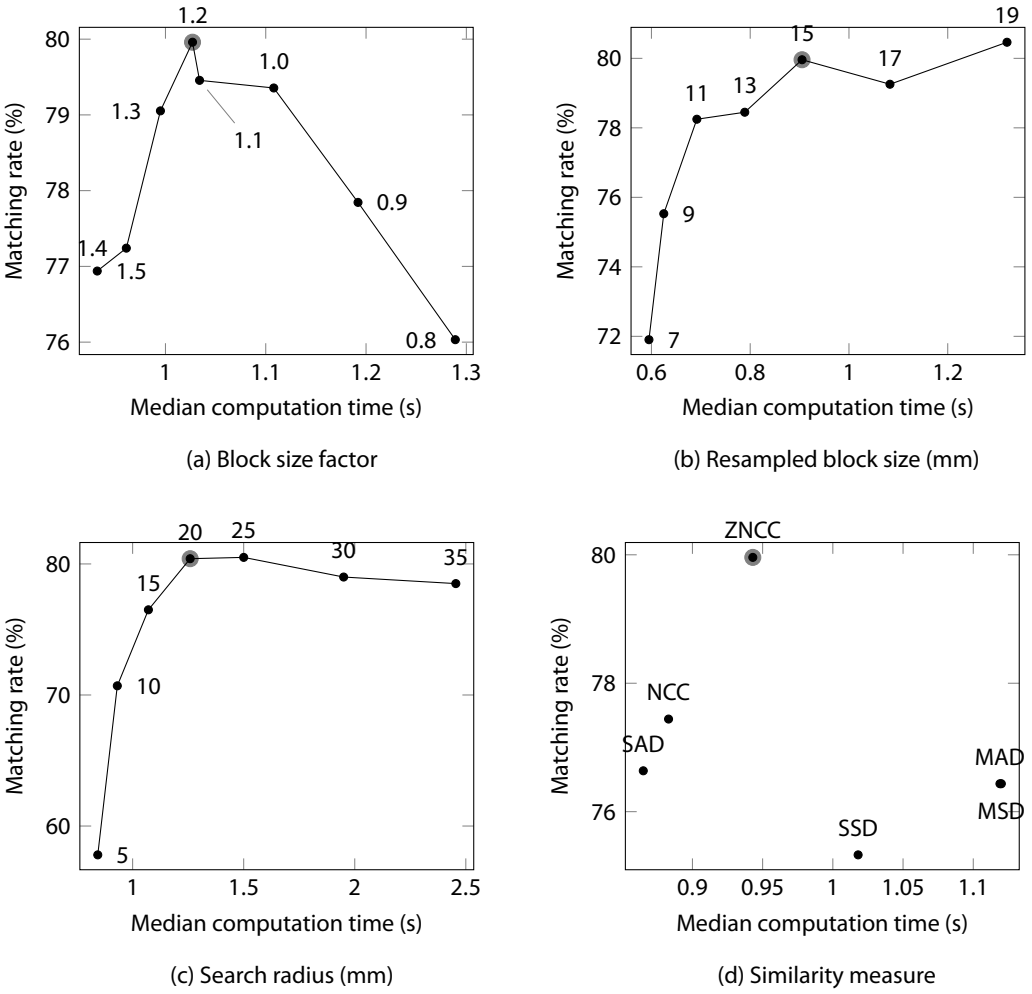
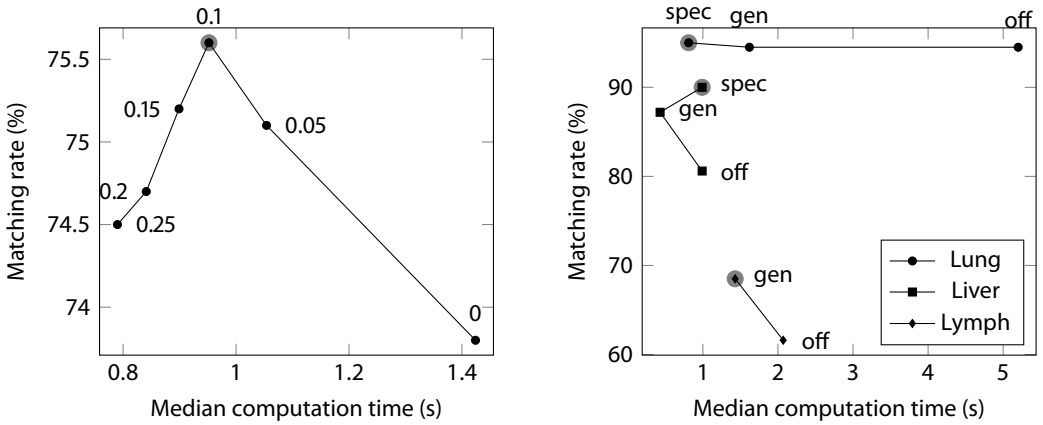


Figure 7.3: Parameter optimization for template matching. Working points are denoted by a gray overlay.

Secondly, I tested the impact of using candidate detection at all and of applying the special handlings for lung and liver lesions as described earlier. In Figure 7.4b, the three possible modes are denoted by candidate detection off, default handling, and special handling.

The most important result for the candidate detection is that it improves the matching rate and the computation time simultaneously. An increase in matching rate of about 7 percentage points can be observed for liver lesions and lymph nodes, whereas for lung nodules the performance is already very good without candidates. Lung nodules, however, show the strongest reduction of computation time because normal lung tissue voxels can be easily discarded by the thresholding criterion and the additional lung segmentation removes many further spurious candidates. With the special thresholding strategy for liver metastases, an even slightly higher matching rate is achieved, but the histogram analysis evens out the speed-up from evaluating less candidates in template matching. It is still faster than lymph nodes, though, which typically produce most candidates.



(a) Quantile arguments

(b) Candidate detection modes

Figure 7.4: Optimization of candidate detection.

	Matching rate (%)	
	Development data	Test data
Lung	95.0	89.2
Liver	90.0	76.9
Lymph	68.5	84.9

Table 7.1: Matching rate (in %) in the development data set (994 lesions) and the independent test data set (209 lesions).

Although the overall matching rate increases, it is also important to check how often the candidate detection actually discards the right lesion. This happens in 11 of the 994 lesions, with roughly equal frequency for each lesion type.

Matching rate by lesion type and amount of change

It was already mentioned that the overall matching rate on the development data set was 80.0 %. Table 7.1 gives the separate numbers for the three lesion types and compares them with the independent test data set (the Kiel data in Table 6.1). Although the average matching rate in the test data (83.3 %) is similar, the numbers reveal that the actual performance of the algorithm depend not only on the lesion type, but can also vary considerably across different data sets.

Looking at the lesion types at first, lung nodules always have the best results at approximately 90 %. For liver metastases and lymph nodes, results vary quite considerably between the two data sets, with differences of more than 15 percentage points. While the matching rate for liver metastases is higher in the development data, lymph nodes perform better in the test data.

For lymph nodes, I conducted a further analysis with respect to their location. Lymph nodes can be spread over large parts of the body, and the performance of a template matching approach may vary. For example, in the neck, lymph nodes are closely packed with muscles and other

	Development data		Test data	
Location	Number	Matching rate (%)	Number	Matching rate (%)
cervical	47	56.0		
axillary	64	68.8	12	75.0
mediastinal	180	80.6	21	95.2
abdominal	109	56.0	16	81.3
– paraaortic	40	60.0		
– mesenteric	32	34.4		
– other	37	70.3		
iliac	62	67.7	12	92.3
inguinal	49	69.4	11	72.7

Table 7.2: Matching rate for lymph nodes grouped by locations.

structures that appear similar in CT, whereas lymph nodes in the inguinal are often surrounded by fat and clearly discernible. Some areas, most notably the abdomen, show a high anatomical variability over time, while the structure of the mediastinum is relatively constant. Table 7.2 shows the distribution and matching rate for lymph nodes grouped by locations. Mediastinal lymph nodes, which form the most frequent group in the data base, clearly showed the best results. The most severe problems occurred for cervical and mesenteric nodes.

As motivated in Chapter 6, the amount of change over time is a critical aspect for template matching. Although it was shown that theoretically matching is robust against a certain amount of change, in practice degrading performance has to be expected with increasing change. This assumption is verified by Figure 7.5a, which correlates volume change and matching rate. The plot shows that the overall matching rate is around 90 % or higher if there is little change (d_v between -0.3 and 0.5). The curve drops quickly for shrinkage, but clearly more slowly for growth, allowing a matching rate of 75 % even for $d_v = 0.8$, which corresponds to a volume change by a factor of 9. This tendency towards better results for growth might be caused by the fact that shrinking lesions get a (relatively) larger partial volume zone, which can be a problem for the candidate detection or the template matching itself, especially due to downsampling.

Similar plots have been created for density change and contrast change (Figures 7.5b and 7.5c). These plots have to be read with care, because the frequency of extreme changes is very low (see box plots in Figures 6.4 and 6.5) and therefore these values may not be representative. The impression one gets from the plots is that more change typically reduces the matching rate, but the correlation is not as clear as for change in volume.

Example results and problem classes

In addition to the quantitative results presented so far, this section will provide some examples showing typical successful applications of the methods, but also problem classes that could be identified. Starting with some positive examples, Figure 7.6 shows some non-trivial cases where lesion tracking succeeds. In reference to Section 6.2, the corresponding volume change (d_v) and contrast change (d_c) are given. Note the combination of two unusually high values in Figure 7.6c.

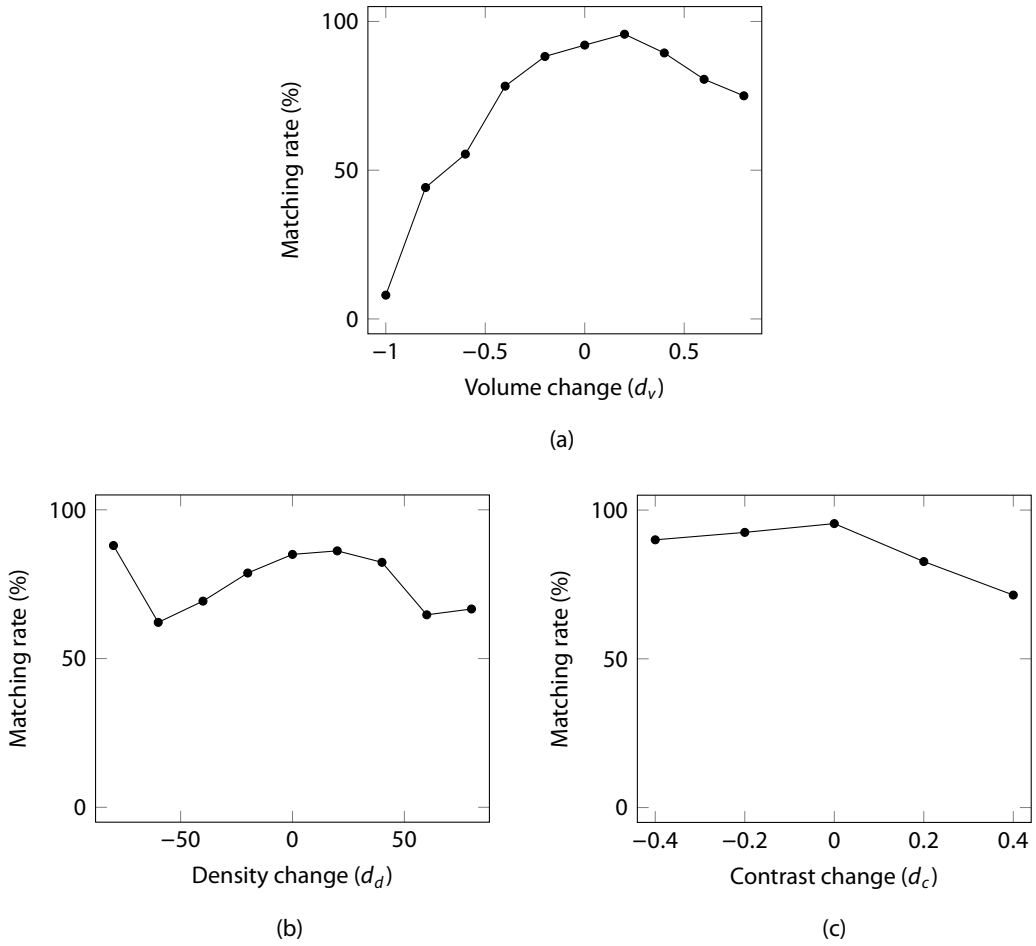
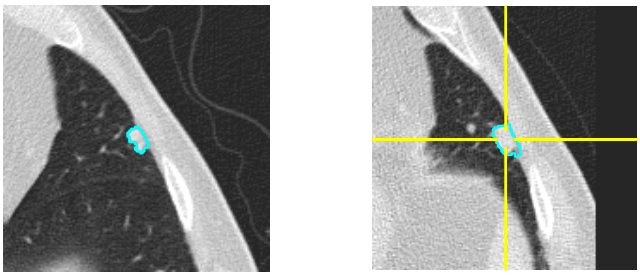


Figure 7.5: Matching rate grouped by change of volume, density, and contrast. Only groups containing at least ten cases are shown.

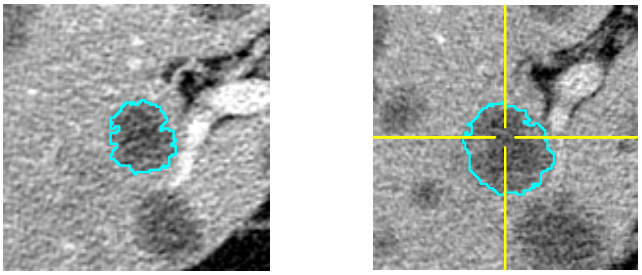
An additional aspect of change that is hard to analyze quantitatively is change in the anatomical relations. For instance, in Figure 7.6a, the two images were taken in differing breathing states so that the lung nodule is closer to the heart in the follow-up image. Another typical problem is the presence of structures that can easily be confounded with the target lesions such as other lesions (Figure 7.6b) or, for lymph nodes, muscles or vessels (Figure 7.6d).

I analyzed the cases where matching failed and identified the most common reasons with approximate frequencies with respect to *all* cases.

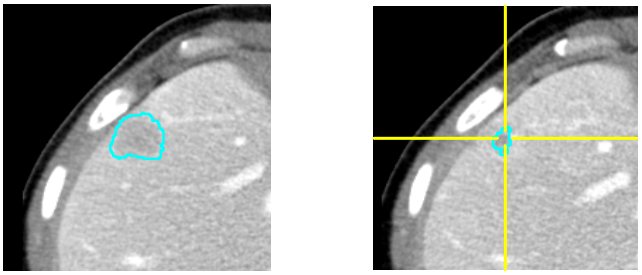
Inaccurate registration (3.5 %): As discussed in Section 7.2, depending on the quality of the registration, there are some cases where the lesion is not contained in the search region. In these cases, the result is typically a random candidate voxel, sometimes belonging to another lesion (Figure 7.7a), or there might even be no candidates at all, in which case an error is raised. Although this can be easily detected by a plausibility check, users cannot understand the problem unless there is a marked anatomical change which prevents registration, e.g. after surgery.



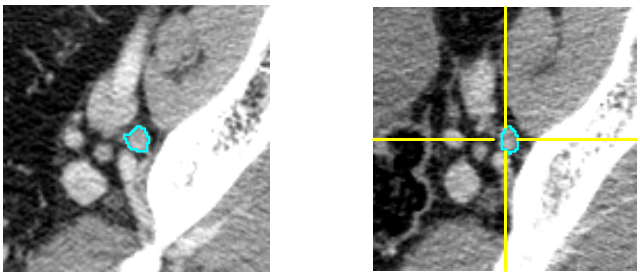
(a) Growing lung nodule with a changing position relative to the heart ($d_v = 0.48$, $d_c = 0.00$).



(b) Growing liver metastasis in a highly metastasized liver ($d_v = 0.32$, $d_c = -0.13$).



(c) Shrinking liver metastasis with increasing contrast ($d_v = -0.85$, $d_c = 0.47$).



(d) Stable lymph node surrounded by several similar structures ($d_v = -0.10$).

Figure 7.6: Examples of successful lesion tracking. User segmentations are shown in cyan, the matching result is indicated by the yellow crosshairs. Left: baseline, right: follow-up.

Insufficient structural information (5 %): Template matching works only if there is some reliable structural information to locate the lesion. This is almost never a problem for lung nodules and liver metastases because the pattern "bright spot on dark background" or vice versa is sufficient. For lymph nodes, however, the patterns are much more complex and also more variable. This is especially true for lymph nodes in the cervical and mesenteric regions. In the former, individual lymph nodes are hard to delineate even visually in "packages" of similar-looking structures in the neck (Figure 7.7b). In the latter, due to the high mobility in the intestines, no reliable anatomical landmarks can be used (Figures 7.7c and 7.7d). These two lymph node regions were confirmed as being the most difficult ones by radiologists.

Strong change in size (8 %): The statistics visualized in Figure 7.5a indicated already that strong change in size, in particular strong shrinkage, is a major reason for template matching failures. The result is sometimes another lesion in the neighborhood whose size in follow-up is similar to the size of the target lesion in baseline (Figure 7.8a). This can be hard to spot both for a human and for an automatic error detector. Most of the time, however, template matching returns a point in a different structure, such as a muscle instead of a lymph node (Figure 7.8b), which should not mislead the user.

Strong change in density (2 %): While correlation is robust to changes in density, candidate detection may fail in some cases. Examples are liver lesions turning from hypodense to hyperdense (Figure 7.8c). Another frequent problem is that shrinking lesions get a relatively larger partial volume zone which leaves few or no candidate voxels in the threshold range (Figure 7.8d).

Computation time

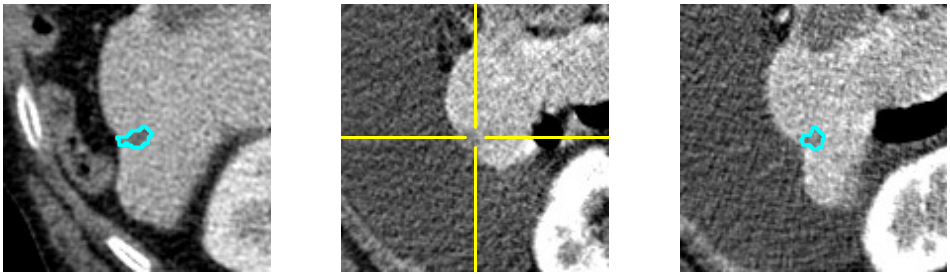
The median computation time of the template matching over all lesions was 1.07 s. A box plot grouped by lesion types is shown in Figure 7.9. The main factor is the number of candidate voxels, which is higher for lymph nodes because they are often surrounded by other structures in the same density range. The lesion size also has an impact. Since the resolution is adapted to the lesion size and the physical size of the search region is constant, there are more voxels to check for smaller lesions. This effect, however, is partially compensated by the higher cost for resampling for larger lesions, so that no clear correlation is visible.

While for 95 % of the cases computation time is less than 3.5 s, there are some outliers with a maximum of 23 s for one case. This is a huge lymph node, which is also the largest lesion in the data base (772 ml).

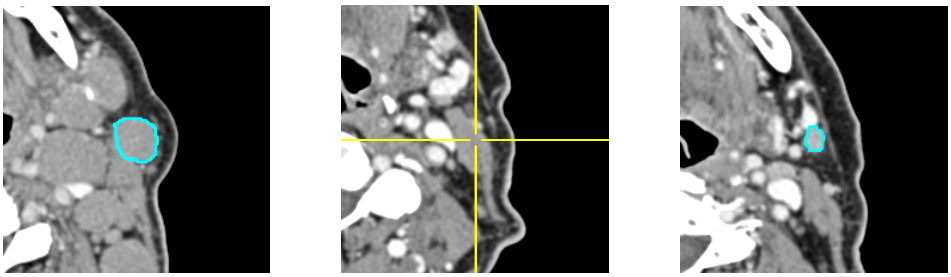
7.3.4 Discussion

This section presented the setup of a template matching procedure for lesion tracking. The optimal parametrization was derived on a data base of almost 1000 lesions, and it was shown that an additional candidate detection step based on thresholding increases the matching rate and reduces the computation time simultaneously. The method achieved a total matching rate of 80.0 % on the development data set and 83.3 % on an independent test data set at a median computation time of approximately 1 s.

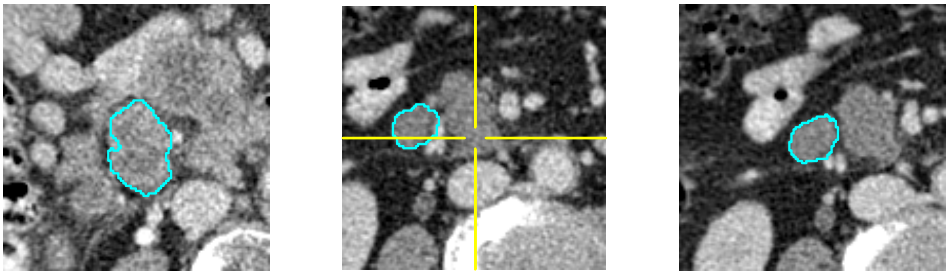
Originally, I planned to incorporate shape features into the candidate detection as well. They were meant to detect structures that have a compact, ellipsoidal shape and approximately the size



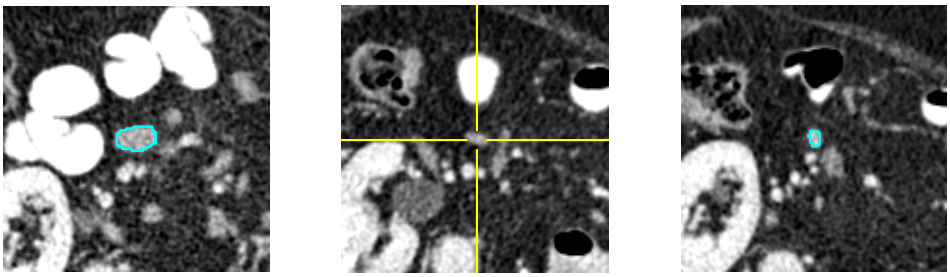
(a) Liver metastasis where registration is inaccurate due to anatomical change and the wrong lesion is found.



(b) Cervical lymph node where matching fails due to insufficient structural information.

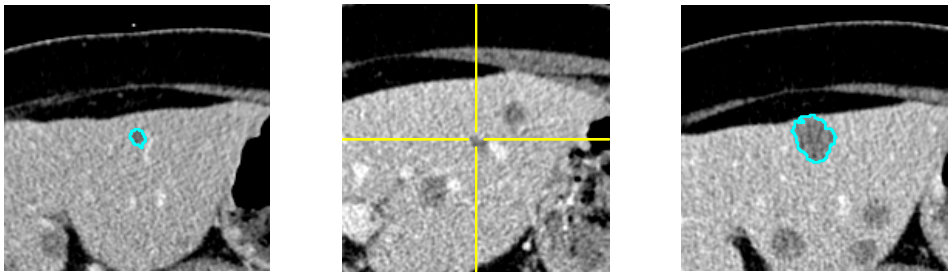


(c) Abdominal lymph node where matching fails due to insufficient structural information.

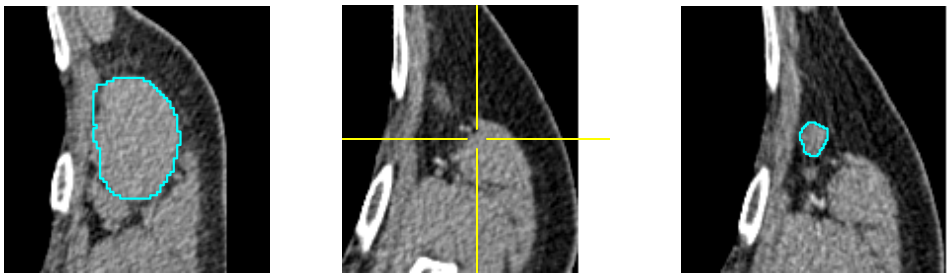


(d) Abdominal lymph node where matching fails due to insufficient structural information.

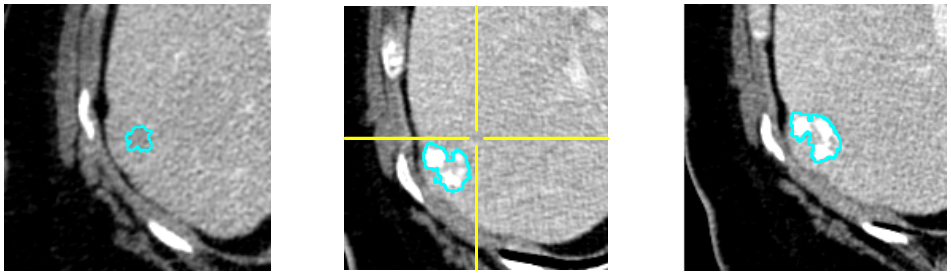
Figure 7.7: Examples of failed lesion tracking. User segmentations are shown in cyan, the matching result is indicated by the yellow crosshairs. Left: baseline, middle: algorithm result in follow-up, right: correct position in follow-up.



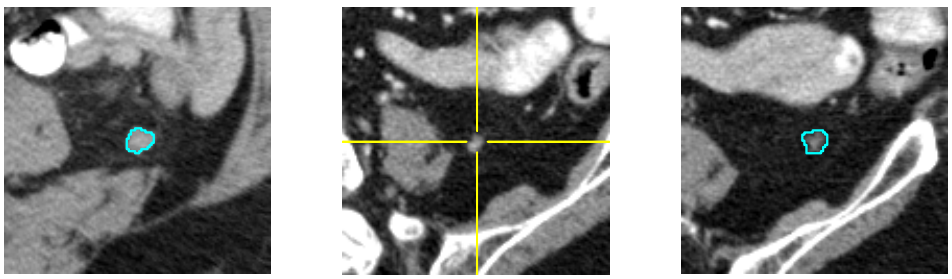
(a) Liver metastasis where matching fails due to strong growth (d_v = 0.87).



(b) Lymph node where matching fails due to strong shrinkage (d_v = −0.97).



(c) Liver metastasis where matching fails due to contrast inversion (chemoembolization) (d_c = 1.75, d_d = 121).



(d) Lymph node where matching fails due to shrinkage and enlarged partial volume zone (d_d = −116).

Figure 7.8: Examples of failed lesion tracking. User segmentations are shown in cyan, the matching result is indicated by the yellow crosshairs. Left: baseline, middle: algorithm result in follow-up, right: correct position in follow-up.

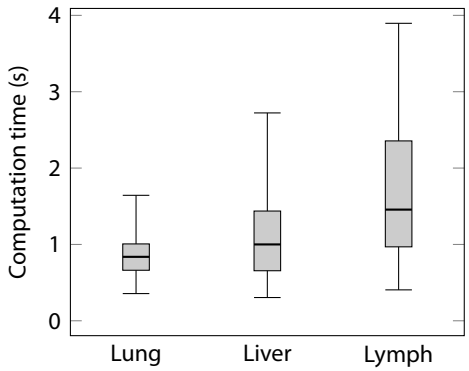


Figure 7.9: Box plot of computation times. The whiskers show the 5 % and 95 % quantiles, respectively.

of the baseline lesion. I expected this to be useful in cases where structures are present that have similar gray values as a lesion, but other geometric properties, such as blood vessels or muscles as compared to lymph nodes. I applied a Hough transform and Hessian-based blobness measures for detecting ellipses as well as local object scale for filtering by size. None of these features, however, improved the performance of the template matching significantly, whereas overall computation time increased, in some cases dramatically. Since the purely threshold-based candidate detection performed well, I did not investigate these additional features in more depth.

Another idea that I discarded during my experiments was a refinement step for the match point. Although the procedure described so far computes a point inside the lesion most of the time, this point is not always located in its center. Reasons for this include the reduced resolution which is used for template matching, the fact that the candidate detection might miss some voxels within the lesion due to noise, or changes in appearance that prevent the similarity measure from exactly detecting the lesion center. For small lesions, the match point is sometimes just outside the lesion. This can, but does not necessarily cause problems in the subsequent segmentation. In my refinement step, I started a region growing from the match point, computed a Euclidean distance map, and moved the match point to its maximum in a limited neighborhood, assuming this to be the center of the structure of interest. Unfortunately, this procedure seemed to be too simple for the cases that actually cause problems, so on average neither the matching rate nor the segmentation quality were improved.

7.4 Stroke propagation and segmentation initialization

7.4.1 Method

Template matching returns a single point inside the follow-up lesion. Since the overall goal is to automatize volumetry in follow-up, the next step is to start a segmentation algorithm. The initialization, which is normally done by the user, has to be performed automatically.

Here, I describe a procedure that has been tailored to our own segmentation methods as the one presented in Part I. They are given a stroke that should be an approximation of the longest diameter of the lesion. Once again, I rely on the similarity assumption and take the baseline

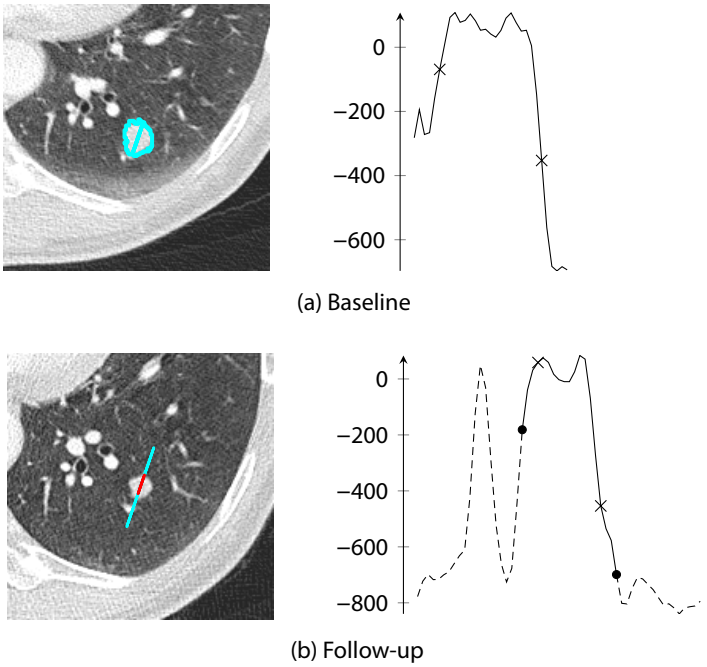


Figure 7.10: Illustration of the stroke propagation method. (a) Baseline: The stroke is extracted from the segmentation, lengthened slightly (beyond the "x" markers) and its profile is computed. (b) Follow-up: Along the lenghtened projected stroke (dashed), the part with highest correlation to the baseline stroke is found (solid). Note that in this example the resulting follow-up stroke is shorter than the baseline stroke.

stroke as a starting point. However, I do not use the original stroke by the user because it is often off-center and slightly too short or too long. Instead, I compute the longest line that fits inside the segmentation mask and passes through its center of gravity. This stroke is then projected to the follow-up image such that the old center of gravity is transformed to the result of the template matching. From this initial stroke candidate, new candidates are generated by lengthening and shortening the stroke by up to 40 % at both ends independently.

The quality of these stroke candidates is measured by computing their gray value profiles and comparing them to the profile under the baseline stroke. Again, normalized cross-correlation is the metric of choice, because it tolerates density variations. In order for these profiles to be useful, they need to include the edge at the lesion boundary, which is the main landmark for the correlation. Therefore the input stroke is lengthened by 20 % at both ends. The resulting stroke then has to be shortened by $\frac{20}{100+20} = 14.3\,\%$ to remove the overhanging part. This procedure is illustrated in Figure 7.10.

In practice, this stroke often differs from what a person would have drawn. Some further modifications to the initialization of the segmentation methods turned out to be useful. First, if the stroke is too short and the lesion is not completely contained in the ROI, segmentation will fail and even manual refinement will not be possible. In order to avoid this frustrating situation, the ROI is always at least as large as in baseline. Furthermore, if the segmentation result touches

the ROI boundary or if the erosion strength in smart opening (Section 3.1) is 100 %, which also indicates that the ROI is too small, the edge length of the ROI is doubled and segmentation is repeated.

Second, since the stroke is lengthened at both ends independently, the center of the resulting stroke may differ from the match point. This can have two effects. On the one hand, if the match point was off-center, this procedure can help to correct it and move the stroke center to the actual lesion center. But on the other hand, there are also cases where the stroke protrudes from the lesion and the stroke center is moved out of the lesion. Since this second effect would cause segmentation to fail although matching was successful, I decided to always use the match point as the seed point for segmentation rather than the stroke center.

7.4.2 Evaluation

The purpose of the evaluation is to compare the results of the segmentation algorithms using the automatically propagated stroke with those using a reference stroke. This stroke is defined as the longest diameter of the reference segmentation that passes through its center of gravity. It should be noted that this stroke does not necessarily yield the optimal segmentation, it is just the stroke that a user with the given reference segmentation in mind would most likely draw.

For all cases where matching was successful, segmentations with both strokes are computed and the volume overlaps with the reference segmentation are compared. Results are illustrated by the scatter plots in Figure 7.11. At first glance, it becomes apparent that there is little difference in segmentation quality for lung nodules, whereas the other lesion types show a clear tendency towards worse results. This can be easily explained by the properties of the segmentation algorithms. The lung nodule segmentation uses only the center of the stroke and, using fixed thresholds, does not depend on it very much. For liver metastases and lymph nodes, on the other hand, the stroke is directly used for threshold computation and marker positioning. Therefore, a moderate decrease in segmentation quality will be unavoidable.

In spite of the visual impression, a statistical analysis of the volume overlap differences reveals that the median is very close to zero for all three lesion types (Figure 7.12). Although there is a wider spread for liver metastases and lymph nodes, this means that in half of the cases the segmentation is at least as good as with a manual stroke.

One may argue that the difference in volume overlap is not always a suitable measure, because a drop from 80 % to 70 % may be a significant degradation, while two segmentations with overlaps of 30 % and 20 % may be considered equally bad by a user. Therefore, Table 7.3 takes a different approach to evaluating the results. Here a comparison is made of how many segmentations exceed a particular overlap threshold. When the automatic stroke is used, there is a notable decrease, e.g. from 78 % to 70 % if an overlap of more than 60 % is required, i.e., there will be slightly more cases where the segmentation result has to be refined manually.

Ideally, the matching point is close to the lesion center and the resulting stroke could have been drawn by a user. Figure 7.13 shows two typical cases where the stroke propagation does not work that well and causes a segmentation that is much worse than with the reference stroke. In Figure 7.13a, the matching point is perfect, but since the two lesions grow closer to each other, the stroke gets too long and protrudes into the neighboring lesion, which is then included in the segmentation. In contrast, the reason for the problem in Figure 7.13b is the matching point being

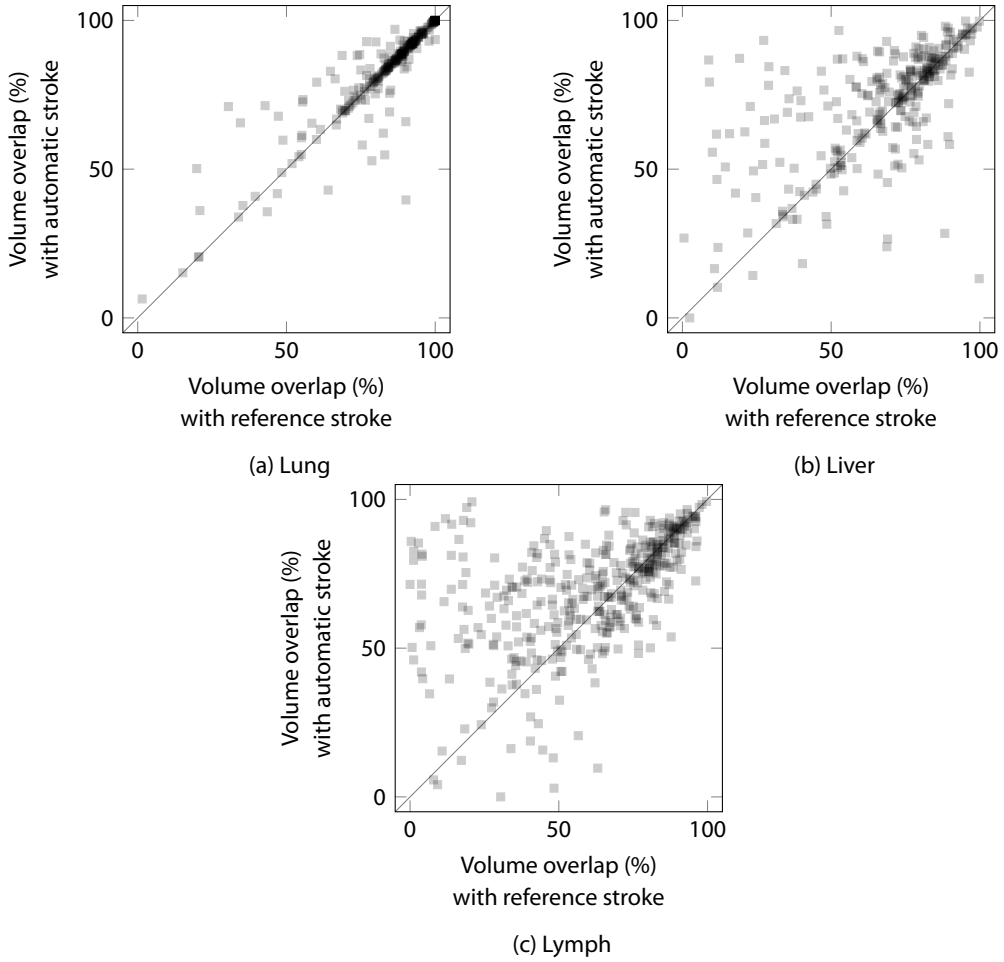


Figure 7.11: Comparison of volume overlap with automatically propagated stroke and reference stroke.

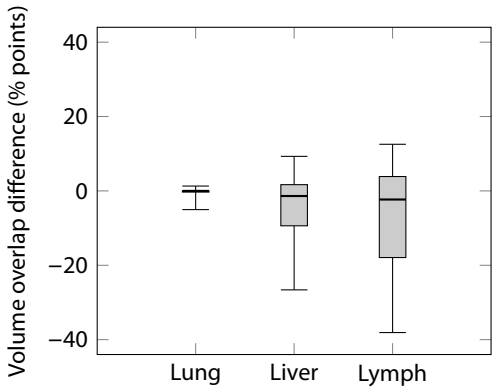
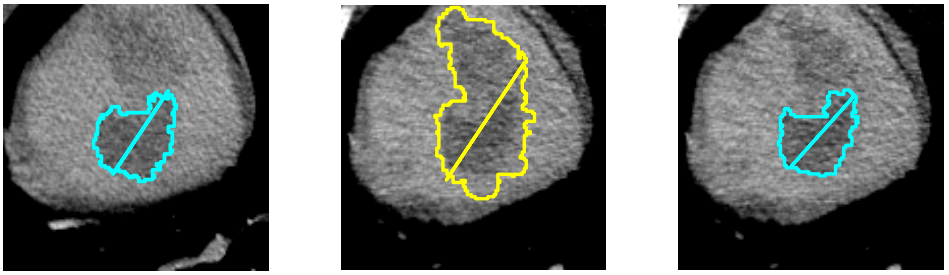


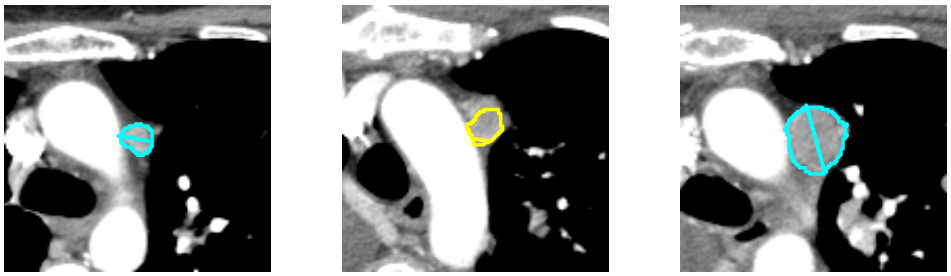
Figure 7.12: Box plot for distribution of volume overlap differences (auto - ref). The whiskers show the 10 % and 90 % quantiles, respectively.

	Frequency	
Overlap	automatic (%)	reference (%)
≥ 60 %	70.0	78.3
≥ 70 %	58.4	65.5
≥ 80 %	41.4	46.2
≥ 90 %	16.9	20.2

Table 7.3: Frequency of segmentations above a given overlap with automatic and reference stroke.



(a) Propagated stroke is too long and causes segmentation to be too large.



(b) Stroke is too short and does not pass through the actual lesion center.

Figure 7.13: Examples of bad segmentation results due to a poorly propagated stroke. Left: baseline with user segmentation, middle: follow-up with propagated stroke and segmentation, right: follow-up with reference stroke and segmentation.

too far away from the actual lesion center. In this case, the stroke is too short and the resulting segmentation covers only a part of the lymph node.

7.4.3 Discussion

In this section, a method for automatic segmentation initialization in follow-up was presented. The approach chosen here can also be regarded as a template matching, where the "template" is the intensity profile under the baseline stroke and the search region is a line in the follow-up image that is parallel to the template stroke. Experimentally, I also created stroke candidates by rotation around the matching point, but that did not improve the results. This can be explained by the fact

that stroke profiles are not rotationally invariant. While most lesions are homogeneous inside, they have various neighboring structures which make rotated profiles look different.

In this method, a decision was made to directly simulate the initialization that the segmentation algorithms typically require from a user. The advantage is that our segmentation methods are applicable without any changes. If, however, other segmentation methods with different initialization mechanisms were used instead, the propagation procedure would have to be changed as well.

This combination of identifying a single point in the lesion, simulating a user interaction based on this point, and starting a segmentation algorithm is in contrast to some procedures in the literature where the baseline segmentation is directly transferred by a registration. Such a procedure typically takes more time and we would also risk to lose the qualities of our specialized segmentation methods. On the other hand, such an approach enables a more direct transfer of information from the baseline segmentation, which is useful if lesions do not change too much. So far, this available information is reduced to the profile along the longest diameter. Improvement might possibly be achieved by propagating more than one stroke, e.g. three orthogonal ones, or by applying template matching to several points in the lesion. While this is certainly worth investigating, it should also be said that all approaches based on the similarity assumption have their limitations when lesions change strongly.

7.5 Plausibility check

Since template matching is just a maximization of a similarity function, it always gives a result, even if the lesion has vanished under therapy, is outside the search region, or there has been so much change that the similarity assumption is violated.

In these cases, the result will be wrong and can be so implausible that users may lose their confidence in the algorithm altogether. Therefore it is better to discard such results automatically rather than show them to the user. In some of these situations, it might be possible to find the correct lesion after all if some parameters of the algorithm are changed.

Since the ultimate goal is to compute a segmentation of a lesion in follow-up, we discard all tracking results where segmentation is not possible. For this, we rely on the error detection of our segmentation algorithms which will detect cases where there is *no lesion* at the detected position.

Since this is only a subgroup of potential errors, I decided to leave the remaining task to a classifier that is trained with my development data. I use the REPTree learner from the machine learning software *WEKA* (Hall et al. 2009). The resulting classifier is a pruned decision tree. The advantage of such a simple classifier is that it can be inspected by a human and checked for plausibility, for example to avoid overfitting. I also tried other classifiers to make sure that the performance of the decision tree is not significantly worse.

The features that are computed for the classifier can be divided into two groups: Some of them refer to a single pair of lesions and some to all target lesion pairs of an image pair. The latter evaluate the geometric relations between the lesions.

7.5.1 Features for lesion pairs

The following features are computed for each lesion pair individually.

- MatchSim: Similarity as computed by the template matching. This mostly helps to detect cases where a lesion has vanished or is outside the search region. If there is nothing in the search region that resembles a lesion, the similarity will be very low. Since similarity is related to change in appearance, however, a low value does not necessarily imply a mismatch.
- NbhSim: Median similarity of the 26 neighbors of the match point. If the match point is correct, it can be assumed that its neighbors also have a high similarity. Otherwise, it might just have been caused by noise. Here, neighbors which were not candidates are assigned a similarity of 0, which is the worst value in terms of correlation.
- NbhSimRank: Median similarity rank of the 26 neighbors of the match point. The rank is 0 for the match point, 1 for the second best similarity in the search region and so on. The rank transform makes the feature less dependent of the optimum similarity which varies over cases. Since the number of voxels in the search region is constant, the results for different cases are better comparable. Non-candidate neighbors are assigend a rank of ∞.
- StrokeSim: Similarity as computed by the stroke propagation. This complements the template matching well as it compares the profile across the detected structure which should also have a characteristic pattern for a lesion.
- InvMatch*: A family of features derived from inverse lesion tracking, using the result of the initial lesion tracking as the reference point. The assumption is that a correct tracking process should be revertible because the lesion does not change a lot or there are no other suitable candidates in the search region. On the other hand, if tracking failed, the correspondence is weaker and inverse tracking will probably find some other "random" point rather than the original lesion.

  Three variants of this feature were used:

  - InvMatch: Is the inverse match point contained in the baseline lesion mask?
  - InvMatchDist: Distance of the inverse match point to the baseline lesion center.
  - InvMatchMaskDist: Distance of the inverse match point to the baseline lesion mask.
- VolDiff: Volume difference between the baseline and follow-up masks. An exceptionally high volume difference is less likely caused by an actual change than by a mismatch, which might cause the segmentation to either return just a few voxels or leak all over the ROI. The classifier will learn from the training data which volume changes are still plausible.

7.5.2 Features for image pairs

The main idea of the features for image pairs is to detect geometric inconsistencies between the lesion positions. As a simple example, Figure 7.14 shows a configuration with three lesions where $\boldsymbol{q}_1$ is incorrect.

We would expect the translation vector between corresponding points to be approximately the same for all pairs. This is based on the assumption that the difference between the two images is essentially described by a translation and there are no significant local deformations. Any major deviation from this model will be regarded as a *geometric inconsistency*.

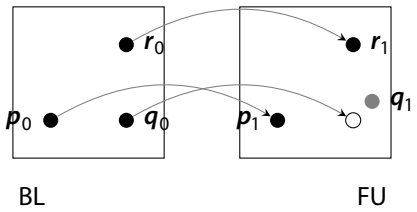


Figure 7.14: Illustration of geometric consistency criteria for several lesions in an image pair. The gray circle symbolizes an incorrect tracking result. The arrows indicate the predicted lesion positions based on the average translation.

So a family of features GeoCon* can be informally described as follows: If the position of a particular lesion is predicted by the average translation of all (other) lesions, what is the error? In the figure, the predicted position of $\boldsymbol{q}_1$ (white) deviates from its actual position (gray), while for $\boldsymbol{p}_1$ and $\boldsymbol{r}_1$ the prediction is correct. Here, the prediction is based on the median of all translations.

Different variations of this feature are possible and can be combined:

- GeoConMedian: The mean is replaced by a median.
- GeoConExcl: The lesion pair that is being tested is excluded from the averaging.
- GeoCon3NN: Averaging is restricted to the three lesions which are closest to the lesion under test.
- GeoConNorm: The computed distance is divided by the baseline lesion radius. This normalization accounts for the fact that the tolerable deviation from the predicted position depends on the lesion size.

7.5.3 Training

When training the classifier, it first has to be decided which feature should be used to mark samples as positive or negative. Possible criteria include:

1. Is the match point contained in the reference segmentation?
2. Is the center of the automatic segmentation contained in the reference segmentation?
3. Is the automatic segmentation of sufficient quality as compared to the reference segmentation?

For a user, a segmentation result is only useful if it is correct or can be corrected quickly. This would be best reflected by criterion 3. However, since a discrepancy measure has to be chosen and a binary threshold has to be defined, this leaves too many degrees of freedom as to be objective. Criterion 2 is much simpler, but also more clearly defined, and it captures quite well whether the segmentation is roughly correct. For instance, if the segmentation leaks strongly, such a case would be labeled as negative. This would not be covered by criterion 1, therefore 2 is used.

Since the geometric consistency features require at least three lesions, training was done twice: once for all cases without GeoCon*, resulting in classifier C_2, and once for the cases with at least three lesions and all features, yielding a classifier C_3.

The classifier training uses all lesions listed in Table 6.1 except the data from Kiel, using 10-fold cross-validation. An independent validation was performed on the Kiel data. Since I suspected overfitting in some of the decision trees during my experiments, I set the maximum depth to 3. This did not deteriorate the performance in cross-validation.

The classifier C_2 uses only the feature InvMatchMaskDist and achieved an F_1 score of 0.864 in cross-validation. The classifier C_3 selected InvMatchMaskDist, NbhSimRank, and VolDiff, resulting in an F_1 score of 0.882.

In my final implementation, I use C_2 for all cases. There are three reasons for this decision. First, it performs almost as well as C_3 in cross-validation. Second, it is easier to implement because it can be computed for each lesion individually without having to accumlate the results for all lesions of an image. And finally, it is preferable for the classifier output to be independent of the number of lesions segmented.

7.5.4 Evaluation

Results for the classifier performance are given by true positives (*TP*), false negatives (*FN*) and so on, where *positive* denotes a match. For the interpretation of the results, note that $TP + FN$ equals the matching rate as reported previously, while $TP + FP$ is the fraction of lesions for which a result will be displayed to the user. Three measures are derived which focus on different aspects in the evaluation of the classifier and the overall performance of the lesion tracking method:

Output rate For how many lesions is a result available? ($TP + FP$)

Output matching rate How many results that are displayed to the user are correct? ($TP/(TP+FP)$, *precision, positive predictive value*)

Efficiency How many correct results are displayed? ($TP/(TP + FN)$, *recall, sensitivity*)

As Table 7.4 shows, the measure that varies most across different lesion types – and is thus most strongly affected by the matching rate – is the output rate. In comparison, the output matching rate and the efficiency are relatively constant. This is the desired behavior according to the specification: The output matching rate is relatively high, i.e., cases that have to be corrected by the user are rare. Results for lung nodules and liver metastases are very reliable with a precision above 96 %, but also for lymph nodes the output matching rate is much higher than the total matching rate. The output rate, on the other hand, is more variable and decreases for difficult classes such as lymph nodes. The efficiency is not directly relevant for users, because they cannot measure it, but interesting from a developer's perspective. A low efficiency would mean that many correct results are "wasted" because they are not shown to the user. In Table 7.4, these numbers do not drop as low as the output rate, but "wasting" more than 10 % of the correct results for lymph nodes might still be considered unsatisfying.

Since the classifier is so simple – a threshold on a single feature –, it is possible to investigate the trade-off between the three criteria more closely. The curve shown in Figure 7.15a is known as a *precision-recall curve* in information retrieval. Here, it is important to note that the precision cannot

	Lung	Liver	Lymph		
	Development data			Test data	Vanishing
TP	90.5	83.3	61.2	76.6	0.0
FP	1.4	1.6	8.3	4.8	4.5
TN	3.6	8.5	23.1	13.9	95.5
FN	4.5	6.6	7.4	4.8	0.0
Output rate	91.9	84.9	69.5	81.4	4.5
Output matching rate	98.5	98.1	88.1	94.1	0.0
Efficiency	95.3	92.7	88.1	94.2	NaN

Table 7.4: Evaluation of the classifier. All values are given in %.

reach 100 %, but has, in this data set, its maximum at 94.7 %. This is because InvMatchMaskDist is 0 mm if inverse matching is successful and thus the feature cannot discriminate cases where this happens. On the other end of the curve, a recall of 100 % reduces the precision to 80.8 %, which is very close to the total matching rate and thus hardly better than without the plausibility check. It also becomes clear that the working point chosen by the classifier is very close to the optimal precision. Figures 7.15b and 7.15c show how the output rate depends on precision and recall.

In order to convey a better understanding of classification failures, Figure 7.16 shows some examples. The main scenario where *false positives* occur is when a lesion changes strongly and another lesion in the neighborhood becomes more similar to it than the lesion itself. In these cases, the inverse matching is likely to make the inverse error than the initial matching and lead back to the correct baseline lesion (Figure 7.16a).

Concerning *false negatives*, the main problem seems to be matching points that are close to the lesion margin. Segmentation works fine in many of these cases, but for the inverse matching errors are more likely. Figure 7.16b shows an example where the inverse matching seems to be guided more by the fat and the kidney than by the actual lesion, which has also changed its size significantly.

Vanishing lesions

The last column of Table 7.4 shows the results for the test set of 65 vanishing lesions. They have to be evaluated separately because they are negative *by definition* and the goal is to have as many true negatives as possible. Overall, three lesions are classified as correct, but the great majority of lesions are discarded. For the three false positives, which are all liver metastases, the corresponding anatomical location is identified correctly and one might argue that the lesion is still faintly visible. Examples are shown in Figure 7.17.

7.5.5 Discussion

In this section, I developed a classifier to detect and discard implausible results automatically. The classifier was trained on the 994 lesions of the development data base with a total number of 24 features (including 16 geometric consistency features which are only applied in cases where at

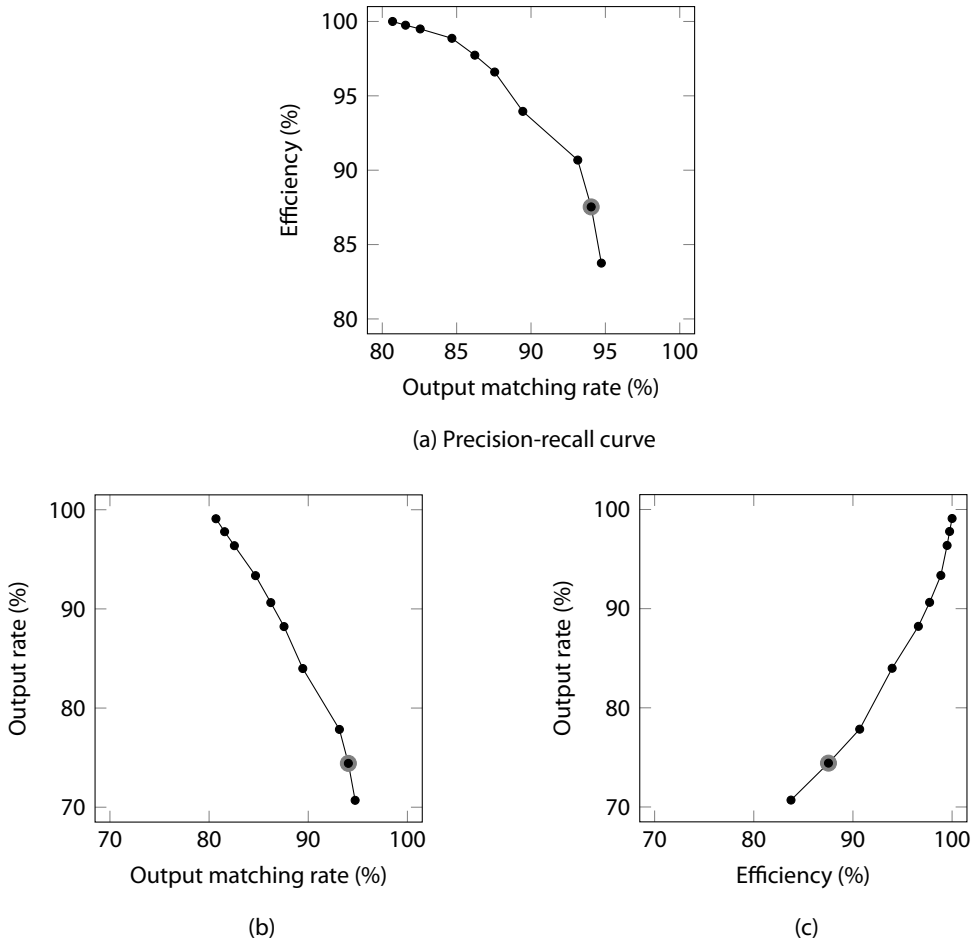
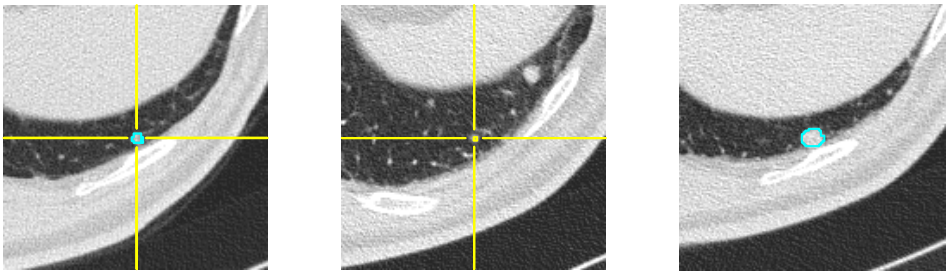


Figure 7.15: Relation of precision, recall, and positive rate for a classifier using InvMatchMaskDist. The gray overlay indicates the working point selected by the classifier.

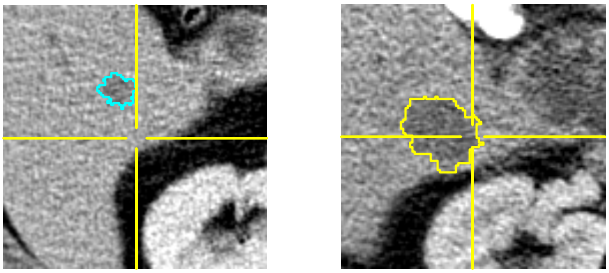
least three lesions are considered). In the end, a single feature proved to be sufficient for achieving an F_1 score of more than 0.86. Adding more features did not improve the classifier performance significantly. Also, InvMatchMaskDist was consistently selected as the first feature during my experiments with different versions of the matching algorithm and the performance was always similarly good, although the other selected features varied considerably.

The fact that the selected feature is based on inverse matching is interesting because it means that the developed algorithm has an inherent capability of checking itself. If it works correctly, it is (approximately) invertible. If the invertibility is not given, the result should not be trusted.

A limitation of the training process is the lack of knowledge about the prevalence of vanishing lesions. Since only a small set of such cases was included, the classifiers might estimate the a priori probability of positive cases too high. A more realistic training can only be conducted with data that were acquired in clinical practice.

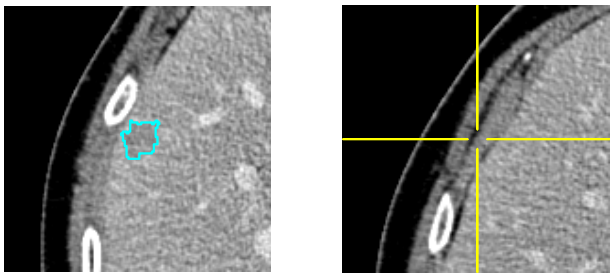


(a) False positive: Matching point is in a similar lesion and inverse matching leads back to the original lesion.

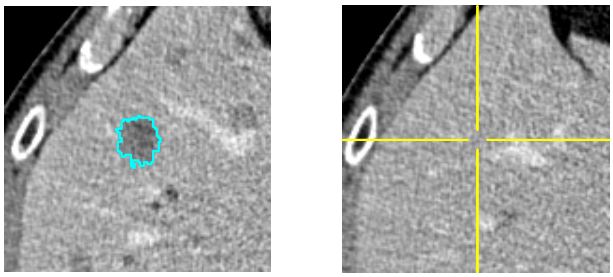


(b) False negative: Matching point close to the boundary of the lesion and inverse matching fails.

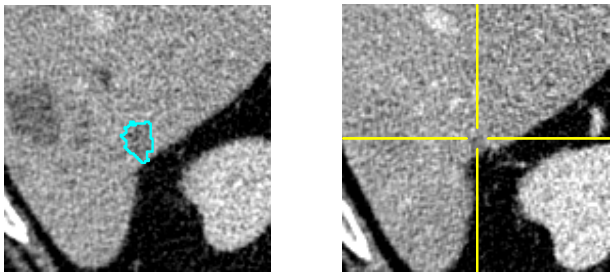
Figure 7.16: Examples of classification failures. Left: baseline, middle/right: follow-up.



(a) True negative: Matching detects a structure outside the liver.



(b) True negative: Matching detects a random noise voxel in the liver.



(c) False positive: The lesion is still faintly visible.

Figure 7.17: Example results for the test data with vanishing lesions. Left: baseline, right: follow-up.

Chapter 8

Workflow-centered evaluation

8.1 Goals

In the previous chapter, I focused on technical aspects of the performance of my algorithm. But at this point, it is still unclear whether automatizating lesion tracking has an actual benefit in clinical practice. Therefore I organized a study with four radiologists to compare the workflow of reading follow-up examinations with and without automatic lesion tracking (ALT). The study was designed such as to verify three hypotheses: With ALT,

(H1) examination time is reduced;

(H2) the inter-reader variability of the results is reduced;

(H3) the quality of the precomputed segmentations is at least as good as if they were initialized manually.

8.2 Materials and methods

8.2.1 Workflow integration

The study was performed using the *Oncology Prototype Software* (see Section 1.4). It is assumed that with a combination of the semi-automatic segmentation and the manual refinement tools it is always possible to achieve an accurate volumetric measurement.

For comparing lesions in baseline and follow-up images, two workflows are possible. When ALT is not used, the results of the baseline examination are displayed and an optional synchronization of the viewers aligns anatomically corresponding slices automatically. The user initializes the segmentation of the target lesion by drawing a stroke.

When ALT is used, it runs as a preprocessing step before a radiologist starts reading a case. The segmentation results for baseline and follow-up are immediately displayed and can be checked by the user. If a wrong lesion has been segmented, the user can discard the result and initialize a new segmentation. The algorithm contains a mechanism that discards implausible results automatically in order to account for lesions that vanish under therapy and for difficult cases with a large number of lesions or marked anatomical changes. In such cases, no pre-computed results are available.

The flowchart in Figure 8.1 illustrates the two workflows that were compared in the study. Although the level of automation is different, the user has complete control over the result in both cases.

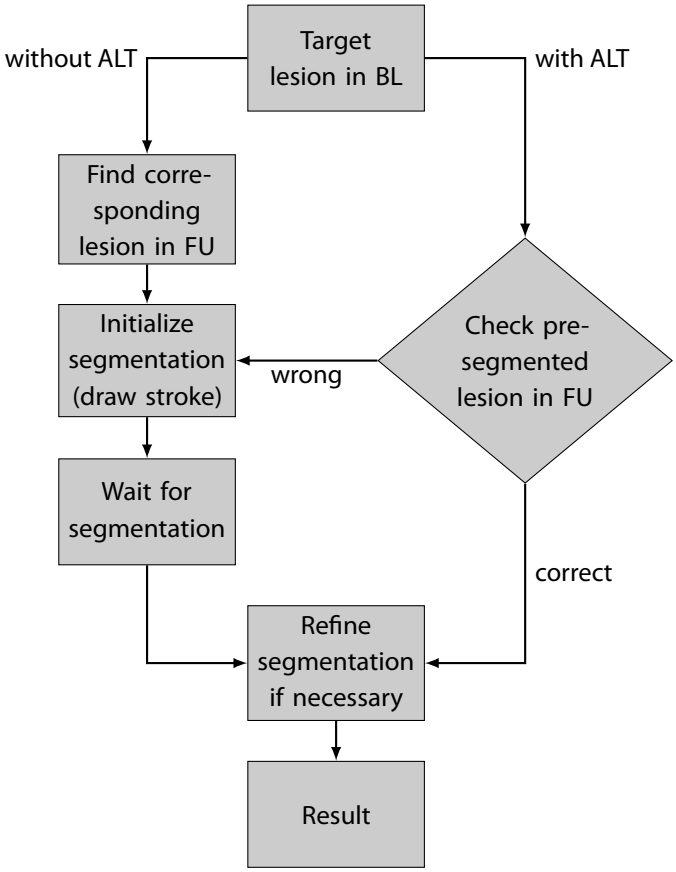


Figure 8.1: Flowchart illustrating the two workflow modes of the study.

8.2.2 Design of the study

Four radiologists participated in the study. They work at four different university hospitals with 3, 6.5, 1.5, and 6 years of experience respectively. All of them had worked with previous versions of the software, but not with the fully integrated ALT. Their task was to perform volumetric follow-up examinations for a set of patients where 1 to 5 (mean 2.7) target lesions had been segmented in the baseline scan, i.e., they had to measure the volumes of the same lesions in the follow-up CT examinations. Checking for new lesions was not required.

The data were collected retrospectively from four different sites and CT systems from three different manufacturers. CT examinations were acquired according to the local protocols, with slice thickness ranging from 1 to 3 mm. The median interval between the two CT examinations was 91.5 days. The two studies were acquired with similar CT parameters. In total, 52 follow-up pairs from 52 different patients under chemotherapy were used. In this data, 139 target lesions were selected: 47 lung lesions, 49 liver lesions and 43 lymph nodes. These lesions spanned a wide range of volumes (0.03 ml to 907.3 ml in baseline, median 1.15 ml) and volume changes (98 % shrinkage to 7900 % growth, median 6 % shrinkage). I only selected lesions which were still visually traceable

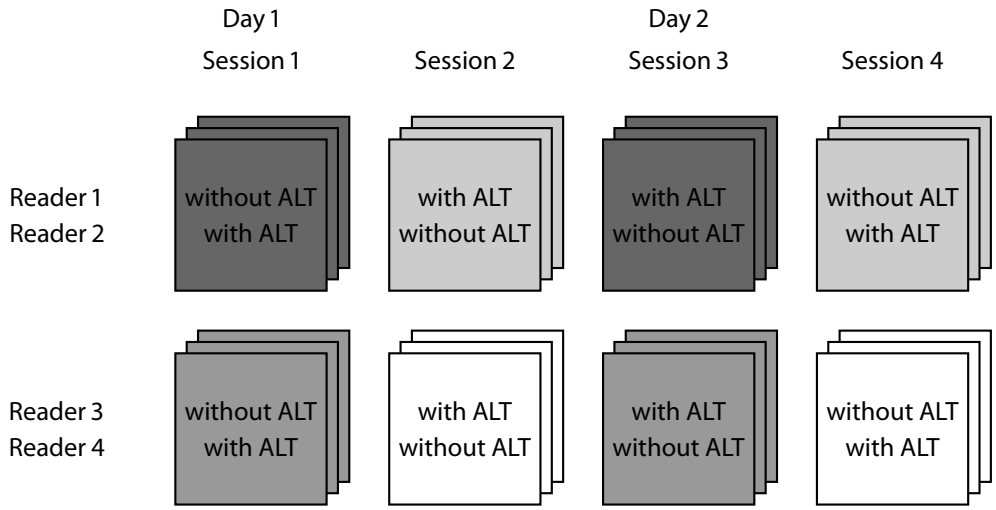


Figure 8.2: Overview of the study setup. The different gray levels indicate the four parts into which the data were divided.

in follow-up. The data for the study consisted mostly of new cases that had not been used during development.

Each case was read by two radiologists twice, once with and once without ALT. In both instances, the users had the option to refine the segmentation result with their method of choice until they deemed the segmentation quality sufficient for volumetric analysis. When ALT was enabled, the users could choose to simply accept the result, refine the contours, or delete them and start a new segmentation by drawing a stroke.

Reading times were measured by technicians sitting next to the radiologists using stopwatches. Measurement was started when the images and the available segmentation results were displayed and stopped when the radiologist had finished analyzing the examination. The final segmentation results as well as all refinement steps were saved for further analysis.

The study was performed on two consecutive days. Since I reckoned that the users would be faster on the second day, half of the cases were read without ALT and the other half with ALT on the first day and vice versa on the second day. The four readers were divided into two groups, each of which worked on half the cases. The complete study setup is visualized in Figure 8.2.

8.3 Results

8.3.1 Availability of precomputations with ALT

For 122 of the 139 target lesions (88 %), a precomputed segmentation in the follow-up CT exam was available. This included 89 % (42/47) of lung nodules, 82 % (40/49) of liver metastases and 93 % (40/43) of lymph nodes. Some examples are shown in Figure 8.3. In the remaining 17 of 139 cases (12 %), the algorithm did not compute a result, most often because there was a strong change in size or in the surrounding anatomy.

Reader	All cases			"With ALT first"		
	Reading time (s)		Time saved (%)	Reading time (s)		Time saved (%)
	without ALT	with ALT		without ALT	with ALT	
1	211	119	43.6	246	155	37.0
2	170	105	38.2	135	110	18.5
3	121	86	28.9	115	116	0.0
4	140	90	36.2	109	86	21.1
Average	159	99	38.0	151	117	22.5

Table 8.1: Statistics of reading times per patient, mean per reader.

Note that the availability of a precomputed result does not imply that the correct lesion was segmented. This will be examined in the following subsections.

8.3.2 Reading time

A graphical representation of all measured reading times for the four readers can be found in Figure 8.4. It can be seen that for all readers in the majority of the cases (82 of 104 = 79 %) lesion tracking led to faster assessment. In order to show that lesion tracking has a stronger effect on reading time than remembering the case, let us consider only those cases that were first examined with ALT. Even in this set, which has an unfavorable bias for the new method, there is a speed-up in 35 of 53 cases (66 %).

For a better quantification of the overall improvement in time taken for assessment during a reading session, I compared the average reading times per patient with and without ALT in Table 8.1. On average over all readers, the mean reading time per patient decreased from 159 s to 99 s. This means that more than one third (38 %) of the time is saved. For the individual readers, relative speed-up varied between 29 % and 44 %. Even if we once again consider the "with ALT first" subset only, there is still a speed-up per patient from 151 s to 117 s (23 %), averaged over the readers.

Here, I always give reading times per patient because we did not measure the time for the individual lesions. On average, however, times per lesion can be computed using the fact that there were 2.7 lesions per patient. This means that on average 22 s were saved per lesion. Since more than one lesion type can occur in a patient, separate reading time statistics for the lesion types cannot be reconstructed.

8.3.3 Inter-reader variability

In this study, inter-reader variability can be measured on two different levels: the choice of the corresponding lesion in follow-up and the measured volume of a particular lesion.

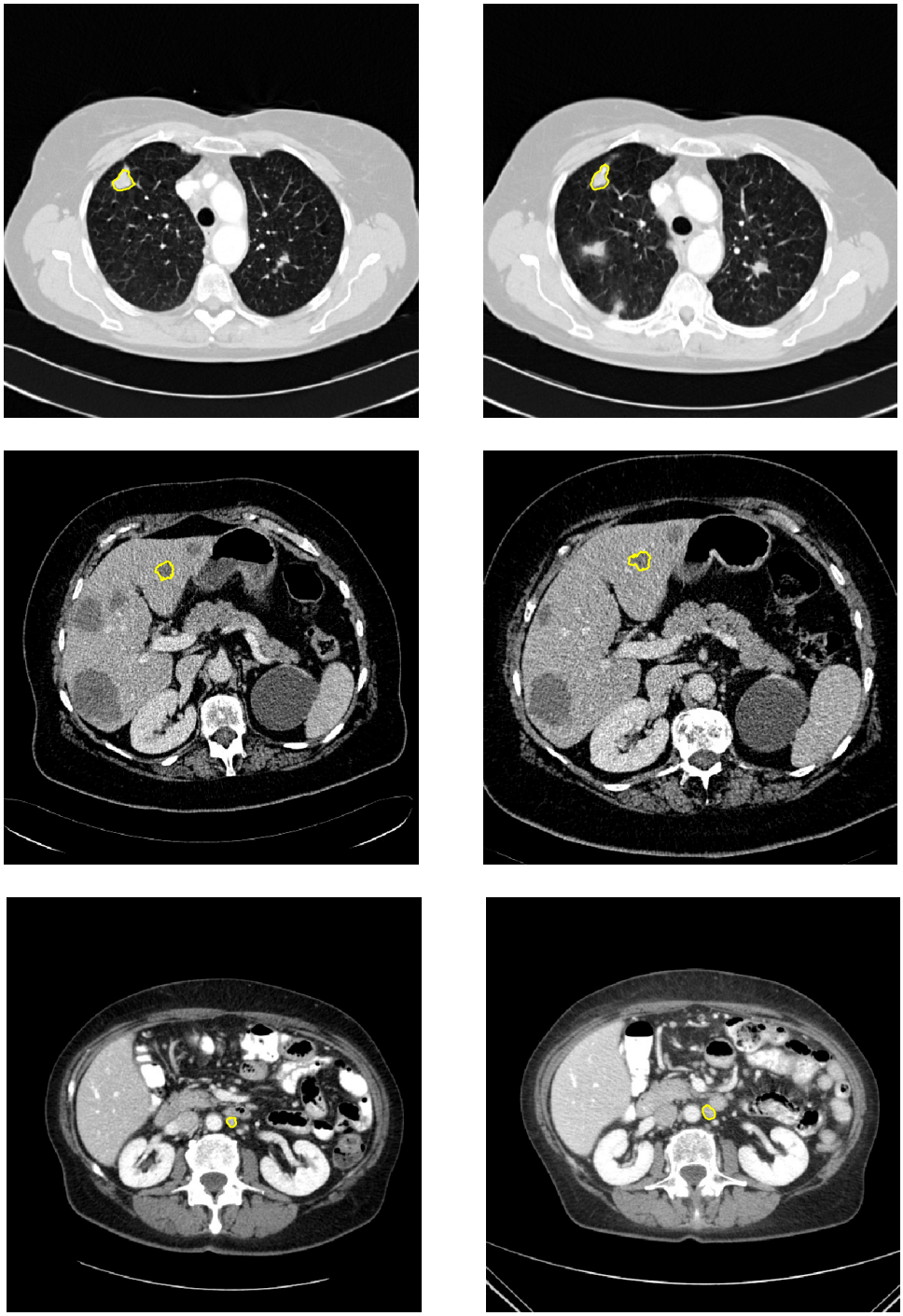


Figure 8.3: Examples of succesfully detected and segmented lesions. Left: baseline, right: follow-up.

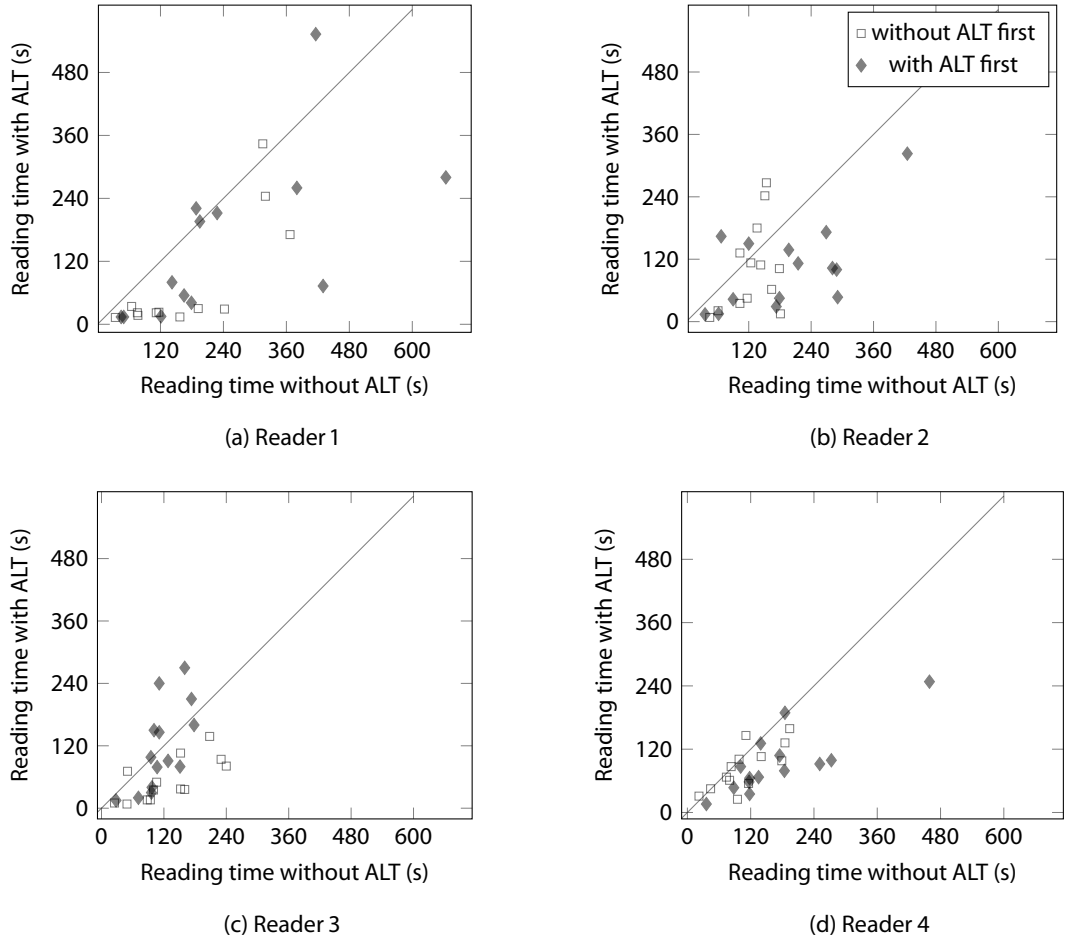


Figure 8.4: Comparison of reading times per patient. For all data points below the gray line, reading with lesion tracking was faster.

Identification of corresponding lesions

For each of the 139 baseline lesions, we had four measurements in follow-up by two readers and in two modes. Since the readers had full control over the result in both modes, we would expect that the same lesion was measured four times. For eight lesions (6 %), however, not all of the four measurements referred to the same lesion. For two of these lesions, no tracking result was available and the readers selected different lesions manually. For four lesions, the readers chose different lesions in manual mode, but both accepted the tracking result. An example of such a case is shown in Figure 8.5. On the other hand, there was one lesion where manual results agreed but one of the readers accepted a differing tracking result. Finally, there was a lesion where both readers discarded the tracking result, but one of them chose a different lesion than in manual mode.

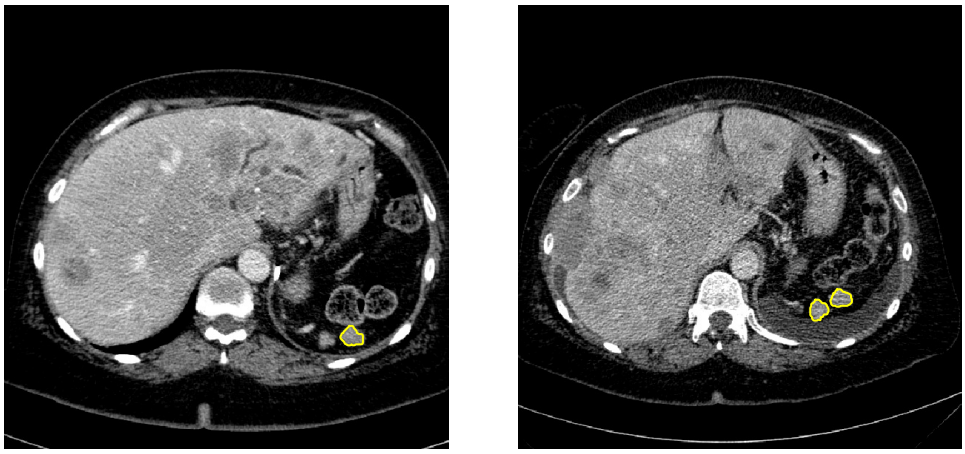


Figure 8.5: Example of a case where the readers chose different follow-up lesions without ALT, but both accepted the correct ALT result (left baseline, right follow-up).

Reader	Number of lesions		Mean variability (%)		Median variability (%)	
	total	variability decr.	without ALT	with ALT	without ALT	with ALT
1/2	55	44	18.6	5.6	6.0	0.0
3/4	62	36	15.8	13.9	8.5	2.5
Average		68.4 %	17.1	10.0	7.8	0.0

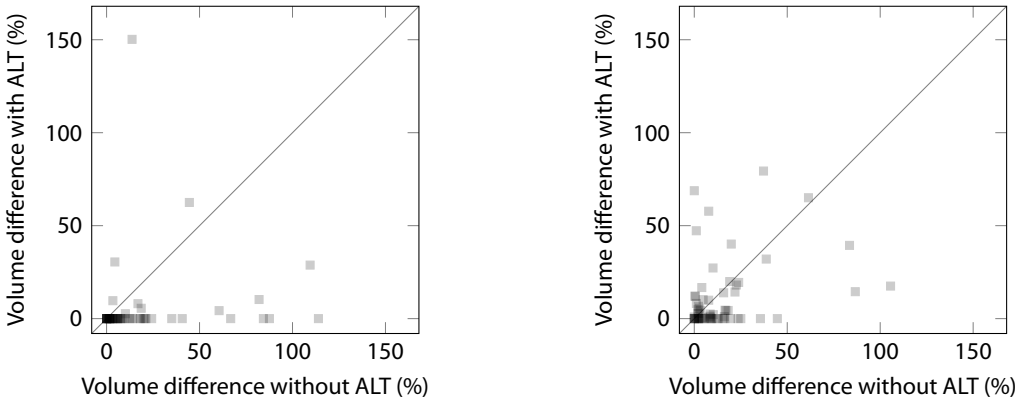
Table 8.2: Volume differences with and without ALT in the two reader groups.

Lesion volumes

The analysis of lesion volumes is restricted to those 117 lesions where an ALT result was available and accepted by at least one user and where four consistent measurements are available. The inter-reader variability of the measured volumes v_1 and v_2 was computed as the relative absolute difference

$$\frac{|v_1 - v_2|}{\frac{1}{2}(v_1 + v_2)}. \tag{8.1}$$

Figure 8.6 shows the variability per lesion with and without ALT. The variability is 0 when both readers accepted the precomputed segmentation or if their volumes were exactly equal by chance. We can see that variabilities close to 0 appear much more often when ALT is used. Still, especially for Readers 3 and 4 there is also a number of lesions where variability is increased. Overall, the mean variability decreased from 17.1 % to 10 %. Further statistical results can be found in Table 8.2. Here, it can be seen that with ALT the median variability goes down to 0, which means that in more than half of the cases the measured volumes of two readers are exactly equal.



(a) Reader 1 vs. 2

(b) Reader 3 vs. 4

Figure 8.6: Relative absolute volume difference between the two pairs of readers with and without lesion tracking. Data points below the gray line indicate that variability was decreased with ALT. Note that the highest concentration of points is close to (0,0), i.e., in both cases the volumes had a very good agreement.

Reader	Lesions	Refined cases		Refinement steps		Volume overlap (%)	
		without ALT	with ALT	without ALT	with ALT	without ALT	with ALT
1	55	11	5	38	20	92.5	92.1
2	55	19	9	117	102	89.0	88.3
3	62	24	16	85	78	87.6	89.6
4	62	31	35	145	189	84.7	86.4
Average		36.3 %	27.8 %	97.4	99.4	88.1	89.0

Table 8.3: Usage of manual refinement of the segmentation result with and without ALT. The table shows the number of refined cases, the number of refinement steps as well as the volume overlap as a measure of how much the initial segmentation was changed by the refinement.

Segmentation quality

In order to compare the quality of the manually initialized segmentations and the precomputed segmentations, I analyze the number of lesions where interactive refinement of the result was necessary and the number of refinement steps that were performed. The results are summarized in Table 8.3.

First, we can observe that the number of lesions where the segmentation result was refined was slightly lower with ALT (65 of 234 = 27.8 %) than without (85 of 234 = 36.3 %). The reduction was particularly strong for the first two readers, while Reader 4 even had a slight increase. It is also interesting that the number of cases where refinement was done differed substantially within the pairs of readers that worked on the same cases. When using ALT, both readers saw the same

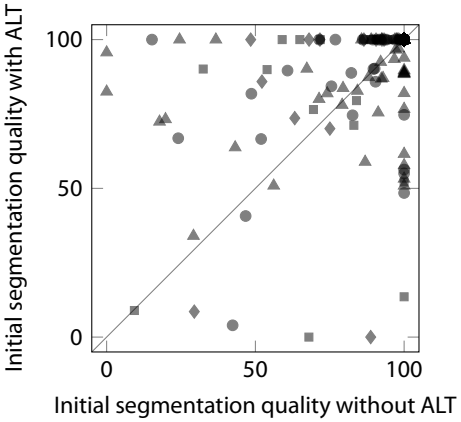


Figure 8.7: Comparison of volume overlap (agreement between automatic and corrected segmentation) with and without lesion tracking. The different symbols represent the four readers. Data points along the gray line indicate that the segmentation quality is similar. Note that the highest concentration of points is close to (100, 100), i.e., in both cases no or only minor corrections were necessary.

precomputed segmentations, but for 25 of 117 lesions (21 %) one of them decided that manual refinement was necessary and the other one did not.

Table 8.3 also shows the absolute numbers of refinement steps that were performed. Although these numbers differ among readers depending on their quality requirements and experience, the effort for achieving an acceptable result was very similar with and without ALT (1.65 and 1.66 steps per lesion, respectively). Since with ALT refinement was used in fewer cases, this means that the cases that actually were refined also needed a higher amount of change.

While the number of refinement steps is most relevant for a user, a more technical measure of segmentation quality is given by the volume overlap between the original and the refined mask. In particular, a volume overlap of 100 % indicates that no refinement was performed, while it is 0 when not a single voxel of the initial segmentation was left after refinement. The latter means that ALT chose a wrong lesion. This happened for one of the 117 lesions. I did not use volume differences here because refinement does not necessarily change the volume when some parts are added and some removed.

Figure 8.7 compares the volume overlap with and without ALT. There are some significant changes in individual cases, but on average the segmentation quality is very similar. This is also reflected in the average volume overlap, which is 88.1 % without ALT and 89.0 % with ALT.

8.3.4 Matching rate

In order to compare my method with others evaluated in the literature, I also computed the number of lesions that were correctly identified in follow-up. This is the case when a precomputed segmentation is available and was not discarded by any of the readers. As already noted, for 122 of 139 lesions a precomputation was available. There was only one lesion for which both readers discarded the ALT result. Another lesion was discarded by only one reader, but since both chose a different lesion in manual mode, we assume that the ALT result was incorrect. This means that

Statement	−−	−	0	+	++
The hit rate (identification of the correct lesion) is sufficient for clinical use.					
– Lung				1	3
– Liver			1	1	2
– Lymph				1	3
The ratio of lesions for which *no* FU lesion was found is acceptable for clinical use.					
– Lung				1	3
– Liver/Lymph				2	2
The ratio of lesions for which *a wrong* FU lesion was found is acceptable for clinical use.				3	1
The quality of the precomputed segmentation is suitable for clinical use.					
– Lung/Liver				2	2
– Lymph				1	3
Using automatic lesion tracking accelerates reading on average.					
– Lung				1	3
– Liver/Lymph					4
Using automatic lesion tracking makes me more certain to choose the correct lesion in FU.					
– Lung/Liver			1	1	2
– Lymph				2	2
The reduced inter-reader variability is a clinically relevant advantage.				2	2
I prefer having fewer lesions presegmented to having to delete wrong lesions.	1			1	2
I would like to use the automatic lesion tracking in clinical routine.					4

Table 8.4: Results of the questionnaire. Answers could be differentiated between the lesion types for all statements. −−: I do not agree at all, ++: I agree completely.

the total matching rate is 86 % (120 of 139). The matching rate is 85 % (40 of 47) for lung nodules, 82 % (40 of 49) for liver metastases, and 93 % (40 of 43) for lymph nodes.

8.3.5 Subjective assessment

In addition to the quantitative measurements presented so far, the readers filled out a questionnaire after the study. They were asked how much they agreed to certain statements on a five-point scale. As mentioned earlier, this was their first opportunity to experience the workflow with precomputed results for follow-up. The results are shown in Table 8.4.

In summary, all readers said they would like to use the automatic lesion tracking in routine clinical practice, that it accelerates reading and that it makes them more confident of having

chosen the correct lesion for follow-up. While overall assessment was very positive, it was also pointed out that there is still some potential for improvement of the method, especially with regard to the matching rate and the detection of mismatched lesions. An aspect that was discussed controversially was the algorithm's error detection mechanism, which suppresses results that seem implausible. Some of the users did not want to see any wrong results and said they did not expect the lesion tracking to work on difficult cases. Others preferred having as many precomputed segmentations as possible, even if they have to discard them.

8.4 Discussion

The goal of the study was to evaluate the clinical benefit of ALT in terms of the three hypotheses stated in Section 8.1. The results show that I could actually verify all three of them.

(H1) Reading time was reduced in most of the cases, with an average reduction of 22 s per lesion. Taking into account that the time needed to measure a lesion with semiautomatic volumetry is only about 2 s, the time saving of 22 s can be relevant in clinical workflow. Typically, up to five lesions per patient are evaluated in oncological baseline and follow-up exams, resulting in a potential average time saving of 110 s (almost 2 min).

Since semiautomatic segmentation does not take long in the first place and the time needed for manual refinement is about the same with and without ALT, it seems that time can mostly be saved in navigating through the 3D dataset in order to find the correct lesion. This is especially true in cases where it can be clearly seen that the correct lesion has been presegmented.

(H2) Inter-reader variability of volume measurements is reduced. In more than 70 % of the lesions, the precomputed segmentation was accepted by both readers and there was no variability at all. Additionally, there were five lesions where a reader accepted the ALT result but segmented a different lesion manually. If we assume the other reader to be correct, ALT has avoided a mistake in four cases and created one in the fifth case. These numbers are too low to draw any statistical conclusions, but they show that even sophisticated software support cannot always avoid that two different lesions are compared to each other in follow-up.

(H3) Average segmentation quality is comparable with and without lesion tracking, although there are also some significant differences. As expected, specific user knowledge is needed in some cases to draw an accurate stroke to initialize the segmentation. However, there were also cases where a better result was achieved with a stroke that was propagated automatically from the baseline segmentation.

Previous publications on automatic lesion tracking methods mostly evaluated a particular algorithm on a technical level, without considering workflow aspects. They showed that lung nodules can be tracked successfully by an algorithm on various kinds of images. In oncological data, matching rates of 86.3 % (Beyer et al. 2004) and 66.7 % (Lee et al. 2007) were reported. An evaluation on screening data gave a matching rate of 92.7 % (Tao et al. 2009). A detailed analysis showed that errors are mostly caused by differing inspiration levels and pathological changes in the lungs. Therefore the selected data can have a strong influence on the result of a study. This is confirmed by the fact that the three mentioned publications achieved such different results although they are based on the same software. Screening data are typically easier to handle by an algorithm because there is less change than in patients undergoing chemotherapy. Our algorithm

showed a similar performance to previous approaches with an overall matching rate of 86 %, with the difference that the same method was applied on three different lesion types.

A high matching rate alone, however, does not guarantee an actual practical benefit. It is not obvious whether a partial automation of a task can really reduce the overall time. Therefore it is important to evaluate a software tool in the workflow context where it is used. One such study has been performed in the context of lung cancer screening (Beigelman-Aubry et al. 2007). In this study, the impact of a computer-aided diagnosis tool on the detection of growing lung nodules was analyzed. The software offered automatic detection and tracking of nodules and proved to increase sensitivity significantly without compromising reading time. Koo et al. (2012) compared the time needed for matching lung nodules in both screening and diagnostic scans with and without software support. They found that automatic lesion tracking was faster in 94 % of the cases and saved an average 2.3 min per patient for matching up to ten nodules.

After these promising results for screening, to my knowledge, I conducted the first study to demonstrate that automatic lesion tracking may also show a benefit in chemotherapy monitoring. This should be further investigated with a larger group of radiologists in a real clinical setting. Checking the exams for new lesions and other relevant findings should additionally be taken into account. Another aspect is hard to evaluate in a study but may be important in practice: In a clinical setting, radiologists are often distracted or interrupted during work and have to regain orientation in the image several times. In this situation, the advantages of a software support for lesion tracking may be even more significant.

So far, I used only chemotherapy data and did not investigate the capabilites of my method in the context of other therapies or screening. Also, I only used data with a relatively high resolution, although a slice thickness of 5 mm is still common. These are tasks for future studies.

In spite of these limitations, I can conclude that the results are promising. I was able to show that automatic lesion tracking has the potential to save examination time, reduce inter-reader variability and increase user satisfaction. This justifies further efforts to automatize tedious procedures in the radiological workflow.

Part III

Uncertainty-aware validation of segmentation algorithms

Contributions An adaptive scoring system for validating segmentation algorithms using multiple reference segmentations.
An analysis of the variability in manual delineation of liver lesions in CT, involving ten experts, using a novel methodology for measuring the variability within a set of segmentations.
A concept and an evaluation study for a tool that allows experts to generate probabilistic reference segmentations.

Acknowledgments The work described in this part was inspired by discussions with numerous colleagues, most notably Jan Rühaak and Stefan Braunewell, as well as Frank Heckel, Sebastiano Barbieri and Lennart Tautz. The implementation and parts of the evaluation of the tool for probabilistic reference segmentations were done by Christiane Steinberg during an internship. The manual liver lesion delineations were created by Christiane Engel, Benjamin Geisler, Markus Hittinger, Michaela Jesse, Ulrike Kayser, Andreas Kießling, Nikolaas Kohlhase, Daniel Pinto dos Santos, Michael Püsken, Asmus Wulff, Gülsen Yanc, and Susanne Zentis.
This work was funded in part by the European Regional Development Fund.

Publications The adaptive score was introduced at *SPIE Medical Imaging 2011* in Lake Buena Vista (Moltz et al. 2011a). The variability analysis for liver lesions has been presented at the *IEEE International Symposium on Biomedical Imaging 2011* in Chicago (Moltz et al. 2011b). The tool for probabilistic reference segmentations was presented at *Medical Image Understanding and Analysis 2013* in Birmingham (Moltz et al. 2013) and received a *Bursary Award*.

Chapter 9

Validation using multiple reference segmentations

9.1 Introduction

The development of segmentation algorithms for different anatomical structures and imaging protocols is one of the central tasks in medical image analysis. In the previous parts, I have shown examples from my own research and from the literature. The validation of these algorithms, however, is often treated as a subordinate task, and the community has just started to establish standard methodologies.

In software engineering, the term *validation* is used for "the process of evaluating software at the end of the software development process to ensure compliance with software requirements" (IEEE 1990) and "determining the fitness or worth of a software product for its operational mission" (Boehm 1984). In the field of medical image analysis, however, the requirements are often not well defined. In contrast to other software development tasks, it is not possible to specify exactly which output is expected for a given input. Different users may expect different results, and different results may be accepted by the same user. So the desired behavior of an image analysis software could be described as generating an output that is acceptable for as many users as possible, but it is not clear how this should be measured. These issues make validation a particular challenge in this field.

Still, validation should be regarded as an integral part of all phases of medical image analysis research. During development, different approaches are compared, parameters are optimized, and sometimes algorithms are automatically trained on example data. All of this requires some means of assessing the quality of a segmentation result. Once a promising method has been found, developers need to make sure that further changes do not degrade the performance that has already been achieved. This is known as *regression testing* and especially important when algorithms have already been shipped as a part of a commercial software and are further developed. Finally, making a statement about the quality of a method is an essential part of sound scientific work. In all the phases mentioned so far, quality assessment is mostly relevant for researchers and programmers.

On the next level, it is important to be able to compare different algorithms and decide which one is most suitable for a particular problem. This is of interest for clinicians, who decide whether to use a particular software, and for product managers, who may want to include an algorithm into a commercial software package. Still, it is often up to the developers to conduct the required experiments to convince potential customers of the quality of their algorithms. Of course, comparing different approaches to a problem is also interesting in itself from a scientific point of view. In the last couple of years, many researchers have participated in *challenges* and let their algorithms compete with others.

This part will focus on the validation of segmentation algorithms such as the one introduced in Part I. This means that segmentation denotes the *delineation* of a specified object rather than partitioning an image into several segments. Also, I will only deal with algorithms that produce a *binary* segmentation, where each voxel is classified as either object or background, as opposed to *fuzzy* segmentation, which assigns each voxel a probability of being part of the object.

Following the terminology by Zhang (1996), I will focus on *empirical* validation methods that look at results on test data rather than *analyze* theoretical properties of an algorithm. In my studies I use *clinical data* because in my experience the behavior of an algorithm on *artificial data* allows only very limited conclusions about the applicability in the clinic where a huge diversity of cases can occur.

Although validation methodology in segmentation is far from being standardized, there is a general consensus about the aspects that should be considered. As summarized by Udupa et al. (2006), three criteria should be taken into account. *Accuracy* measures the agreement between the segmentation and the true extent of the object. *Precision* or *reproducibility* denotes the agreement between repeated segmentations. *Efficiency* captures the time or resources needed for the segmentation.

These criteria apply not only to segmentation algorithms, but also to manual segmentation. Thus, they motivate why segmentation algorithms are used in the first place. Typically, computers offer a higher precision and efficiency than humans can achieve when it comes to data processing. This can be verified for a particular task by simple experiments. Accuracy, on the other hand, is a much more complicated issue.

Measuring accuracy requires knowledge of the truth, but a true segmentation is only available for phantoms. In clinical data, the correct result is unknown. It is common sense to use manual segmentations by experts as a *surrogate of truth* and require an algorithm to compute a similar result. The problem with this approach is that experts are neither efficient nor precise. Outlining a liver lesion on all slices of a CT image can take several minutes, and when it is done twice, either by the same or two different persons, the results will virtually never be exactly equal. There is always *uncertainty* about the true segmentation.

This uncertainty is often ignored and a single expert segmentation is considered to be the "ground truth". Often this is the only feasible approach due to limited expert resources, but one should realize that this will have an effect on the validation outcome. In recent years, awareness of this uncertainty has risen and it has become a much-discussed problem. Multiple reference delineations are acquired in order to increase the reliability of the results. There are different ways to deal with the additional information that they provide. An obvious approach is to fuse the masks into a single one, which should then give a better estimate of the true segmentation. Techniques for mask fusion range from simple voxel-wise majority voting to more sophisticated methods like *STAPLE* (Warfield et al. 2004) or *shape-based averaging* (Rohlfing and Maurer Jr. 2007).

Currently, a new point of view is emerging: to abandon the idea of a "ground truth" altogether, to accept the uncertainty and even regard it as a source of information. The variability between expert segmentations can be used to calibrate the expected quality of a segmentation algorithm. This does not only make validation fairer, it also allows an easier interpretation of the results and should be a better reflection of the way experts would perceive the quality of the method.

The remainder of this part is guided by a number of questions that arose during my efforts to make reliable statements about the accuracy of the algorithm presented in Part I:

- What is the effect of using a single reference segmentation on validation results?
- What is a suitable validation measure that incorporates the uncertainty of the true result?
- How many expert segmentations should be acquired?
- How can the necessary number of experts be reduced?

These questions are investigated using the problem of liver tumor segmentation in CT as an example. For this purpose, additional manual delineations have been acquired. Although the analysis was performed on a single problem class, the results should generalize well.

9.2 Related work: A critical review

In order to get an overview of how validation is usually done in the literature, I reviewed publications on liver tumor segmentation in CT from the last ten years with regard to the validation techniques that were applied to assess the quality of the algorithms. A total number of 17 papers was found to provide a substantial evaluation, including the editorial of the *3D Liver Tumor Segmentation Challenge* at MICCAI 2008 which provided an evaluation framework for all participants. Table 9.1 gives an overview of these works, including the number of liver lesions segmented, the number of manual reference segmentations, and the evaluation measures. For the definitions of these measures, refer to Heimann et al. (2009) and the publications mentioned in Table 9.1.

Often, a single reference segmentation is used for accuracy analysis. In papers that incorporate multiple references, many develop their own validation scheme because for a long time no standard had been established. Only recently, owing to several segmentation challenges at MICCAI conferences since 2007, a methodology has been more widely adopted. These challenges contributed to spread the awareness of the importance of validation and the problems associated with it. The framework they provided is a good step towards objective and meaningful evaluation, but it still has some drawbacks which will be discussed in the next section.

Let us first have a look at those papers that did not use the MICCAI framework. Yim and Foran (2003) compared the reproducibility of manual and semi-automatic area measurements from repeated reading by the same expert. Popa et al. (2006) obtained four reference segmentations by different readers and estimated a ground truth using the *STAPLE* algorithm. Their principal approach is stated as follows: "We defined an accurate result as where the semi-automated segmentation measurements and comparison values are similar with the measurements and comparison values of one radiologist." To evaluate this, comparison metrics were computed for one of the references and the algorithmic result versus the estimated ground truth. Zhao et al. (2006) computed *concordance correlation coefficients* (CCC) for volume measurements of three readers and their algorithm as well as an overall CCC to assess the accuracy of their results. They also analyzed the concordance in the three measurements of one reader. Ray et al. (2008) used references of four readers from three reading sessions each. They analyzed the development of the volume measurements between the sessions and compared the measured volumes with those from the

Publication	# lesions	# reference seg.	Volume overlap	Volume error	Surface distance	False pos./neg. ratio	Remarks
Yim and Foran (2003)	10	2 · 1		•			2D
Popa et al. (2006)	4	4	•	•	•		STAPLE
Zhao et al. (2006)	59	3 · 1 + 2		•			
Cai et al. (2007)	63	1		•			
Li and Jolly (2008)	15	1		•	•		
Jolly and Grady (2008)	159	1	•	•			
Massoptier and Casciaro (2008)	38	1	•			•	
Deng and Du (2008)	30	2	•	•	•		MICCAI framework
Ray et al. (2008)	13	3 · 4		•			
Szilágyi et al. (2009)	3?	1	•	•			
Behnaz et al. (2010)	10	1		•			
Smeets et al. (2010)	61	2	•	•	•		MICCAI framework
Zhou et al. (2010)	37	2	•	•	•		MICCAI framework
Drechsler et al. (2011)	7	1	•		•		
Su et al. (2011)	29	4					NPRI, multiple methods
Häme and Pollari (2012)	31	2	•	•	•		MICCAI framework

Table 9.1: Summary of validation methods in recent publications on liver tumor segmentation in CT. For the number of reference segmentations used, "$x \cdot y$" denotes x sessions and y readers, where $x = 1$ if not stated otherwise.

algorithm in a qualitative way: They checked whether the algorithmic volume was within the range of the manual volumes. A comparison of multiple segmentation methods was performed by Su et al. (2011). The *normalized probabilistic rand index* (NPRI) was used in order to take the variability of four reference segmentations into account. It measures the pixel-wise agreement of labels and is mostly used for multi-label segmentation.

9.2.1 MICCAI Grand Challenge framework

The *MICCAI Grand Challenge framework* was developed for the *Segmentation of the Liver* challenge in 2007 and described in this context by Heimann et al. (2009). Deng and Du (2008) adapted it for the *3D Liver Tumor Segmentation Challenge* in 2008. Its main contribution is the fact that it melts several comparison metrics into a single score with a predefined range of 0 to 100. The five metrics used are volume overlap error (1 – volume overlap), relative volume difference, average symmetric surface distance, root mean square symmetric surface distance, and maximum symmetric surface distance (Hausdorff distance). In order to allow averaging of these metrics, they are calibrated

with regard to typical values that a human observer could achieve. For this purpose, two reference masks are acquired for each case. One mask M_1 is confirmed by both readers as correct and is then used as the "ground truth" to compare the algorithmic segmentations with. The other mask M_2 is used to compute a reference value for each of the metrics. This reference value is the deviation of M_2 from M_1, averaged over all cases. For an algorithmic segmentation A, a score of 100 is given if A completely agrees with M_1. If A has the same discrepancy from M_1 as M_2 has from M_1, the score will be $100 - \alpha$. The value of α was initially 25, but changed to 10 for liver tumors.

Formally, let $\varepsilon(A, M_1)$ be the discrepancy between masks A and M_1 according to a particular metric and $\bar{\varepsilon}$ the constant reference value. Then the score is defined as

$$\phi_{\mathrm{MICCAI}}(A, M_1) = \max\left(100 - \alpha \cdot \frac{\varepsilon(A, M_1)}{\bar{\varepsilon}}, 0\right). \tag{9.1}$$

The essential part is the quotient, while the rest of the equation is just a normalization. The choice of α is only significant for clamping at zero, i.e. errors that are more than $100/\alpha$ times as large as the reference value will all be given a score of zero. The total score is the mean of the five scores obtained from the different metrics. Note that this is in effect a comparison with a single reference mask. M_2 is only incorporated into the global reference value $\bar{\varepsilon}$ for each metric. In the remainder of this part, I will refer to ϕ_{MICCAI} as the *MICCAI score*.

The advantage of this framework is that it combines different metrics into a single score that is easy to interpret because it relates the actual accuracy of an algorithm to the accuracy that "can be expected". Thereby, it is an important step towards a more objective validation of segmentation results. Furthermore, the fact that it was used in several challenges spread the awareness of validation problems in the community. Still, the MICCAI score does not account for the uncertainty of human measurements completely.

One problem is that it can only be used for lesions where the readers are able to agree on a "perfect" segmentation. The other cases were discarded for the MICCAI challenges, so the problem of validating segmentations where expert opinions differ and no actual ground truth is available was left aside. In practice, however, such cases are not rare, so the validation result may sometimes depend strongly on which reference mask is chosen.

Figure 9.1 shows a graphical example where one reference mask is just a dilation of the other, i.e., one reader produced a mask that is systematically larger. Although this scenario is simplified for illustration, results like this can easily occur when the two readers use different window settings or when the contrast between the object and the background is low. In the figure, the gray area covers all masks that would get a score above $100 - \alpha$ if the segmentation drawn in red was used as the "ground truth". The areas with high scores are significantly different depending on the choice of the "ground truth" as is shown in Figures 9.1a and 9.1b. Given the information from both reference masks, a symmetric distribution as in Figure 9.1c would be more desirable. If both segmentations were drawn by independent experts, there is no reason to prefer one of them.

But even if the readers are able to agree on a "ground truth", there is still a degree of uncertainty about the true segmentation which is reflected in the independent segmentation of the other reader. This information, however, is effectively discarded by the MICCAI scoring system since the direction of the deviation is not taken into account. Since the two segmentations are not treated equally, the first one is more likely to be accepted or only slightly modified even if the second

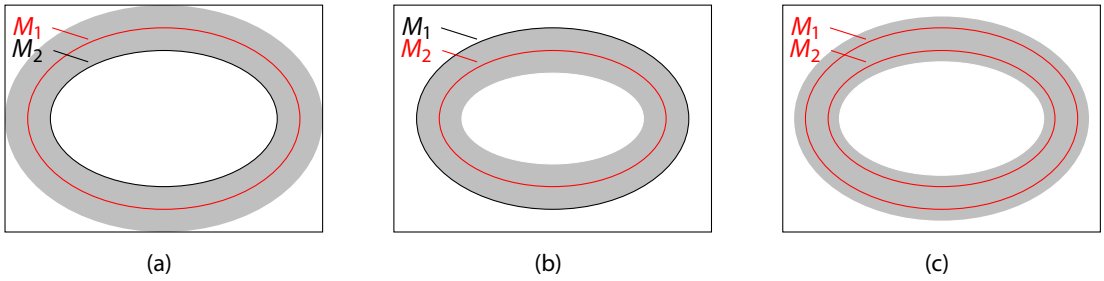


Figure 9.1: Graphical illustration of some problems with the MICCAI scoring system. Contours denote manual segmentations, red ones are used as the "ground truth". The gray area covers the segmentations that would get a score above $100 - \alpha$. (a) M_1 is used as the ground truth. (b) M_2 is used as the ground truth. (c) M_1 and M_2 are treated equally.

reader produced a different result initially. This effect has been investigated by Sensakovic et al. (2010) and often observed by myself.

Furthermore, while the framework aims at taking inter-expert variability into account, it does so on a global level only. The reference values for transforming the comparison metrics into scores have been determined empirically from a set of test data of the particular segmentation problem. This implicitly assumes that the degree of variation between different experts is independent of the specific objects that are segmented. While this assumption may be justified for organ segmentation, in tumor segmentation factors such as size, contrast, noise and the anatomical position have an impact on the variability and also on the segmentation quality that a clinician would expect. In particular, the interpretation of surface distance measures depends on the size of the object, which varies considerably for tumors. Using global reference values here will produce misleading results.

9.2.2 Williams' index

If more than two reference segmentations are available and all of them should be treated equally, a possible approach is to think of *Williams' index*. It was introduced by Williams (1976) as a general statistical tool to compare the agreement of a single rater with a group and the agreement within that group. Agreement denotes an equal classification of a particular item. Let $P_{i,j}$ be the agreement between raters i and j and rater 0 be the one to be evaluated. Then Williams' index is defined as

$$I_0 = \frac{\frac{1}{n}\sum_{j=1}^{n} P_{0,j}}{\frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{k=j+1}^{n} P_{j,k}}. \tag{9.2}$$

Let p be the probability that two random raters (excluding 0) give the same classification. Then the probability that rater 0 agrees as well is $I_0 \cdot p$.

It is not immediately clear how this method can be applied to the problem of segmentation validation while keeping up the statistical reasoning behind it. Chalana and Kim (1997) defined a modification of the index where they replaced the proportion of agreement by the inverse of the Hausdorff distance between the two segmentations. This way, however, the probabilistic meaning of the index is lost and the inverted distances do not have an intuitive interpretation.

9.3 A score for uncertainty-aware validation

The previous section discussed existing approaches for validation using multiple reference segmentations and highlighted their shortcomings. Based on the observations made there, I propose the following three criteria for a consistent validation framework:

1. all reference masks are treated equally a priori and no consensus between the experts is demanded, because it is not always possible or would distort the results;
2. the algorithmic performance is evaluated in relation to the inter-reference variability per case, i.e., more tolerantly in cases where the experts disagree about the true segmentation and more restrictively where they concur;
3. the results are comparable for different test data, possibly even for different segmentation problems.

In this section, I will introduce a new validation concept that was constructed such as to fulfill these criteria as far as possible. It is based on the MICCAI score, which should make it easy to understand and help to spread it in the community. It is independent of the segmentation problem, but liver tumor segmentation in CT will serve as an example.

The MICCAI score and Williams' index, defined in Section 9.2, are two concepts that complement each other well. There is already a structural similarity in the definitions, which allows to interpret both approaches as special cases of a general validation paradigm. Rewriting Equation (9.2) in the notation of Equation (9.1), it becomes

$$\frac{\frac{1}{n}\sum_{j=1}^{n}\varepsilon(A,M_j)}{\frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{k=j+1}^{n}\varepsilon(M_j,M_k)} = \frac{\hat{\varepsilon}(A;M_1,\ldots,M_n)}{\bar{\varepsilon}(M_1,\ldots,M_n)} = \frac{\text{discrepancy value}}{\text{reference value}}. \tag{9.3}$$

The resulting formula has two major differences compared to the MICCAI score. Starting from Equation (9.1), a single comparison of A and M_1 was replaced by an average discrepancy between A and $\{M_1,\ldots,M_n\}$ which is denoted by $\hat{\varepsilon}$. Furthermore, the reference value is now a function of the reference masks. So the definition of a new adaptive score becomes

$$\phi_{\text{adaptive}}(A;M_1,\ldots,M_n) = \max\left(100-\alpha\cdot\frac{\hat{\varepsilon}(A;M_1,\ldots,M_n)}{\bar{\varepsilon}(M_1,\ldots,M_n)},0\right). \tag{9.4}$$

For the subsequent experiments, the same five metrics as in the MICCAI score are used and averaged to compute the final score. In concordance with the MICCAI liver tumor segmentation challenge, α is set to 10. This means that a score of 90 is given to a segmentation result that is considered as good as a manual delineation.

Equation (9.4) is a general formulation that requires computing discrepancy measures between sets of masks. Since there is no standard way to do this, different implementations are possible. I tested two variations which are illustrated in Figure 9.2. The first variation averages pairwise

comparisons between masks:

$$\hat{\varepsilon}_{\text{pair}}(A; M_1, \ldots, M_n) = \frac{1}{n} \sum_{j=1}^{n} \varepsilon(A, M_j); \tag{9.5}$$

$$\bar{\varepsilon}_{\text{pair}}(M_1, \ldots, M_n) = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \varepsilon(M_j, M_k). \tag{9.6}$$

This is similar to the approach proposed by Chalana and Kim (1997), but it is more general and has a clearer interpretation.

The second implementation uses comparisons with a ground truth estimate GT from the reference masks:

$$\hat{\varepsilon}_{\text{GT}}(A; M_1, \ldots, M_n) = \varepsilon(A, GT(M_1, \ldots, M_n)); \tag{9.7}$$

$$\bar{\varepsilon}_{\text{GT}}(M_1, \ldots, M_n) = \frac{1}{n} \sum_{j=1}^{n} \varepsilon(M_j, GT(M_1, \ldots, M_n)). \tag{9.8}$$

The ground truth estimates can again be computed with different methods like majority voting, STAPLE, and shape-based averaging. I used majority voting in the experiments, because it is the simplest method and does not introduce any artificial effects. The two respective scores will be denoted by ϕ_{pair} and ϕ_{GT}.

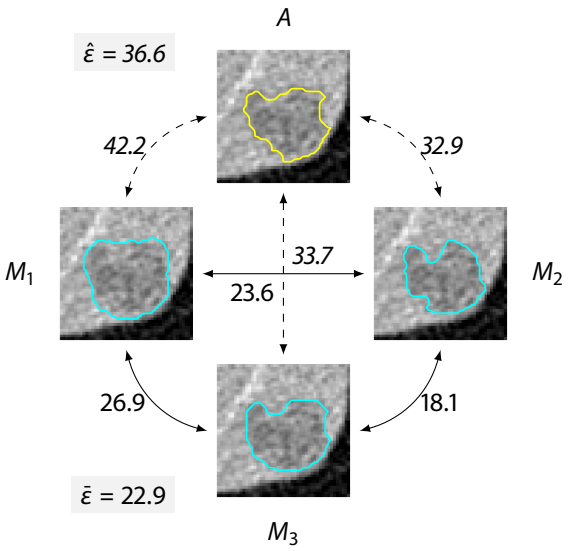
9.4 Experiments

In order to examine the results of this new methodology in practice, I used the data from Section 4.2. I calculated the scores according to the MICCAI criteria three times, using one of the manual segmentations as the ground truth at a time. This is not exactly in accordance with the MICCAI framework since no consensus about the "perfect segmentation" was achieved, but it gives an impression about the variability of the scores that are possible within this framework, depending on what data are available. Then, I computed the adaptive scores and compared the results with pairwise comparison and with comparison to a ground truth estimate.

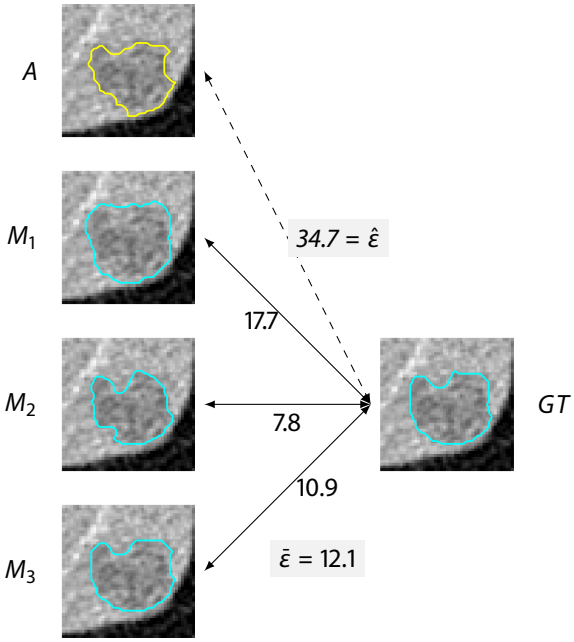
9.5 Results

All computed scores for all cases are shown in Figure 9.3. The distribution of the three MICCAI scores corresponds roughly to that of the individual metrics as shown in Figure 4.11, again with significant differences in some cases, depending on the chosen reference segmentation. The adaptive scores are often, but not always, higher than the MICCAI scores, which means that the required quality for being regarded "as good as an expert" has been lowered on average. Furthermore, ϕ_{pair} is always higher than ϕ_{GT}. To understand these relations better, it is helpful to have a look at the reference values $\bar{\varepsilon}$ that are used by the different scores.

The mean values and standard deviations of $\bar{\varepsilon}$ are shown in Table 9.2. Two important observations can be made. First, in agreement with the previous observations, the pairwise reference values



(a) Pairwise comparison



(b) Comparison with a ground truth estimate

Figure 9.2: Illustration of the two implementation variants of the adaptive score, exemplified by the volume overlap error.

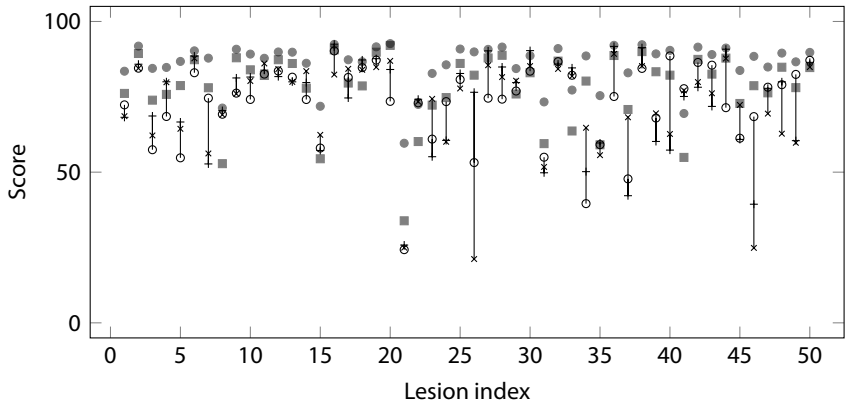


Figure 9.3: MICCAI scores for each reference segmentation (o, x, +) and adaptive scores using pairwise comparison (filled circle) and comparison with a ground truth estimate (filled square).

	$\bar{\varepsilon}_{\text{pair}}$	$\bar{\varepsilon}_{\text{GT}}$	$\bar{\varepsilon}_{\text{MICCAI}}$
Volume overlap error (%)	31.4 ± 12.5	17.3 ± 7.4	12.94
Volume difference (%)	24.6 ± 16.8	14.0 ± 9.1	9.64
Average surface distance (mm)	0.76 ± 0.51	0.38 ± 0.25	0.40
RMS surface distance (mm)	1.10 ± 0.74	0.69 ± 0.41	0.72
Max surface distance (mm)	4.64 ± 4.27	3.46 ± 3.09	4.00

Table 9.2: Comparison of adaptive reference values for the 50 cases of the study (mean ± standard deviation) and global MICCAI reference values for liver tumors.

seem to be systematically higher than those from the ground-truth-based computation and from the MICCAI criteria. This is due to the fact that the estimated ground truth is a kind of average of the masks, so the mean difference to the ground truth is lower than the average difference between the original masks. In the MICCAI framework, the reference mask is also an approximation of a "ground truth" because it required confirmation by both experts. The MICCAI reference values are slightly lower for volumetric measures and slightly higher for surface distances. This might be explained by the fact that surface distances are affected by the lesion size, so their average value may differ strongly depending on the data.

As a second observation, the standard deviations indicate a great variation in the reference values across the different cases. This confirms once more that the variability between experts depends on the individual cases and cannot be captured by global comparison values, especially when objects of various sizes are considered.

A different perspective on the results is taken in Figure 9.4, showing scatter plots of the average of the three MICCAI scores and both variants of the adaptive score. Additionally, the uncertainty about the true segmentation is encoded. From the plots, it can be seen that a high adaptive score is more likely if either the average MICCAI score or the uncertainty is high. Cases with a medium or low MICCAI score and high certainty are rather dragged down. These properties are true for both ϕ_{pair} and ϕ_{GT}.

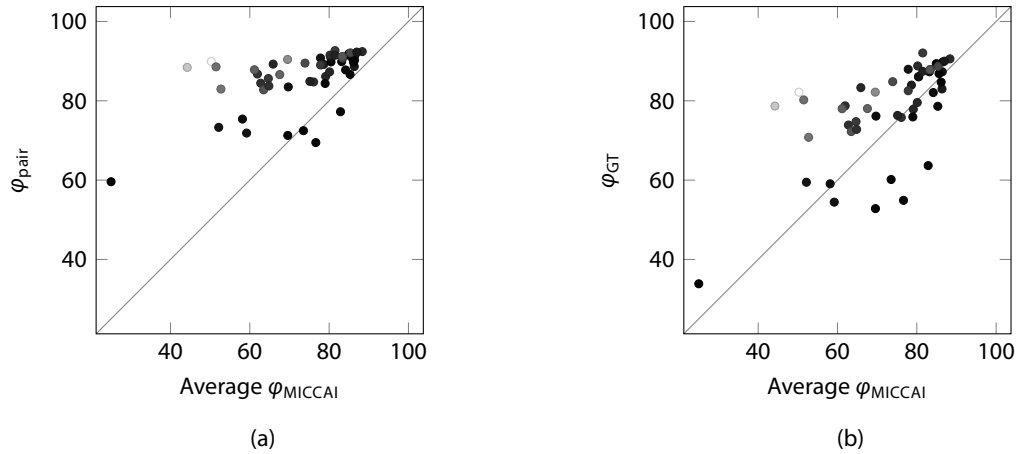


Figure 9.4: Relation between the average MICCAI score and the variants of the adaptive score. The filling of each point indicates the difference between minimum and maximum MICCAI score (the darker, the more certain the "ground truth").

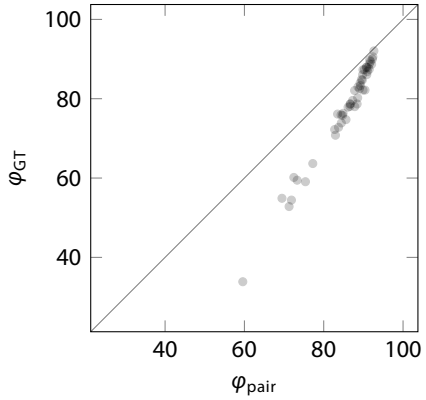
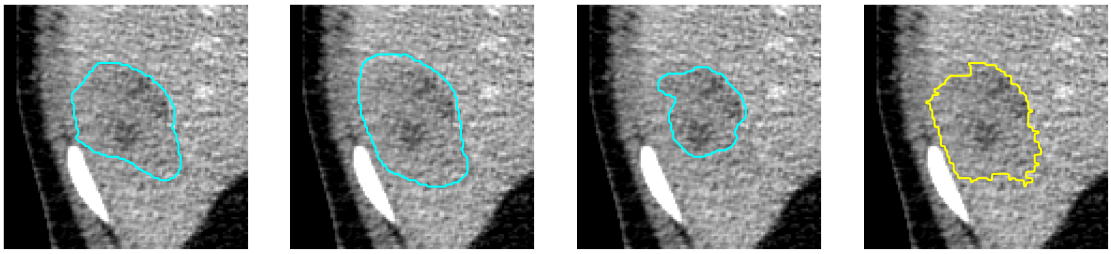


Figure 9.5: Correlation between φ_{pair} and φ_{GT}. Note that 90 points are approximately equivalent for both scores.

The values of ϕ_{GT} are consistently lower than those of ϕ_{pair}, but there is a strong correlation of 0.99 (Figure 9.5). This means that the two variants are essentially equivalent, which is not surprising given that one computes the average discrepancy to all masks and the other the discrepancy to the average mask. Interestingly, 90 points are approximately equivalent for both scores, so the interpretation of being "as good as an expert" is roughly the same. The extrapolations of the curves, however, differ.

I will now present some cases from the study that exemplify the principles of the new approach and its benefit. In Figure 9.6, a liver tumor with three highly different manual segmentations is depicted. Due to the low signal-to-noise ratio the delineation of this lesion is not clear. The table in Figure 9.6e shows that the global reference values ($\bar{\varepsilon}$) from the MICCAI validation framework are not suitable here to get an assessment of the algorithmic result in relation to the quality of the



(a) M_1: 25.04 ml (b) M_2: 34.59 ml (c) M_3: 18.22 ml (d) A: 34.98 ml

	$\hat{\varepsilon}_{\text{pair}}$	$\bar{\varepsilon}_{\text{pair}}$	φ_{pair}	$\hat{\varepsilon}_{\text{GT}}$	$\bar{\varepsilon}_{\text{GT}}$	φ_{GT}
Volume overlap error (%)	39.4	38.4	89.7	37.3	21.2	82.2
Volume difference (%)	32.4	41.9	92.3	34.5	21.7	84.1
Average surface distance (mm)	2.93	2.38	87.7	2.53	1.16	78.2
RMS surface distance (mm)	3.30	3.59	90.8	3.11	1.91	83.8
Max surface distance (mm)	17.60	15.56	89.0	17.01	9.83	82.7

(e) Discrepancy values, reference values and according scores

M_1	M_2	M_3		
φ_{MICCAI}			φ_{pair}	φ_{GT}
53.2	76.5	21.2	89.9	82.2

(f) Total scores

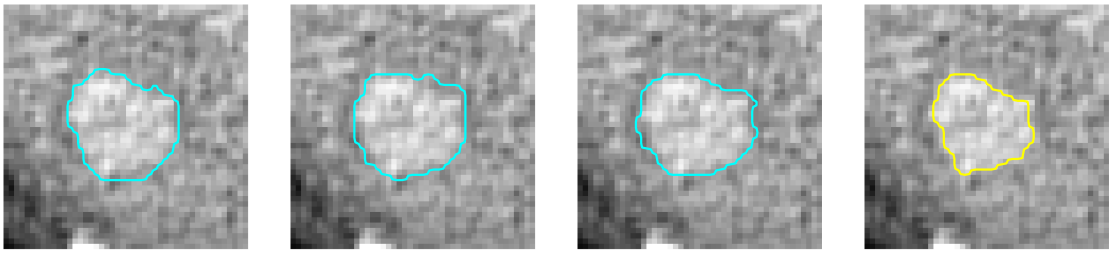
Figure 9.6: Example from the study where M_1, M_2 and M_3 differ significantly and A provides a visually satisfying compromise.

manual segmentations. The adaptive reference values are much higher, for the pairwise comparison even higher than the discrepancy values for the algorithm ($\hat{\varepsilon}$) which results in scores above 90 points. This is confirmed by the visual impression of Figure 9.6d which shows that A is a good compromise between the differing manual segmentations.

The second example in Figure 9.7 shows the opposite effect. Here the inter-reference metrics are lower than the MICCAI reference values (except for the volume overlap error) due to the high agreement between M_1, M_2 and M_3, but also due to the small size of the tumor. In comparison, A is too small, which results in lower scores according to the new method. This is again confirmed by visual inspection and comparison of the computed lesion volumes in Figures 9.7a to 9.7d.

9.6 Discussion

The validation of a validation methodology is very difficult or even impossible. Since the goal of this work is to increase the objectivity compared to a validation based on the opinion of a single expert, it does not make sense to correlate the adaptive scores with an expert rating or with scores computed by other methods. Instead an axiomatic approach was chosen: by specifying criteria that a validation framework should fulfill and constructing a methodology accordingly. I motivated



(a) M_1: 1.32 ml (b) M_2: 1.23 ml (c) M_3: 1.34 ml (d) A: 0.96 ml

	$\hat{\varepsilon}_{\text{pair}}$	$\bar{\varepsilon}_{\text{pair}}$	φ_{pair}	$\hat{\varepsilon}_{\text{GT}}$	$\bar{\varepsilon}_{\text{GT}}$	φ_{GT}
Volume overlap error (%)	28.2	16.8	83.2	27.4	8.8	68.0
Volume difference (%)	29.6	5.6	47.3	29.4	3.5	14.9
Average surface distance (mm)	0.77	0.32	84.2	0.76	0.16	63.8
RMS surface distance (mm)	0.92	0.51	85.4	0.89	0.35	78.3
Max surface distance (mm)	2.40	1.73	86.1	2.32	1.38	82.6

(e) Discrepancy values, reference values and according scores

M_1	M_2	M_3		
φ_{MICCAI}			φ_{pair}	φ_{GT}
82.2	84.6	81.8	77.2	63.7

(f) Total scores

Figure 9.7: Example from the study where M_1, M_2 and M_3 have a very good agreement whereas A is too small.

the criteria and verified the assumptions by theoretical considerations and practical results. In particular, I showed that reference segmentations by different experts may exhibit a high variability and that the degree of variability depends on several properties of the image and the object to be segmented. Therefore comparison with a single expert segmentation and global reference values for dissimilarity metrics can convey a wrong impression of the quality of an algorithm or – if used during development – carry the risk of overly adapting an algorithm to a specific expert.

Looking back at the criteria one by one, I can clearly state that the first requirement is fulfilled. All reference masks are treated equally and independently. Even if STAPLE is used for ground truth estimation, the masks may be given different weights during the computation, but not a priori. The method also complies with the second criterion. By determining the reference values individually for each case, the tolerance of the score is adapted to the level of variability within the set of reference segmentations. At this point, however, a further improvement may be possible. The variability is estimated for each case, but typically the variability is restricted to certain parts of an object. This could be captured by using a fuzzy ground truth that assigns each voxel a probability of being part of the object.

The third criterion is quite difficult to fulfill. Validation results from different test images can never be directly comparable and a fair comparison of methods should be based on a common

data set. But at least the adaptive score tries to compensate for different levels of "difficulty" of the segmentation tasks. If a test comprises only clearly delimitable objects on high-contrast data, it can be expected that the inter-expert variability is low and even slight deviations will be penalized by the score.

With regard to the different variants of the score, it was shown that they are essentially equivalent and have the same interpretation for 90 points, but ϕ_{GT} decreases faster. At this point, it is not clear whether any of the variants should be preferred. Maybe this question can be answered by further experiments.

In the experiments I used the same combination of comparison metrics as proposed by the MICCAI framework. I think, however, that this should also be examined more closely. In particular, the three surface distance measures seem to have a high correlation and the fact that they contribute 60 % of the final score is also debatable.

So far, I only applied the new method to liver tumor segmentation and used only three reference segmentations. A thorough investigation is only possible if the community can be encouraged to perform further tests for other segmentation problems. Studies should also be extended to a higher number of expert segmentations and analyze the effects of using different ground truth estimation methods.

Chapter 10

Variability of manual segmentations

10.1 Introduction

In the previous chapter, it was illustrated how the variability between reference segmentations affects validation results, but also how incorporating this variability into the validation methodology can help to make more meaningful statements about the quality of an algorithm.

A question that remains open is how many expert segmentations should be acquired for a validation study. This is important because creating reference segmentations requires a lot of effort from the experts, which is not always feasible. Therefore it is necessary to have an idea of the typical variability in manual delineations and of the number of references that provides a good compromise between completeness and viability. This may be a step towards giving concrete recommendations on how evaluation studies should be performed.

For this chapter, ten expert segmentations were acquired for a set of liver tumors. The goal of the analysis was to find out whether each of the ten experts actually contributed additional information or whether a subset of a particular size is sufficient.

10.2 Related work

The variability of manual segmentations has been an important field of research in radiation oncology for about ten years. In radiotherapy planning, the segmentation of the target structure, often called *target volume delineation* or *contouring*, is an essential prerequisite and is done manually in clinical practice. A good overview of variability studies has been assembled by Jameson et al. (2010). In particular, they investigated which methods have been applied to measure the differences between contours. The main organs of interest are prostate, lung, brain, and breast, and no studies with focus on liver tumors are mentioned. The methods for measuring variability are not standardized and therefore the results of different studies are hardly comparable. Metrics used include the coefficient of variation of the volume or the *concordance index*, which is the same as the volume overlap, but is sometimes also applied to the union and intersection of a set of segmentations. A new measure of contour deviation was introduced by Deurloo et al. (2005). A mean contour is defined by the set of voxels that have been segmented by at least 50 % of the readers. At each point of the mean contour, the perpendicular distance to the original contours is measured. It was found that these distances at each point roughly follow a Gaussian distribution and can thus be represented by their standard deviation. The local standard deviations at all surface points can be visualized and the overall variation can be measured by their mean, which is just called SD in the paper. The interpretation of this value is that about two thirds of the contours lie within this distance around the mean contour.

A review of the cited papers reveals that the intra- and inter-observer variability in target volume delineation in CT is generally regarded to be a critical issue among radiation oncologists. Looking at non-small-cell lung cancer (NSCLC) as an example, Bowden et al. (2002) report an average coefficient of variation of 20 % for the volume of six tumors delineated by six readers, ranging from 5 % to 42 %. A larger study with 22 tumors and 11 readers was performed by Steenbakkers et al. (2006). They measured an average concordance index of 0.17. In fact, it was zero in four cases where the intersection of all contours was empty. The *SD* value as defined above was 1.02 cm. The figures in this paper give a visual impression of significant variation.

For liver tumors, which will be the focus of my following investigations, Bellon et al. (1997) examined differences in manual and semi-automatic delineation. They used six manual delineations by three experts on 14 images, but only one slice was considered per image. The coefficient of variation of the area amounted to 10.2 %. A visual example is shown, but no interpretation of this result is given.

Interestingly, there seems to have been little interaction between the radiological and image analysis communities so far. This may be due to the fact that the underlying questions are different. For a clinician it is most important to investigate the effects of variability on treatment success, although it is not feasible to have several physicians delineate the same lesion in order to enhance accuracy. In image analysis, on the other hand, we are more interested in how to measure the quality of our algorithms given this variability, but so far this has not been examined systematically.

One of the few papers that analyzed variability with validation of algorithms in mind was published by Tingelhoff et al. (2008). They analyzed 20 manual delineations of the maxillary sinus and the ethmoid sinuses in a single CT scan. Coefficients of variation for the volume measurements are given as 5.1 % and 35.3 % for the two respective structures. The authors conclude that "manual segmentation is not adequate as gold standard" and suggest that the combination of multiple expert segmentations might be a solution.

10.3 Data

For my experiments, I used a collection of 13 CT images from 13 different patients, seven hospitals, and scanners by four vendors. Slice thickness varies between 0.8 and 1.5 mm. In each image, a liver tumor was selected and a region of interest was cut out. Ten manual segmentations were obtained by radiologists and experienced radiology technicians. They drew the outlines of the tumors in all axial slices in the original image resolution. The experts were allowed to adjust the window settings individually in order to include this typical source of variation into the analysis.

The image data and three expert segmentations were taken from the data set specified in Section 4.2. I tried to choose a representative set of tumors with respect to size, anatomical location, contrast to parenchyma, resolution, noise etc. However, I selected mostly smaller tumors in order to reduce the workload of the experts. Figure 10.1 shows all tumors used in the study.

10.4 Methodology

The analysis of the manual segmentations will be divided into two parts. First, a descriptive analysis of the variability is performed, based on an established metric, the coefficient of variation (COV)

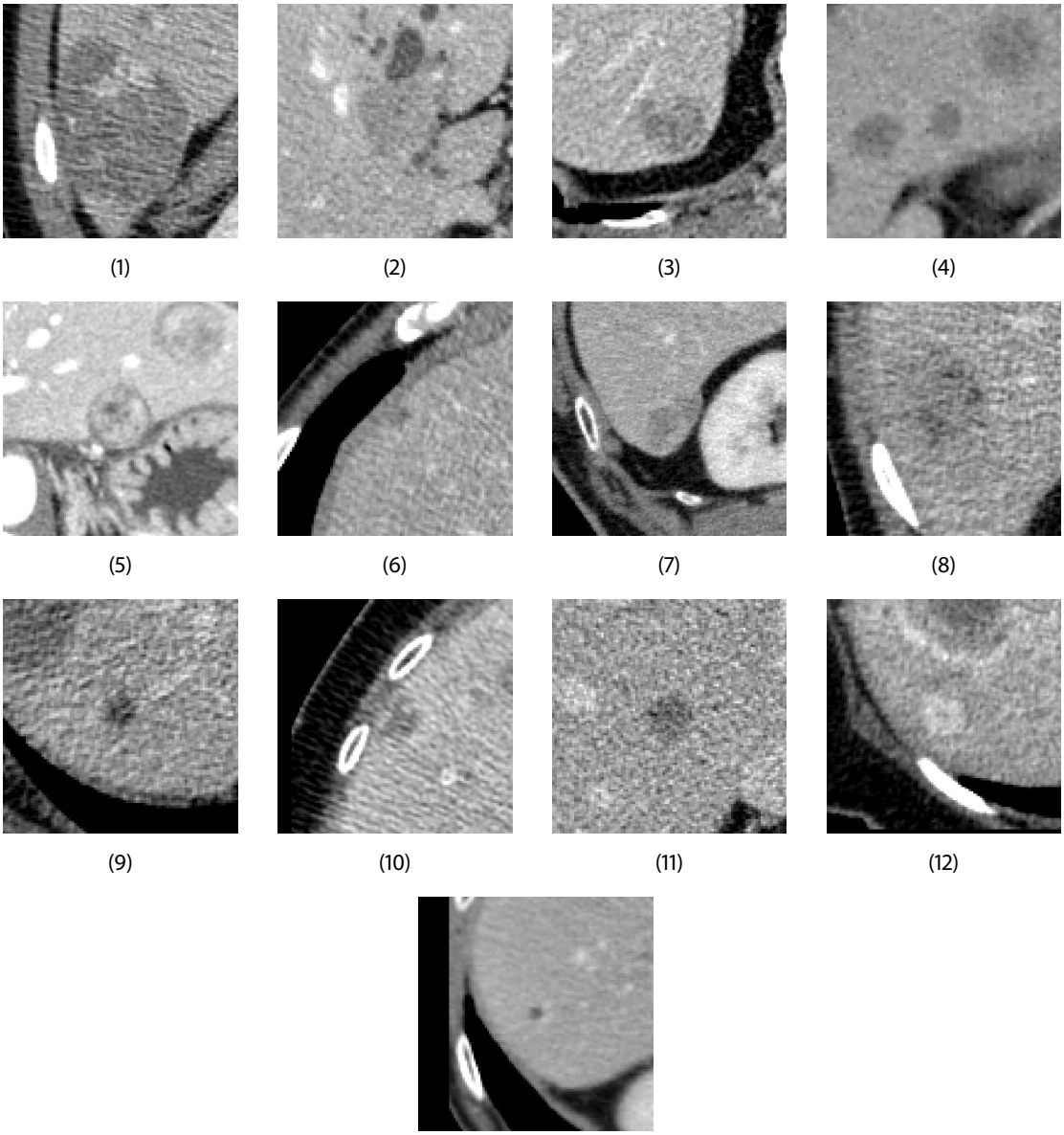


Figure 10.1: Tumors that were used for the study. The tumors are in the center of the slices shown here. All ROIs have the same physical size and are shown with the same window settings.

of the volume. Although variations in the volume do not necessarily reflect all possible variations in the outlines themselves, this simple metric has some advantages. Since there is no standard methodology to measure the variability within a set of segmentations, any choice of a metric would bear the risk of introducing artificial effects. Different metrics might generate different results, and it would be difficult to interpret these differences. Therefore I decided to keep this study as simple as possible and restricted my examinations to the volumes of the masks defined by the manual delineations.

From a purely *descriptive* analysis, however, it cannot be concluded how many manual delineations are necessary to cover the typical variability and allow a meaningful validation of a segmentation algorithm. *Inductive* statistics try to make general statements from a limited amount of data. In my study, ten segmentations are available for 13 cases. This is a large number for this particular problem, but a very low number for making inductive statistics. It is impossible to derive a reliable estimate of the distribution of manual delineations by experts from such a small sample.

Rather than trying to make statements about the complete "population" of manual delineations on liver tumors from the available data, I decided to simulate this process one level lower. I set up an *urn model* where the ten delineations for the 13 tumors are the complete population. Then, subsets of this population can be investigated by drawing from the urn. If the simulated population is representative of the real population, the conclusions can be generalized and give an idea of how much information is lost when a subset of a certain size is used.

For the urn experiment, the average volume of all ten experts is considered as the true volume v_{true}. For a subset, the average volume v_{sub} of all contained segmentations is computed. The volume error is then defined by

$$\varepsilon = \frac{|v_{\text{sub}} - v_{\text{true}}|}{v_{\text{true}}}. \tag{10.1}$$

Now consider the set of volumes measured by the ten experts for a particular tumor. Each subset of size $k = 1, \ldots, 9$ represents a random choice of k experts and defines a volume estimate and an error according to the formula given above. If k segmentations are drawn from an urn, all $\binom{10}{k}$ combinations have the same probability. Therefore the expected value of the error made by a selection of k experts is just the mean of the errors of all subsets containing k elements. This expected error can be plotted against k. A different point of view on the data can be taken by computing the probability of exceeding a particular error depending on k.

10.5 Results

The manual segmentations are displayed as overlaid contours (Figure 10.2) and summarized in probability maps (Figure 10.3), along with the average volume of the segmentations and the corresponding coefficients of variation. Figure 10.4 visualizes the absolute volumes of the segmentations as well as the relative differences to the mean volumes.

The COV has a mean of 21.83 % and ranges from 3.56 % to 48.46 %. The extreme values occurred for the largest and the smallest lesion, respectively. The data set is too small to prove a correlation between size and variability, but it is plausible since differences between segmentations occur mostly at the boundary, and for smaller objects the relative amount of boundary voxels is higher.

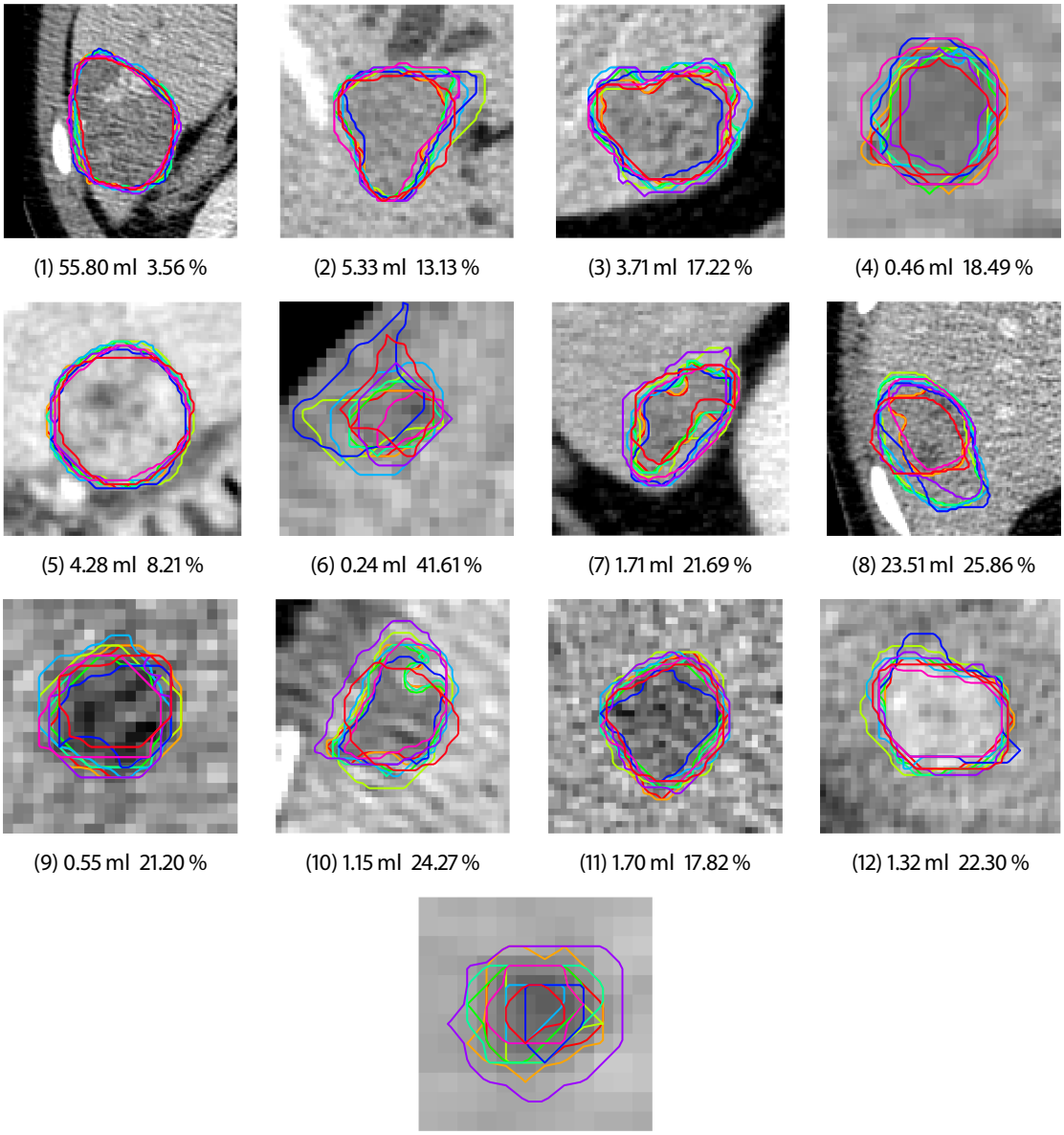


Figure 10.2: Tumors from Figure 10.1 with ten manual segmentations, shown as colored contours. Images are zoomed for better visibility of the contours. For each tumor, the mean volume and the coefficient of variation of the volume are given.

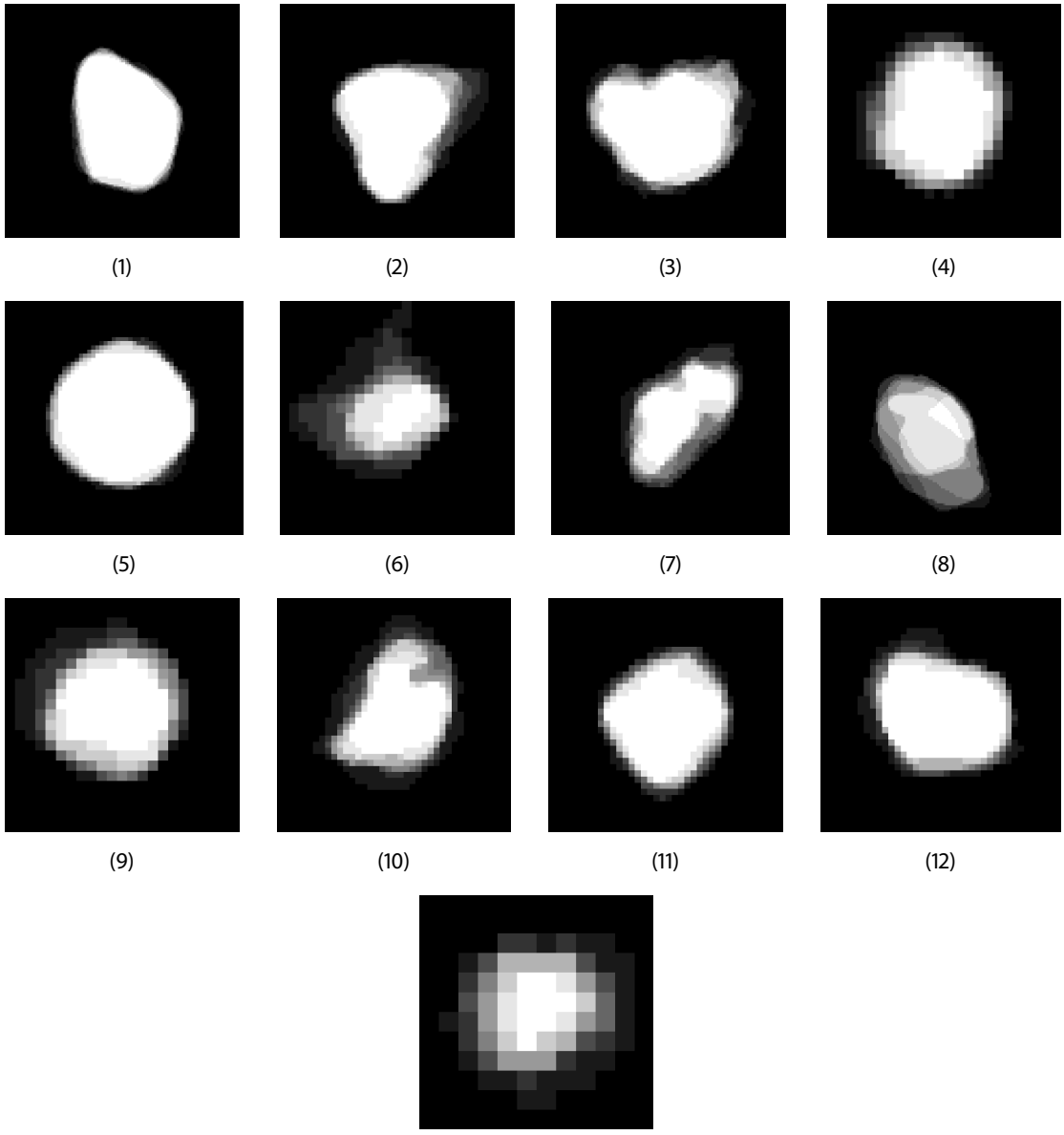


Figure 10.3: Probability maps for the segmentations in Figure 10.2.

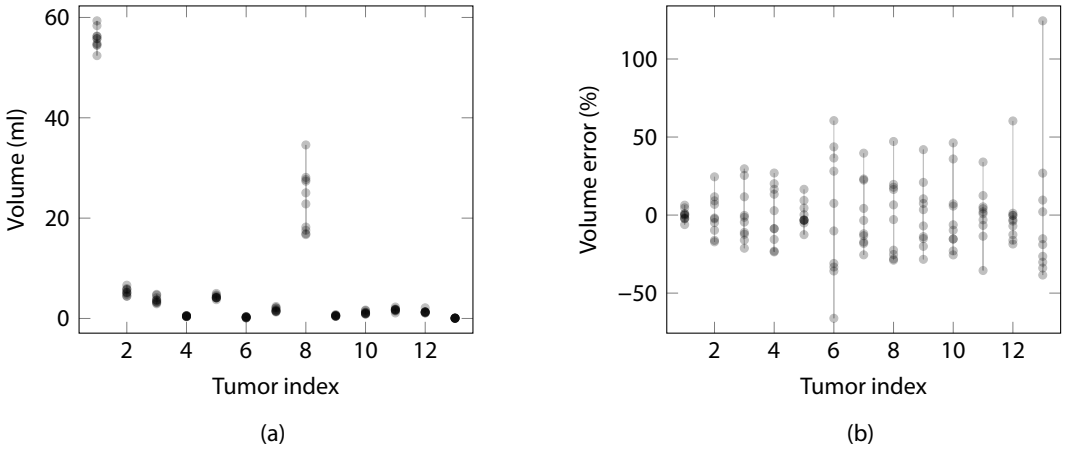


Figure 10.4: Volumes and volume errors (compared to the mean volumes) of the ten manual delineations for 13 tumors.

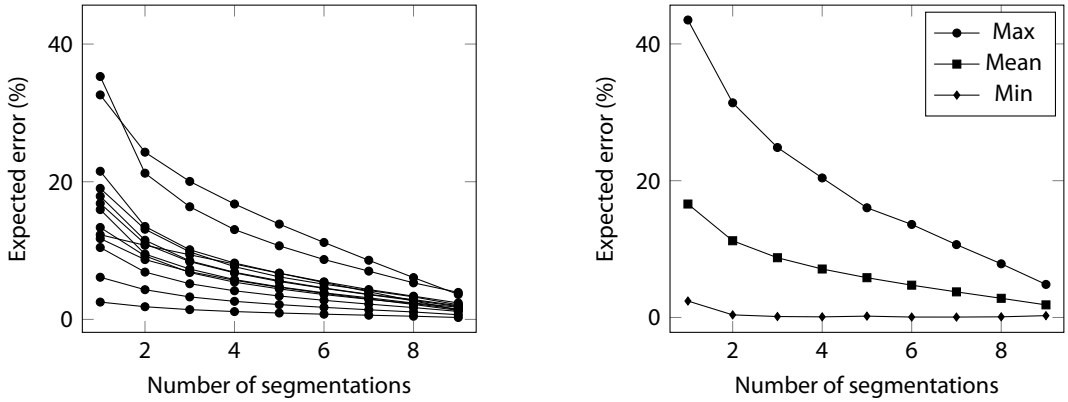
Tumor 13 has the highest COV. Since it is very small, the differences are likely to be caused by differing inclusion of the partial volume zone and drawing inaccuracies. In tumor 1, on the other hand, the differences between the segmentations are only minor drawing inaccuries which hardly influence the volume. In both cases, it seems that in principle all experts agree about the shape of the tumor and its rough extents and that the differences can be explained by *statistical* deviation.

Tumors 7 and 8 are examples of an average COV. Visually, however, there are more significant differences between the delineations than in the previous cases, which cannot be explained by drawing inaccuracy alone. The readers actually seem to disagree about the extent of the tumors. In tumor 8, this can be explained by the weak contrast to the liver tissue and the heterogeneity of the tumor. Although in tumor 7 the contrast is stronger, there are some regions which are not clearly hypodense but, probably inferred by expert knowledge, are still regarded tumor by some readers.

The results of the urn model experiment are displayed in Figure 10.5a where the expected errors for a random sample of k segmentations are shown. The errors for $k = 1$ correspond roughly to the COVs. The curves for all tumors decrease monotonically as expected and for higher k show a tendency to converge to each other. If a single segmentation is drawn randomly, the error is 16.6 % on average and up to 35.3 % for tumors with a high variability. The average error is already reduced significantly for $k = 2$ (11.2 %) and almost halved for $k = 3$ (8.7 %). Then it decreases more slowly until being halved again at $k = 6$ (4.7 %). These relations are approximately true for the curves of all individual tumors as well, so the relative error reduction does not depend on the COV.

Figure 10.5b presents summarized results for all tumors. It shows the mean expected error, i.e., the mean of all curves in Figure 10.5a, along with the minimum and maximum errors, again averaged over all tumors. The minimum error is a best-case estimate, i.e., it results from selecting those segmentations that are closest to the true segmentation. It can be seen that if the two "best" experts are selected, the error is already zero. Analogously, the maximum error represents the worst choice of segmentations and is more than twice the mean error for any number of experts.

The curves in Figure 10.6 show which k should be chosen if a certain error ε should not be exceeded or should only have a low probability of being exceeded. For instance, for being



(a) Individual tumors

(b) Average over tumors

Figure 10.5: Expected value of the volume error if $k = 1, \ldots, 9$ segmentations are randomly selected. (a) Curves for all 13 tumors. (b) Expected, maximum and minimum error, averaged over all tumors.

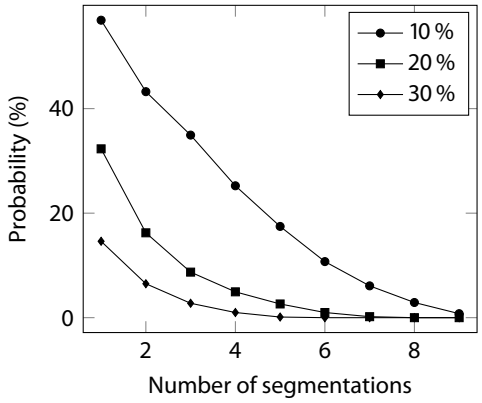


Figure 10.6: Probability of making a volume error $> \varepsilon$ if $k = 1, \ldots, 9$ segmentations are randomly selected, for $\varepsilon = 10, 20, 30$ %, averaged over all tumors.

sure to have an error of less than 30 % in average over all tumors, at least six of the ten manual segmentations are necessary. While for $k = 3$ the probability of $\varepsilon > 30$ % is still only 2.7 %, reducing the error bound to $\varepsilon > 10$ % increases the probability to 34.9 %. An average error of less than 10 % cannot even be guaranteed for $k = 9$.

10.6 Discussion

The analysis of the manual delineations of ten experts shows that there is a considerable variability with a mean volume COV of 21.83 %. This variability differs strongly between tumors and reaches a maximum of 48.46 % for a very small tumor.

Visual inspection of the delineations revealed that two kinds of uncertainty should be distinguished. *Statistical uncertainty* can be modeled by a mean contour and an uncertainty margin

of a particular width. It can be caused by differing perception of the tumor size depending on the window settings. If the contrast is low, some readers may tend to draw the outline around everything that might be tumor, while others mark only the region that certainly belongs to the tumor. *Semantic uncertainty*, on the other hand, cannot be modeled by deviation around a mean contour. Instead, larger regions, not just narrow bands of voxels, have been included by some experts and excluded by others, resulting in a fuzzy segmentation with distinct areas of a particular probability.

In an urn experiment, I simulated the effects of using samples of the ten segmentations for volume estimation. Assuming that the available set of tumors is representative of practical cases, one can say that if a single manual segmentation is used to estimate the volume of a tumor, the expected error is around 17 %, but might get as high as 35 %. For comparison, a volume increase of 73 % or a decrease of 66 % are used to classify the response of a tumor to treatment according to the RECIST criteria (Eisenhauer et al. 2009). This means that the error of volume estimation is already a quarter of a clinically significant threshold for volume change on average and up to a half in some cases.

This uncertainty also has an impact on the validation of segmentation algorithms. When comparing the result of an algorithm to a single reference segmentation and computing volume errors, the expected volume error of the reference needs to be taken into account. Since mostly absolute values of errors are used, it cannot be determined whether the two errors accumulate or cancel each other out. Again for comparison, the evaluation of a state-of-the-art algorithm for liver tumors by Smeets et al. (2010) reports a volume error by comparison with a single reference segmentation of 17.9 %, which is just the expected volume error that I found for the reference. Therefore it is questionable if this kind of validation is reliable.

However, the error decreases with increasing numbers of segmentations. When three are used instead of one, the average error is almost halved. Taking the efforts required for obtaining manual delineations into account, a number of three could be a reasonable compromise in practice. But since my study was limited, this is only a first impression and it is important to gather a larger collection of data for drawing more general conclusions.

Chapter 11

A tool for creating probabilistic expert segmentations

11.1 Introduction

In the previous chapter, the variability of manual tumor delineations was analyzed. It turned out that in some cases the variability cannot be explained by drawing inaccuracy alone. This variability captures the uncertainty of a set of experts about the true segmentation. My observations show that it does not make sense to insist on having a hard "ground truth". Instead, the uncertainty of the experts should be incorporated into the validation methodology. This requires a quantification of the uncertainty and a distinction between two different aspects. Often it can be assumed that several experts *mean* the same contour but draw it slightly differently. Any of their segmentations and *anything in between* can be considered correct. In these cases, it makes sense to compute an average contour and a standard deviation which models the *statistical uncertainty*. However, there are also cases where the experts actually seem to have different ideas of the correct segmentation. Then, an average contour does not make sense and a good algorithm should be close to at least one of the expert delineations. This may be called *semantic uncertainty*.

The main problem in taking the uncertainty of the true segmentation into account is the effort that is required from experts. Even large validation initiatives such as LIDC (Armato III et al. 2011) collected "only" four segmentations per case. Most individual researchers do not have access to more than one or two experts. A common restriction, however, is that experts are usually asked to draw a single contour as their best estimate of the true segmentation. Variability is then measured in terms of the differences between the best estimates of multiple experts. An aspect that is mostly disregarded is the uncertainty of *each individual* expert. Before drawing a contour, each reader has to make two decisions: where to draw the most probable boundary within an often blurred margin and whether or not to include ambiguous regions which may or may not be part of the tumor.

The hypothesis of this chapter is that the variability between multiple contours can in part be reproduced by a lower number of experts, if they are given a tool to express their uncertainty. Such a tool will be presented and evaluated in the following sections. The evaluation uses the same tumors as the variability study in Chapter 10 and compares the results of three users with the new tool to those of ten users drawing conventional contours. As in that study, liver tumor segmentation in CT is used as an example, but the methodology is easily generalised.

11.2 Related work

A related approach was presented by Restif (2007). He introduced a framework called *Comets* that allows a single user to create a probabilistic reference segmentation. It was specifically developed for 2D cytometry images where blurred boundaries and connected objects are common problems.

The user draws the most probable outline and adds inner and outer limit pixels which are definitely inside or outside the object, but as close to the border as possible. From this input a confidence map is computed by setting 0 on the drawn outline, ±1 on the limit pixels and interpolating on all other pixels.

As compared to Restif's work, this article presents three additional contributions. First, the focus will be on 3D images. While transferring the concept to 3D is straightforward in principle, efficiency becomes an issue when contours have to be drawn in each slice. The concept of limit pixels may not be intuitive for all users and it might take some time to define them on all slices. Therefore, I opted for a simpler and more efficient interaction based on contours.

Second, Comets does not distinguish statistical and semantic uncertainty but covers both by a single method and blends them together in the confidence map. For validation purposes, however, it is advantageous to separate these two aspects. This is done explicitly in the new tool.

Finally, Restif does not compare Comets to other ways of generating reference segmentations. Since my work was motivated by the goal to reduce the number of necessary experts without losing information, I conducted a user study to evaluate this.

11.3 Workflow

With the new tool, segmentation is done in two phases. In the first phase, the most probable contour is drawn. The statistical uncertainty is modeled by a rim around this contour. The inner boundary of the rim delineates all voxels which are definitely part of the tumor. Analogously, all voxels outside the outer boundary definitely belong to the background. The width of the uncertainty rim is set by the user before drawing the contour. For simplicity, this setting is applied globally on each slice, but can be adapted locally afterwards. The current width is visualized as the diameter of a circle displayed at the cursor position and can be changed by turning the mouse wheel (Figure 11.1a).

Once the user has finished drawing, the inner and outer contours are generated by applying a distance transform to the user-defined contours and adding or subtracting the uncertainty radius. These contours are displayed and can be edited. Although in many cases a global uncertainty radius is reasonable, there are cases where a different value should be set locally. For example, a tumor may have a blurred boundary to the liver parenchyma, but a clearly defined one to a structure outside the liver. Editing is achieved by drawing new partial contours which are inserted into the existing ones.

Now the contours are transformed into a probability map (Figure 11.1b). Voxels are assigned 1 if they are inside the inner contour and 0 if they are outside the outer contour. Between the contours, probabilites are linearly interpolated. Note that, unlike Restif (2007), the values are limited to $[0, 1]$ and do not decrease further outside the outer contour.

In the optional second phase, additional regions can be outlined and assigned a confidence of belonging to the tumor (Figure 11.1c). For these regions, no uncertainty margin is defined because that seemed to be too confusing for users, although technically it would not be a problem. Regions are included in the probability map by using the maximum of the value assigned earlier and the confidence set by the user (Figure 11.1d). Alternatively, the results of the two phases can be stored separately for further analysis.

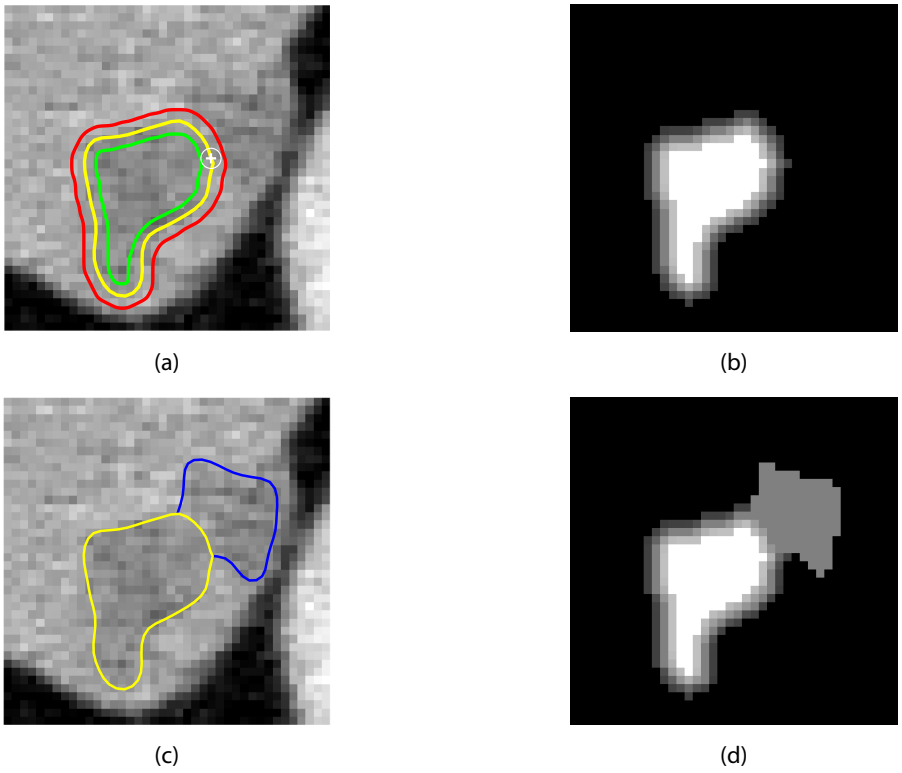


Figure 11.1: Illustration of the workflow of the new tool and the results it produces. (a) User-drawn contour (yellow) and inner and outer contours (green and red) constructed from the radius of the circle. (b) Probability map. (c) Additional region with confidence 0.5 (blue). (d) Probability map.

11.4 Evaluation

The newly developed tool was evaluated in a study with three experts (one radiologist and two radiology technicians) and the same 13 liver tumors that were used in Chapter 10. The created probability maps, averaged over the readers, are shown in Figure 11.2.

The usage of the features offered by the tool varied across the participants. Readers 1 and 2 adapted the uncertainty width in each case, whereas Reader 3 always used the same value (in voxels). Reader 3 also did not draw any additional regions. The two others added three and eight regions, respectively, in total affecting eight of the 13 tumors. I compared the new results to the earlier ones (Figure 10.2) and found a high visual similarity for many of the tumors. The chosen uncertainty widths correspond well to the statistical uncertainty among ten experts as illustrated by tumors 12 and 13. Still, some interesting effects can be seen. In tumor 1, for instance, a region was left out by one of the three readers although it had been included by all ten readers in the earlier study. For tumor 8, on the other hand, there was slightly more variability among ten readers than could be reproduced by three.

For a more quantitative analysis, we define a metric that captures the variability encoded in a probabilistic segmentation. It is based on the fuzzy volume overlap, where the volume of a

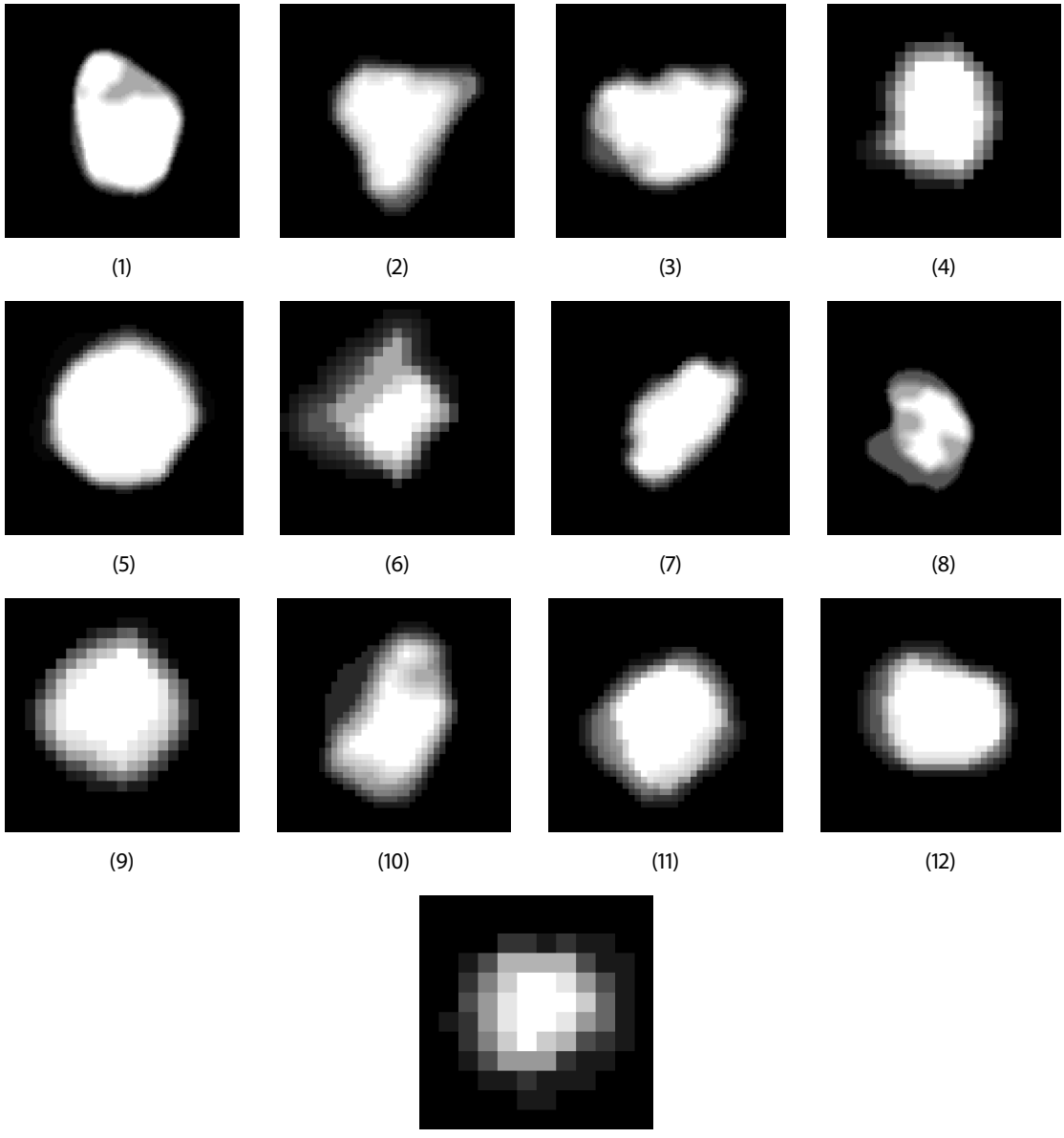


Figure 11.2: Averaged probability maps created by the three study participants.

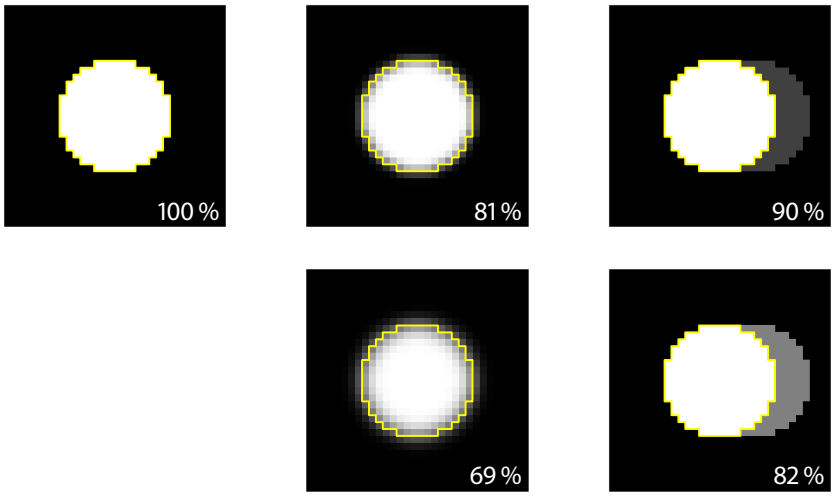


Figure 11.3: Illustration of properties of the fuzzy self-overlap on academic examples. The binarization is shown in yellow. The value is 100 % for a binary segmentation and drops with increasing uncertainty about the boundary and about additional regions.

segmentation is the sum of the probabilities of all voxels, with intersection and union being defined by the voxel-wise minimum and maximum (Crum et al. 2006). The fuzzy overlap of two segmentations compares two aspects, the mean segmentations and the spread of probabilities around them. Applying the fuzzy overlap to a probabilistic segmentation and its own mean segmentation, defined by thresholding at 0.5, measures the variability. We call this the *fuzzy self-overlap*. It is 100 % for a binary segmentation and gets lower the more the probabilities are spread. Figure 11.3 illustrates the properties of this metric.

Figure 11.4 compares the variability in averaged segmentations created from the ten conventional segmentations of my earlier study and from the three probabilistic ones of the present study. In the plot, it is clearly visible that with the new tool more information can be acquired using fewer experts. One expert using the new tool could replace three experts drawing conventional contours. Together, the three experts in this study generated more variability than ten in the previous study.

After the study, the participants were interviewed. They said that they felt unfamiliar with expressing their uncertainty because usually they have to make a crisp decision. While, however, the uncertainty width was adopted easily, the readers had difficulties defining additional regions and quantifying their confidence. This shows that users need some training to get used to the new way of thinking the tool requires. The reader who achieved the best results was already interviewed in the development phase and probably had the best understanding of the concepts at the time of the study.

11.5 Discussion

The motivation for this work was to be able to reduce the number of experts needed for a validation study without losing information and without increasing the workload per expert too much. A

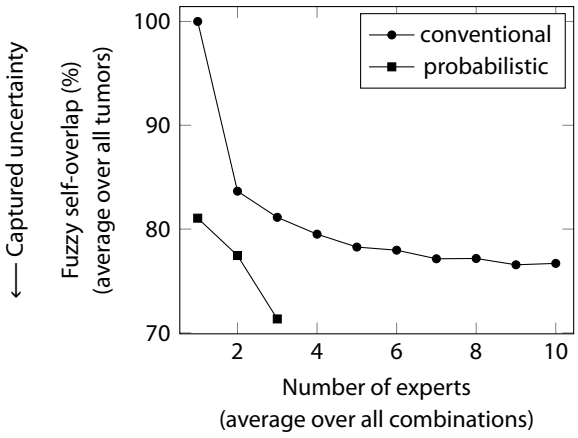


Figure 11.4: Variability in combined segmentations by different numbers of experts, using conventional and probabilistic expert segmentations. The lower the fuzzy self-overlap, the higher the variability.

basic decision was made to separate statistical and semantic uncertainties explicitly, both for reducing the effort and for making it available for further analysis. In the study, the statistical uncertainties were captured well at virtually no additional cost because the uncertainty width was set very quickly. A possible disadvantage of the conceptual separation, however, is the fact that users typically decide to add a uncertainty region in the first phase, but have to wait for the second phase before they can actually draw it. This requires a high concentration and memory capacity and might be a reason why not many uncertainty regions were added. A workflow that allows alternating the two phases on each slice might improve this. As a further improvement, one might think about not just adding, but also subtracting uncertainty regions from the initial segmentation. This might be more intuitive than leaving out regions with a very high confidence in the first phase and adding them later.

The results of the study show that using the new tool expert uncertainty can be recovered with a lower number of experts as compared to conventional contours. This was confirmed both visually and quantitatively. It is interesting to see that in some cases uncertainty regions were used that have no correspondence among ten experts. This shows that the explicit capturing of uncertainty can actually gather additional information compared to just averaging over a large number of segmentations. But on the other hand, there are also some cases where the complete variation cannot be reproduced with a lower number of readers. In Figure 11.2, tumors 1 and 8 illustrate this duality.

The processing time was not measured, but from our observations during the study it can be said that the new method allows a considerable reduction of efforts. Assuming that segmentation took 25 % longer than pure outlining, which is a very conservative estimation since uncertainty regions are typically small and cover only a couple of slices, the overall person time was still reduced by almost two thirds.

Future work is necessary to investigate how these probability maps can be used for algorithm validation. Since they are not inherently binary, many common approaches are not directly applicable. Some widely used metrics like the volume overlap can be easily generalized for probabilistic

segmentations, whereas for surface distances there is no obvious solution and different proposals have been made. Crum et al. (2006) discuss their application in medical image analysis. They focus, however, on the case where the algorithm result is probabilistic rather than the reference segmentation. Further experiments should provide insight into how suitable these methods are for validation. Also, common methods are not able to make use of the explicit distinction between statistical and semantic uncertainty. The additional information that is becoming available calls for a completely new validation paradigm that works not only on (a set of) random expert delineations, but builds up knowledge about plausible and implausible segmentations.

Chapter 12

Conclusion

This final chapter summarizes the contributions of the thesis and discusses directions for future research.

In Part I, a method for segmenting liver lesions in CT was presented. It aims primarily at metastases, but is also able to handle primary tumors as long as they are homogeneous or consist of a core and a rim. During development, special attention was paid to the computation time because that is a paramount criterion for clinical acceptance. In the end, a median runtime of less than 1 s could be achieved, which is considerably faster than all previously published methods. This was achieved by a combination of computationally modest methods such as region growing and morphological processing with a sophisticated threshold selection based on appearance models for different lesion types and the corresponding histograms.

The method was evaluated on 371 lesions not available during development with one reference segmentation per case. This is the most substantial validation of a liver tumor segmentation algorithm published so far. A median volume overlap of 62.8 % and a median Hausdorff distance of 4.5 mm were achieved. These values, however, are hard to interpret because the expected quality is unclear and a comparison with other methods is not possible if different data are used.

On a subset of 50 lesions, three reference segmentations were compared to each other, yielding a median pairwise volume overlap of 67.7 % and a median pairwise Hausdorff distance of 3.2 mm. On the same subset the algorithm achieved a volume overlap of 64.3 % and a Hausdorff distance of 3.8 mm, averaging over the three reference segmentations. This indicates that the results are quite close to the optimum when taking the uncertainty about the true result into account.

The segmentation algorithm has been integrated into a commercial software, which also includes an efficient editing tool, and is available to a large number of radiologists. Unfortunately, it is not often used in practice because it still takes longer than just measuring the longest diameter and the additional benefit of tumor volumetry has not been proved yet. So, while significant improvements of the algorithm itself are probably no longer possible, the next challenge is to find ways to establish it in the clinic.

One possibility to increase the acceptance of segmentation is to provide not only the lesion volume, but further parameters. These might include the fractions of vital and necrotic tissue in the tumor or textural parameters such as homogeneity. These measures are expected to become more important with the advent of new therapies which do not primarily reduce the tumor size, but change its internal structure.

A different approach is to try to further speed up the measurements. For example, this could be done by using the segmentation result only for computing the longest diameter. This would still improve the reproducibility of the measurements, but reduce the effort for manual editing significantly.

The process could also be accelerated by combining a fully automatic lesion detection and segmentation which precomputes the results and minimizes the waiting time. If no previous results are available, a possible workflow would be to detect lesion candidates with a highly sensitive, but not necessarily very specific detection algorithm, precompute the segmentations and display them as soon as the user clicks on a lesion. A solution for follow-up cases has been proposed in Part II of this thesis.

This part presented a comprehensive framework for automatic lesion tracking which automatizes the complete process and only requires the user to check the results and correct them if necessary. The algorithm consists of three steps. First, a template matching is used to detect a point in the lesion of interest. Then, the segmentation algorithm is initialized automatically, simulating the user input. Finally, a classifier detects implausible results and discards them. In contrast to previous approaches, it has been designed to work without incorporating any organ-specific assumptions. It was, however, specifically optimized for lung nodules, liver metastases, and lymph nodes.

A database of 994 lesion pairs was used to motivate the adequacy of the template matching approach in theory and to optimize the parametrization in practice. An independent set of 209 lesion pairs was used for evaluation. On average, a matching rate of 81.4 % was achieved, with lung nodules showing the best and lymph nodes the worst results. The median segmentation quality was the same as with manual initialization, and the classifier had an F_1 score of 0.88.

In addition to the technical evaluation of the individual components, I conducted a workflow-oriented evaluation of the framework as a whole with four radiologists. That study showed that automatic lesion tracking may actually have a benefit in clinical routine. Both reading time and intra-reader variability of the measured volumes were reduced.

While these results are very promising and, according to the study participants, the tool would already be helpful in the clinic, it would be desirable to further increase the matching rate. This would come at a higher computational cost, but that is tolerable since the computations are meant to be performed in preprocessing. An idea that should be investigated in the future is how the plausibility check can be used not only to detect failures, but also to fix them. This could be done by changing parameters, relaxing assumptions or triggering more expensive procedures. Humans often use landmark structures such as vessels in the neighborhood of a lesion for orientation. Such an approach could be mimicked by an algorithm.

While the first two parts presented methods to solve a particular image analysis problem, Part III focused on the validation of those methods from a more theoretical point of view. Starting from the observation that the variability of manual reference segmentations has an impact on the quality assessment of algorithms, the concept of uncertainty-aware validation was introduced. The main paradigm is to skip the idea of a "ground truth" and use the uncertainty about the true segmentation to calibrate validation metrics.

A new adaptive score was introduced as a generalization of the MICCAI Grand Challenge framework. It quantifies the individual uncertainty of each case and assesses the algorithm more strictly if experts agree and more tolerantly if they disagree. Some examples were shown to illustrate properties of the new score, but it is hard to verify that it is actually a good measure of algorithm quality. I am planning a case study where at least two algorithms are compared using at least three reference segmentations and ideally experts rate the results and discuss their uncertainty. With such a setup, it might become possible to prove the adequacy of the new score.

An important factor in this context is the number of expert segmentations that should be used. It might be assumed that above a particular number no additional information is acquired. However, experiments with ten experts showed that this is not the case. On the other hand, it was also shown that moving from one to two or three experts makes the results clearly more robust.

Given this result, I tried to find a way to acquire the same amount of information from fewer experts. A tool was developed that allows experts to express their individual uncertainty during segmentation, making an explicit distinction between statistical and semantic uncertainty. A study revealed that three experts using this tool captured more uncertainty than ten users drawing simple contours. So this new way of generating reference segmentations might have two advantages: It allows a reduction of expert effort and a better understanding of which segmentations are plausible.

I think these are also the main goals of future research in this field. Expert resources are limited, so it is necessary to capture as much knowledge as possible while they are available. Ideally, a validation method should be able to predict how an expert would rate a particular segmentation, which deviations from their own manual segmentation would be tolerated and which would be corrected. This thesis pointed out some ideas to make validation more objective and efficient, but it is up to the medical image analysis community to join this kind of research and share their data so that large-scale studies can be performed and algorithms developed at different institutions can be compared.

Bibliography

Armato III, S., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Beek, E. J. R. van, Yankelevitz, D. F., Biancardi, A. M., Bland, P. H., Brown, M. S., Engelmann, R. M., Laderach, G. E., Max, D., Pais, R. C., Qing, D. P.-Y., Roberts, R. Y., Smith, A. R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G. W., Jude, C. M., Munden, R. F., Petkovska, I., Quint, L. E., Schwartz, L. H., Sundaram, B., Dodd, L. E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Vande Casteele, A., Gupte, S., Sallam, M., Heath, M. D., Kuhn, M. H., Dharaiya, E., Burns, R., Fryd, D. S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B. Y., and Clarke, L. P. (2011). *The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans*. In: Medical Physics 38(2), pp. 915–931.

Behnaz, A. S., Snider, J., Eneh, C., Esposito, G., Wilson, E., Yaniv, Z., Cohen, E., and Cleary, K. (2010). *Quantitative CT for volumetric analysis of medical images: initial results for liver tumors*. In: SPIE Medical Imaging, 76233U–1–6.

Beigelman-Aubry, C., Raffy, P., Yang, W., Castellino, R. A., and Grenier, P. A. (2007). *Computer-aided detection of solid lung nodules on follow-up MDCT screening: evaluation of detection, tracking, and reading time*. In: American Journal of Roentgenology 189(4), pp. 948–955.

Bellon, E., Feron, M., Maes, F., Van Hoe, L., Delaere, D., Haven, F., Sunaert, S., Baert, A. L., Marchal, G., and Suetens, P. (1997). *Evaluation of manual vs semi-automated delineation of liver lesions on CT images*. In: European Radiology 7(3), pp. 432–438.

Beyer, F., Wormanns, D., Novak, C., Odry, B. L., Kohl, G., and Heindel, W. (2004). *Clinical evaluation of a software for automated localization of lung nodules at follow-up CT examinations*. In: Fortschritte Röntgenstrahlen 176(6), pp. 829–836.

Boehm, B. W. (1984). *Verifying and validating software requirements and design specifications*. In: IEEE Software 1(1), pp. 75–88.

Bornemann, L., Dicken, V., Kuhnigk, J.-M., Wormanns, D., Shin, H.-O., Bauknecht, H.-C., Diehl, V., Fabel, M., Meier, S., Kress, O., Krass, S., and Peitgen, H.-O. (2007). *OncoTREAT: a software assistant for cancer therapy monitoring*. In: International Journal of Computer Assisted Radiology and Surgery 1(5), pp. 231–242.

Bowden, P., Fisher, R., Mac Manus, M., Wirth, A., Duchesne, G., Millward, M., McKenzie, A., Andrews, J., and Ball, D. (2002). *Measurement of lung tumor volumes using three-dimensional computer planning software*. In: International Journal of Radiation Oncology*Biology*Physics 53(3), pp. 566–573.

Cai, W., Yoshida, H., and Harris, G. J. (2007). *Dynamic-thresholding level set: a novel computer-aided volumetry method for liver tumors in hepatic CT images*. In: SPIE Medical Imaging, 65142W–1–8.

Chalana, V. and Kim, Y. (1997). *A methodology for evaluation of boundary detection algorithms on medical images*. In: IEEE Transactions on Medical Imaging 16(5), pp. 642–652.

Crum, W. R., Camara, O., and Hill, D. L. G. (2006). *Generalized overlap measures for evaluation and validation in medical image analysis*. In: IEEE Transactions on Medical Imaging 25(11), pp. 1451–1461.

Deng, X. and Du, G. (2008). *Editorial: 3d segmentation in the clinic: a grand challenge II - liver tumor segmentation*. URL: http://grand-challenge2008.bigr.nl/proceedings/pdfs/lts08/00_Editorial.pdf.

Deurloo, K. E. I., Steenbakkers, R. J. H. M., Zijp, L. J., Bois, J. A. de, Nowak, P. J. C. M., Rasch, C. R. N., and Herk, M. van (2005). *Quantification of shape variation of prostate and seminal vesicles during external beam radiotherapy*. In: International Journal of Radiation Oncology*Biology*Physics 61(1), pp. 228–238.

Drechsler, K., Strosche, M., and Laura, C. O. (2011). *Automatic ROI identification for fast liver tumor segmentation using graph-cuts*. In: SPIE Medical Imaging, 79622S–1–7.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., and Verweij, J. (2009). *New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)*. In: European Journal of Cancer 45, pp. 228–247.

Erasmus, J. J., Gladish, G. W., Broemeling, L., Sabloff, B. S., Truong, M. T., Herbst, R. S., and Munden, R. F. (2003). *Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response*. In: Journal of Clinical Oncology 21(13), pp. 2574–2582.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). *The WEKA data mining software: an update*. In: ACM SIGKDD Explorations Newsletter 11(1), pp. 10–18.

Häme, Y. and Pollari, M. (2012). *Semi-automatic liver tumor segmentation with hidden Markov measure field model and non-parametric distribution estimation*. In: Medical Image Analysis 16(1), pp. 140–149.

Heckel, F., Moltz, J. H., Bornemann, L., Dicken, V., Bauknecht, H.-C., Fabel, M., Hittinger, M., Kießling, A., Meier, S., Püsken, M., and Peitgen, H.-O. (2009). *3d contour based local manual correction of tumor segmentations in CT scans*. In: SPIE Medical Imaging, pp. 72593L–1–9.

Heckel, F., Konrad, O., Hahn, H. K., and Peitgen, H.-O. (2011). *Interactive 3d medical image segmentation with energy-minimizing implicit functions*. In: Computers and Graphics 35(2), pp. 275–287.

Heimann, T., Ginneken, B. van, Styner, M. A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., Bello, F., Binnig, G., Bischof, H., Bornik, A., Cashman, P. M. M., Chi, Y., Córdova, A., Dawant, B. M., Fidrich, M., Furst, J. D., Furukawa, D., Grenacher, L., Hornegger, J., Kainmüller, D., Kitney, R. I., Kobatake, H., Lamecker, H., Lange, T., Lee, J., Lennon, B., Li, R., Li, S., Meinzer, H.-P., Németh, G., Raicu, D. S., Rau, A.-M., Rikxoort, E. M. van, Rousson, M., Ruskó, L., Saddi, K. A., Schmidt, G., Seghers, D., Shimizu, A., Slagmolen, P., Sorantin, E., Soza, G., Susomboon, R., Waite, J. M., Wimmer, A., and Wolf, I. (2009). *Comparison and evaluation of methods for liver segmentation from CT datasets*. In: IEEE Transactions on Medical Imaging 28(8), pp. 1251–1265.

Heußel, C. P., Meier, S., Wittelsberger, S., Götte, H., Mildenberger, P., and Kauczor, H.-U. (2007). *Follow-up CT measurement of liver malignoma according to RECIST and WHO vs. volumetry*. In: Fortschritte Röntgenstrahlen 179(9), pp. 958–964.

IEEE (1990). *IEEE standard glossary of software engineering*. In: IEEE Std 610.12-1990.

Jameson, M. G., Holloway, L. C., Vial, P. J., Vinod, S. K., and Metcalfe, P. E. (2010). *A review of methods of analysis in contouring studies for radiation oncology*. In: Journal of Medical Imaging and Radiation Oncology 54, pp. 401–410.

Jolly, M.-P. and Grady, L. (2008). *3d general lesion segmentation in CT*. In: International Symposium on Biomedical Imaging, pp. 796–799.

Keil, S., Behrendt, F. F., Stanzel, S., Sühling, M., Koch, A., Bubenzer, J., Mühlenbruch, G., Mahnken, A. H., Günther, R. W., and Das, M. (2008). *Semi-automated measurement of hyperdense, hypodense and heterogeneous hepatic metastasis on standard MDCT slices. Comparison of semi-automated and manual measurement of RECIST and WHO criteria*. In: European Radiology 18(11), pp. 2456–2465.

Keil, S., Plumhans, C., Behrendt, F. F., Stanzel, S., Sühling, M., Mühlenbruch, G., Mahnken, A. H., Günther, R. W., and Das, M. (2009). *Semi-automated quantification of hepatic lesions in a phantom*. In: Investigative Radiology 44(2), pp. 82–88.

Keil, S., Plumhans, C., Nagy, I., Schiffl, K., Soza, G., Behrendt, F. F., Mahnken, A. H., Günther, R. W., and Das, M. (2010a). *Dose reduction for semi-automated volumetry of hepatic metastasis in MDCT studies*. In: Investigative Radiology 45(2), pp. 77–81.

Keil, S., Bruners, P., Ohnsorge, L., Plumhans, C., Behrendt, F. F., Stanzel, S., Sühling, M., Günther, R. W., Das, M., and Mahnken, A. H. (2010b). *Semiautomated versus manual evaluation of liver metastases treated by radiofrequency ablation*. In: Journal of Vascular and Interventional Radiology 21, pp. 245–251.

Koo, C. W., Anand, V., Girvin, F., Wickstrom, M. L., Fantauzzi, J. P., Bogoni, L., Babb, J. S., and Ko, J. P. (2012). *Improved efficiency of CT interpretation using an automated lung nodule matching program*. In: American Journal of Roentgenology 199(1), pp. 91–95.

Kuhnigk, J.-M., Dicken, V., Bornemann, L., Bakai, A., Wormanns, D., Krass, S., and Peitgen, H.-O. (2006). *Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans*. In: IEEE Transactions on Medical Imaging 25(4), pp. 417–434.

Lee, K. W., Kim, M., Gierada, D. S., and Bae, K. T. (2007). *Performance of a computer-aided program for automated matching of metastatic pulmonary nodules detected on follow-up chest CT*. In: American Journal of Roentgenology 189(5), pp. 1077–1081.

Li, K. and Jolly, M.-P. (2008). *Simultaneous detection of multiple elastic surfaces with application to tumor segmentation in CT images*. In: SPIE Medical Imaging, 69143S–1–11.

Li, Y., Hara, S., and Shimura, K. (2006). *A machine learning approach for locating boundaries of liver tumors in CT images*. In: International Conference on Pattern Recognition, pp. 400–403.

Massoptier, L. and Casciaro, S. (2008). *A new fully automatic and robust algorithm for fast segmentation of liver tissue and tumors from CT scans*. In: European Radiology 18(8), pp. 1658–1665.

Moltz, J. H., Schwier, M., and Peitgen, H.-O. (2009a). *A general framework for automatic detection of matching lesions in follow-up CT*. In: IEEE International Symposium on Biomedical Imaging, pp. 843–846.

Moltz, J. H., Bornemann, L., Kuhnigk, J.-M., Dicken, V., Peitgen, E., Meier, S., Bolte, H., Fabel, M., Bauknecht, H.-C., Hittinger, M., Kießling, A., Püsken, M., and Peitgen, H.-O. (2009b). *Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans*. In: IEEE Journal of Selected Topics in Signal Processing 3(1), pp. 122–134.

Moltz, J. H., Rühaak, J., Hahn, H. K., and Peitgen, H.-O. (2011a). *A novel adaptive scoring system for segmentation validation with multiple reference masks*. In: SPIE Medical Imaging, pp. 796214–1–10.

Moltz, J. H., Braunewell, S., Rühaak, J., Heckel, F., Barbieri, S., Tautz, L., Hahn, H. K., and Peitgen, H.-O. (2011b). *Analysis of variability in manual liver tumor delineation in CT scans*. In: IEEE International Symposium on Biomedical Imaging, pp. 1974–1977.

Moltz, J. H., D'Anastasi, M., Kießling, A., Santos, D. Pinto dos, Schülke, C., and Peitgen, H.-O. (2012). *Workflow-centered evaluation of an automatic lesion tracking software for chemotherapy monitoring by CT*. In: European Radiology 22(12), pp. 2759–2767.

Moltz, J. H., Steinberg, C., Geisler, B., and Hahn, H. K. (2013). *A tool for efficient creation of probabilistic expert segmentations*http://dx.doi.org/. *A tool for efficient creation of probabilistic expert segmentations*. In: Medical Image Understanding and Analysis, pp. 7–12.

Opfer, R., Brenner, W., Carlsen, I., Renisch, S., Sabczynski, J., and Wiemker, R. (2008). *Automatic lesion tracking for a PET/CT based computer aided cancer therapy monitoring system*. In: SPIE Medical Imaging, pp. 691513–1–10.

Popa, T., Ibanez, L., Levy, E., White, A., Bruno, J., and Cleary, K. (2006). *Tumor volume measurement and volume measurement comparison plug-ins for VolView using ITK*. In: SPIE Medical Imaging. Vol. 6914, 69141B.

Puesken, M., Juergens, K. U., Edenfeld, A., Buerke, B., Seifarth, H., Beyer, F., Suehling, M., Osada, N., Heindel, W., and Weßling, J. (2009). *Accuracy of liver lesion assessment using automated measurement and segmentation software in biphasic multislice CT (MSCT)*. In: Fortschritte Röntgenstrahlen 181(1), pp. 67–73.

Puesken, M., Buerke, B., Fortkamp, R., Koch, R., Seifarth, H., Heindel, W., and Wessling, J. (2011). *Liver lesion segmentation in MSCT: effect of slice thickness on segmentation quality, measurement precision and interobserver variability*. In: Fortschritte Röntgenstrahlen 183(4), pp. 372–380.

Ray, S., Hagge, R., Gillen, M., Cerejo, M., and Shakeri, S. (2008). *Comparison of two-dimensional and three-dimensional iterative watershed segmentation methods in hepatic tumor volumetrics*. In: Medical Physics 35(12), pp. 5869–5881.

Restif, C. (2007). *Revisiting the evaluation of segmentation results: Introducing confidence maps*. In: Medical Image Computing and Computer-Assisted Intervention, pp. 588–595.

Ritter, F., Boskamp, T., Homeyer, A., Laue, H., Schwier, M., Link, F., and Peitgen, H.-O. (2011). *Medical image analysis: a visual approach*http://dx.doi.org/. *Medical image analysis: a visual approach*. In: IEEE Pulse 2(6), pp. 60–70.

Rohlfing, T. and Maurer Jr., C. R. (2007). *Shape-based averaging*. In: IEEE Transactions on Image Processing 16(1), pp. 153–161.

Sensakovic, W. F., Starkey, A., Roberts, R., Straus, C., and Armato III, S. G. (2010). *The influence of initial outlines on manual segmentation*. In: Medical Physics 37, pp. 2153–2158.

Shao, M.-Z. and Badler, N. (1996). *Spherical sampling by Archimedes' theorem*. Tech. rep. University of Pennsylvania.

Shi, J., Sahiner, B., Chan, H.-P., Hadjiski, L., Zhou, C., Cascade, P. N., Bogot, N., Kazerooni, E. A., Wu, Y.-T., and Wei, J. (2007). *Pulmonary nodule registration in serial CT scans based on rib anatomy and nodule template matching*. In: Medical Physics 34(4), pp. 1335–1347.

Smeets, D., Loeckx, D., Stijnen, B., De Dobbelaer, B., Vandermeulen, D., and Suetens, P. (2010). *Semi-automatic level set segmentation of liver tumors combining a spiral-scanning technique with supervised fuzzy pixel classification*. In: Medical Image Analysis 14(1), pp. 13–20.

Sofka, M. and Stewart, C. V. (2010). *Location registration and recognition (LRR) for serial analysis of nodules in lung CT scans*. In: Medical Image Analysis 14(3), pp. 407–428.

Stawiawski, J., Decenière, E., and Bidault, F. (2008). *Interactive liver tumor segmentation using graph-cuts and watershed*. URL: http://grand-challenge2008.bigr.nl/proceedings/pdfs/lts08/02_cmm.pdf.

Steenbakkers, R. J. H. M., Duppen, J. C., Fitton, I., Deurloo, K. E. I., Zijp, L. J., Comans, E. F. I., Uitterhoeve, A. L. J., Rodrigus, P. T. R., Kramer, G. W. P., Bussink, J., De Jaeger, K., Belderbos, J. S. A., Nowak, P. J. C. M., Herk, M. van, and Rasch, C. R. N. (2006). *Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis*. In: International Journal of Radiation Oncology*Biology*Physics 64(2), pp. 435–448.

Su, Z., Deng, X., Chefd'hotel, C., Grady, L., Fei, J., Zheng, D., and Chen, N. (2011). *Quantitative evaluation of six graph based semi-automatic liver tumor segmentation techniques using multiple sets of reference segmentation*. In: SPIE Medical Imaging, pp. 796619–1–6.

Sun, S., Rubin, G. D., Paik, D., Steiner, R. M., Zhuge, F., and Napel, S. (2007). *Registration of lung nodules using a semi-rigid model: method and preliminary results*. In: Medical Physics 34(2), pp. 613–626.

Szilágyi, T., Verhoek, M., and Noble, J. A. (2009). *Detecting early response to therapy in liver cancer treatment: 3d metastases segmentation using graph-cuts with a modified prior*http://dx.doi.org/. *Detecting early response to therapy in liver cancer treatment: 3d metastases segmentation using graph-cuts with a modified prior*. In: Medical Image Understanding and Analysis, pp. 84–88.

Taieb, Y., Eliassaf, O., Freiman, M., Joskowicz, L., and Sosna, J. (2008). *An iterative bayesian approach for liver analysis: tumors validation study*. URL: http://grand-challenge2008.bigr.nl/proceedings/pdfs/lts08/05_HUJI-liver.pdf.

Tao, C., Gierada, D. S., Zhu, F., Pilgram, T. K., Wang, J. H., and Bae, K. T. (2009). *Automated matching of pulmonary nodules: evaluation in serial screening chest CT*. In: American Journal of Roentgenology 192(3), pp. 624–628.

Therasse, P., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L., Verweij, J., Van Glabbeke, M., Oosterom, A. T. van, Christian, M. C., and Gwyther, S. G. (2000). *New guidelines to evaluate the response to treatment in solid tumors*. In: Journal of the National Cancer Institute 92(3), pp. 205–216.

Tingelhoff, K., Eichhorn, K. W. G., Wagner, I., Kunkel, M. E., Moral, A. I., Rilk, M. E., Wahl, F. M., and Bootz, F. (2008). *Analysis of manual segmentation in paranasal CT images*. In: European Archives of Oto-Rhino-Laryngology 265(9), pp. 1061–1070.

Udupa, J. K., LeBlanc, V. R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L. M., Hirsch, B. E., and Woodburn, J. (2006). *A framework for evaluating image segmentation algorithms*. In: Computerized Medical Imaging and Graphics 30(2), pp. 75–87.

Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). *Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation*. In: IEEE Transactions on Medical Imaging 23(7), pp. 903–921.

Weßling, J., Puesken, M., Koch, R., Kohlhase, N., Persigehl, T., Mesters, R., Heindel, W., and Buerke, B. (2012). *MSCT follow-up in malignant lymphoma: comparison of manual linear measurements*

with semi-automated lymph node analysis for therapy response classification. In: Fortschritte Röntgenstrahlen 184(9), pp. 795–804.

Wiemker, R., Hoop, B. de, Kabus, S., Gietema, H., Opfer, R., and Dharaiya, E. (2008). *Performance study of a globally elastic locally rigid matching algorithm for follow-up chest CT*. In: SPIE Medical Imaging, pp. 691706–1–6.

Williams, G. W. (1976). *Comparing the joint agreement of several raters with another rater*http://dx.doi.org/. *Comparing the joint agreement of several raters with another rater*. In: Biometrics 32(3), pp. 619–627.

Wong, D., Liu, J., Fengshou, Y., Tian, Q., Xiong, W., Zhou, J., Qi, Y., Han, T., Venkatesh, S. K., and Wang, S.-c. (2008). *A semi-automated method for liver tumor segmentation based on 2D region growing with knowledge-based constraints*. URL: http://grand-challenge2008.bigr.nl/proceedings/pdfs/lts08/09_NUS-I2R-team1.pdf.

Wulff, A. M., Bolte, H., Fischer, S., Freitag-Wolf, S., Soza, G., Tietjen, C., Biederer, J., Heller, M., and Fabel, M. (2012). *Lung, liver and lymph node metastases in follow-up MSCT: comprehensive volumetric assessment of lesion size changes*. In: Fortschritte Röntgenstrahlen 184(9), pp. 820–828.

Xu, J., Greenspan, H., Napel, S., and Rubin, D. L. (2011). *Automated temporal tracking and segmentation of lymphoma on serial CT examinations*. In: Medical Physics 38(11), pp. 5879–5886.

Yan, J., Zhao, B., Curran, S., and Zelenetz, A. (2007). *Automated matching and segmentation of lymphoma on serial CT examinations*. In: Medical Physics 34(1), pp. 55–62.

Yim, P. J. and Foran, D. J. (2003). *Volumetry of hepatic metastases in computed tomography using the watershed and active contour algorithms*. In: IEEE Symposium on Computer-Based Medical Systems, pp. 329–335.

Yu, P., Sheah, K., and Poh, C. L. (2012). *Automating the tracking of lymph nodes in follow-up studies of thoracic CT images*. In: Computer Methods and Programs in Biomedicine 106(3), pp. 150–159.

Zhang, Y. J. (1996). *A survey on evaluation methods for image segmentation*. In: Pattern Recognition 29(8), pp. 1335–1346.

Zhao, B., Schwartz, L. H., Jiang, L., Colville, J., Moskowitz, C., Wang, L., Leftowitz, R., Liu, F., and Kalaigian, J. (2006). *Shape-constraint region growing for delineation of hepatic metastases on contrast-enhanced computed tomograph scans*. In: Investigative Radiology 41(10), pp. 753–762.

Zhou, J.-Y., Wong, D. W. K., Ding, F., Venkatesh, S. K., Tian, Q., Qi, Y.-Y., Xiong, W., Liu, J. J., and Leow, W.-K. (2010). *Liver tumour segmentation using contrast-enhanced multi-detector CT data: performance benchmarking of three semiautomated methods*. In: European Radiology 20, pp. 1738–1748.

Zhou, J., Xiong, W., Tian, Q., Qi, Y., Liu, J., Leow, W. K., Han, T., Venkatesh, S. K., and Wang, S.-c. (2008). *Semi-automatic segmentation of 3d liver tumors from CT scans using voxel classification and propagational learning*. URL: http://grand-challenge2008.bigr.nl/proceedings/pdfs/lts08/10_NUS-I2R-team2.pdf.