Nephroblastoma

in

MRI Data

A dissertation submitted towards the degree

Doctor of Engineering (Dr.-Ing.)

of the Faculty of Mathematics and Computer Science
of Saarland University

by
Sabine Müller

Saarbrücken, 2019

Day of Colloquium:
09.07.2020
Dean of Faculty:
Prof. Dr. Thomas Schuster
Chair of the Committee:
Prof. Dr. Thorsten Herfet
Reviewers:
Prof. Dr. Joachim Weickert
Prof. Dr. Christian Daul
Prof. Dr. Norbert Graf
Prof. Dr. Andreas Keller
Academic Assistant:
Dr. Pascal Peter

Short Abstract

The main objective of this work is the mathematical analysis of nephroblastoma in MRI sequences. At the beginning we provide two different datasets for segmentation and classification. Based on the first dataset, we analyze the current clinical practice regarding therapy planning on the basis of annotations of a single radiologist. We can show with our benchmark that this approach is not optimal and that there may be significant differences between human annotators and even radiologists. In addition, we demonstrate that the approximation of the tumor shape currently used is too coarse granular and thus prone to errors. We address this problem and develop a method for interactive segmentation that allows an intuitive and accurate annotation of the tumor.
While the first part of this thesis is mainly concerned with the segmentation of Wilms' tumors, the second part deals with the reliability of diagnosis and the planning of the course of therapy. The second data set we compiled allows us to develop a method that dramatically improves the differential diagnosis between nephroblastoma and its precursor lesion nephroblastomatosis. Finally, we can show that even the standard MRI modality for Wilms' tumors is sufficient to estimate the developmental tendencies of nephroblastoma under chemotherapy.

Kurzzusammenfassung

Die massgebliche Zielsetzung dieser Arbeit ist die mathematische Analyse des Nephroblastoms basierend auf MRT Sequenzen. Zu Beginn stellen wir zwei verschiedene Datensätze zur Verfügung die in der Segmentierung und respektive der Klassifizierung ihre Anwendung finden. Basierend auf ersterem Datensatz analysieren wir die gängige klinische Praxis bezüglich Therapieplanung aufbauend auf Annotationen eines einzelnen Radiologen. Wir können anhand unseres Benchmarks zeigen, dass diese Herangehensweise nicht optimal ist und es auch beträchtliche Abweichungen zwischen einzelnen menschlichen Annotatoren und insbesondere Radiologen geben kann. Darüber hinaus weisen wir nach, dass die aktuell verwendete Annäherung der Tumorform zu grobgranular und damit fehleranfällig ist. Wir addressieren dieses Problem und entwickeln eine Methode zur interaktiven Segmentierung, die eine intuitive und akkurate Annotation des Tumors erlaubt.
Während der erste Teil dieser Arbeit sich massgeblich mit der Segmentierung von Wilms' Tumoren befasst, wenn wir uns im zweiten Teil der Diagnosesicherheit sowie der Planung des Therapieverlaufs zu. Der zweite von uns zusammengestellte Datensatz erlaubt es uns die Differentialdiagnostik zwischen Nephroblastom und Nephroblastomatose dramatisch zu verbessern. Schlussendlich können wir zeigen, dass selbst die Standardmodalität für MRI Bildgebungen des Wilms' Tumors bereits ausreicht um die Entwicklungstendenzen des Nephroblastoms unter Chemotherapie abschätzen zu können.

Abstract

Today, cancer is one of the most common and unfortunately deadly diseases, with approximately 17 million new diagnoses every year [172]. It is particularly tragic when infants and newborns are already affected: It is completely incomprehensible to them why they have to undergo painful and often protracted therapies.
The most common malignant kidney tumor in childhood is the Wilms' tumor, also known as nephroblastoma. Since approximately 75% of patients are younger than five years (with a peak between two and three years), optimal therapy is of particular importance [47, 137].

This work is therefore dedicated to the analysis and optimization of current therapy planning for the treatment of Wilms' tumors - viewed from the perspective of medical image processing. During the course of therapy, images of the tumor are taken at certain times with the aid of magnetic resonance tomography. These include the time of diagnosis, the completion of chemotherapy and, if necessary, further imaging after additional chemotherapy or radiation. We use this imaging in our work to identify opportunities and potential for improvement.
In the first part of this dissertation we deal with fundamental aspects of therapy planning and begin to evaluate the current clinical practice for determining tumor volume. Since we need a meaningful dataset for this, we start with the composition of a benchmark dataset. This enables us to address various aspects. On the one hand, it allows us to quantify the differences between individual human experts. On the other hand, we show that the current clinical practice for the determination of tumors is very error-prone. In addition, this data collection also enables us to test automatic and semi-automatic methods for segmentation and thus also for a proper quantification of the tumor volume.

We take this next step and develop a semi automatic approach to image segmentation. The human interaction in semi automatic approaches allows to include prior knowledge of a human expert. Typically, an user coarsely labels several regions in the image to generate an initial seed population. Especially in this interactive setting, it is essential that a method is insensitive to information that is not optimal or even wrong: humans usually do not fully agree on an objects' outline resulting in inter-rater variability. We show this ability in various contexts: from direct human interaction to seed labels generated in a fully automatic fashion.

This robust and intuitive approach then allows us to focus in the second part of our work on the diagnosis of kidney disease and therapy prognosis. For this purpose we provide another data set. This contains over 200 patients and provides a comprehensive basis for the possible tumor entitats of the nephroblastoma. A further important aspect of this data collection is that over 50 patients with nephroblastomatosis could also be included. Since this extremely rare diagnosis of Wilms' tumor is neither malignant nor shows invasive tendencies, no further therapy is usually necessary (apart from an extraction). Obviously, a reliable diagnosis of these two diseases is indispensable. We have addressed this problem and developed a new approach to reliably differentiate these two diseases.

Finally, we are working on predicting the development of nephroblastoma under the influence of chemotherapy. Since during the administration of chemotherapeutics the Wilms' tumor changes in its histology and may mutate into more aggressive subtypes, it is of immense importance to determine this evolution at the earliest possible time in order to be able to adapt the therapy accordingly.
For this purpose we created the average visual appearance of all subtypes and used it to train a classification algorithm. Although our attempts are only a proof of concept on this subject, we made some remarkable observations: In all our experiments the accuracy is always much better than the chance. This is especially interesting as the imaging of nephroblastoma is not standardized and show a high parameter noise.

Zusammenfassung

In der heutigen Zeit zählt Krebs mit jährlich ca. 17 Millionen Neudiagnosen zu den häufigsten und leider auch tödlichsten Krankheiten [172]. Besonders tragisch ist es wenn bereits Kleinkinder und Neugeborene betroffen sind: Es ist für sie vollkommen unverständlich, wieso sie sich teilweise schmerzhaften und langwierigen Therapien unterziehen müssen.
Mit der häufigste abdominale Tumor im Kindesalter ist der Wilms' Tumor, der auch als Nephroblastom bezeichnet wird. Da circa 75% der Patienten jünger als fünf Jahre alt sind (mit einer vermehrten Häufigkeit zwischen zwei und drei Jahren), ist eine optimale Therapie von besonderer Bedeutung [47, 137].

Aus diesem Grund widmet sich diese Arbeit der Analyse und Optimierung der aktuellen Therapieplanung für Wilms' Tumore- betrachtet aus der Perspektive der medizinischen Bildverarbeitung. Während des Therapieverlaufs werden zu bestimmten Zeitpunkten Aufnahmen des Tumors mit Hilfe der Magnetische Resonanztomographie erstellt. Diese umfassen den Diagnosezeitpunkt, den Abschluss der Chemotherapie und gegebenenfalls eine weitere Bildgebung nach einer zusätzlichen Chemotherapie oder Bestrahlung. Wir verwenden diese Bildgebungen in unserer Arbeit um Möglichkeiten und Verbesserungspotenzial aufzuzeigen.
Im ersten Teil dieser Dissertation befassen wir uns mit grundsätzlichen Aspekten der Therapieplanung und beginnen damit die aktuelle klinische Praxis zur Bestimmung des Tumorvolumens zu evaluieren. Da wir dafür einen aussagekräftigen Datensatz benötigen, erstellen wir zu Beginn einen Benchmark Datensatz. Dieser erlaubt es uns, diverse Aspekte zu betrachten: Zum einen ist es uns hiermit möglich die Abweichungen zwischen einzelnen menschlichen Experten zu quantifizieren. Zum anderen zeigen wir, dass die aktuelle klinische Praxis zur Bestimmung des Tumors sehr fehleranfällig ist.
Darüber hinaus bringt uns diese Datensammlung in die Position auch automatische und semi-automatische Verfahren zur Segmentierung und damit auch zur Quantifizierung des Tumorvolumens zu testen.

Wir gehen diesen nächsten Schritt und entwickeln einen semi-automatischen Ansatz zur Bildsegmentierung. Die Grundidee bei dieser Art Methode ist, dass ein menschlicher Experte eine grobe Annotation des Bildes in tumor- und nicht-tumor Bereiche vorgibt. Durch diese menschliche Interaktion ist es essentiell, dass sie unempfindlich gegenüber nicht optimalen oder auch falschen Informationen ist, die durch den jeweiligen menschlichen Experten vorgegeben werden. Wir zeigen diese Fähigkeit in diversen Kontexten: Von der direkten menschlichen Interaktion bis hin zu Kombinationen mit anderen (voll-automatischen) Prozess-Schritten.

Dieser robuste und intuitive Ansatz erlaubt es uns dann im zweiten Teil unserer Arbeit unseren Fokus auf die Differentialdiagnostik sowie die Therapieprognose zu legen. Zu diesem Zweck stellen wir einen weiteren Datensatz zur Verfügung. Dieser beinhaltet über 200

Patienten und bildet eine umfassende Basis der möglichen Tumorentitäten des Nephroblastoms ab. Ein weiterer wichtiger Aspekt dieser Datensammlung ist, dass ebenfalls über 50 Patienten mit einer Nephroblastomatose inkludiert werden konnten. Da diese überaus seltene Vorgängerlesion des Wilms' Tumors weder bösartig ist noch invasive Tendenzen zeigt, ist meist (ausser einer Extraktion) keine weitere Therapie notwendig. Offenkundig ist daher eine verlässliche Differentialdiagnostik dieser beiden Krankheiten unerlässlich. Wir haben uns dieses Problems angenommen und einen neuen Ansatz entwickelt, mit welchem wir diese beiden Krankheiten zuverlässig unterscheiden können.

Schlussendlich befassen wir uns damit die Entwicklung des Nephroblastoms unter dem Einfluss der Chemotherapie zu prognostizieren. Da sich während der Gabe von Chemotherapeutika der Wilms' Tumor in seiner Histologie verändert und gegebenenfalls zu aggressiveren Subtypen mutiert, ist es von immenser Bedeutung, diese Evolution zum frühst möglichen Zeitpunkt festzustellen um die Therapie dementsprechend anpassen zu können.
Wir haben zu diesem Zweck das durchschnittliche visuelle Erscheinungsbild aller Subtypen erstellt und es dazu benutzt einen Klassifizierungsalgorithmus zu trainieren. Obwohl unsere Versuche nur eine Machbarkeitsstudie zu diesem Thema darstellen, können wir durchaus erstaunliche Feststellungen machen. So ist beispielsweise in allen unseren Experimenten die Genauigkeit immer deutlich besser als das Zufallsniveau. Dies ist insbesondere bemerkenswert als dass die Bildgebungen von Wilms' Tumoren nicht standardisiert sind und dadurch eine deutliche Abweichung der MRT Parameter vorhanden ist.

Acknowledgments

First of all I would like to thank my supervisor Prof. Dr. Joachim Weickert. First, he has always provided outstanding support at the right time. Second, he has offered me a fruitful and inspiring working environment. I am very grateful for his excellent advice.
Next I would like to thank Prof. Dr. Norbert Graf. He has also arranged a pleasant and motivating working environment. Being a medical doctor and looking at situations from a different point of view from mine, his advice and views have always been very instructive and enlightening.

Of course I particularly thank my co-authors. Iva Farag deserves special mention here. Without her help, exceptionally in the area of manual registration of MRI sequences, some of the tasks would have become much more difficult. Especially Prof. Dr. Peter Ochs with his deep understanding of convex optimization has contributed a lot to give me insights in his interesting research area.
Prof. Dr. Margret Keuper was an important and very patient discussion partner in the area of machine learning for me. Even though we never had the opportunity to work together on a research question, the conversations with Dr. Pascal Peter and Dr. Matthias Augustin were also inspiring and brought me new ideas every time.
Dr. Pascal Peter as well as Dr. Dominik Straßel deserve here special thanks for proof-reading this work and giving many helpful suggestions.

However, there are still many people who have had an indirect influence on my thesis. My special thanks go to my current and former colleagues in the MIA group as well as to those from the paediatric oncology department. An important help was also Peter Franke. He fulfilled all my special requests and was always ready to find unique solutions for me and the large amounts of data I had to deal with.

Not unmentioned should be the essential contribution of Ellen Wintringer, secretary of the MIA Group, and Elisabeth Friedel, secretary of the pediatric oncology department: Both have tirelessly ensured that my interaction with the inevitable bureaucracy has been reduced to a minimum.

Finally, I would like to thank my family and friends. In some phases of this doctoral thesis, dealing with me was not that easy. I am very grateful for the friendship of Julia Jakobs and Davinia Eder and their invaluable support in every life situation. Also I would not like to miss Elena and Edyta Firmery - without their regular help in horse care some aspects would have been much more difficult. I am also very grateful to Heidemarie Wilhelm for taking care of my dog so often when I had too little time.

Last but not least, this work would not have been possible without my parents Brigitte and Otto Müller as well as my sister Nicole. Without their unconditional and continuous support and love I would not have been able to complete this thesis.

To my father.

Contents

1 Introduction

"But in my opinion, all things in nature occur mathematically."

– René Descartes

Contents

In recent decades, the influence of computer science and mathematics in the medical field has steadily increased. Since the first attempts to analyze medical aspects with the help of computers in the 1950's [116], their influence has intensified enormously. It has become common practice to use magnetic resonance tomography and computer tomography to create images of internal organs, or to plan or perform surgical interventions using algorithms and computer-assisted systems.
Modern medicine today is an interplay of different disciplines: from molecular biology to classical medicine, computer science and mathematics. Especially mathematical models enable a more precise analysis of the medical situation and an evaluation of the course of therapy due to the large amount of medical data collected.

A large number of researchers is dealing with the problem of cancer - one of the most frequent causes of death in our time. Progress is being made in the treatment of this serious disease: Developments in diagnosis, medical imaging, treatment plans and basic understanding have a significant impact on the number of cancer fatalities. Although the mortality rate has decreased every year since the 2000s, it has not yet been fully understood. There are over a hundred different types of cancer, usually named after the organ or cell type that formed them [125] and each of these varieties is in principle a different disease. Although scientists have developed theories about how this fatal illness is caused [72], much is still unknown.

Medical image processing and analysis play a major role in cancer research and enable the investigation of the structure as well as the function of the tissue to be analyzed. This includes various aspects: From segmentation to feature extraction and classification or measurement of anatomical and physiological parameters. The fields of research in this area are manifold. From image guided surgery [110], deformation analysis based on biomechanical models [81], to predictions of cancer extension and development [163].

At the moment, however, the majority of research is still dealing with cancers for which large amounts of data are available. This facilitates the development of appropriate algorithms and allows faster improvement in the treatment and course of therapy of a few types of cancer. Rare forms or varieties with data sets that are difficult to obtain are

unfortunately less considered leading to the prognosis remaining below its capabilities.

Children are fortunately under-represented in the group of cancer patients, but also receive less attention. In addition, it is generally difficult to collect a sufficiently large amount of data from children - both for data protection and practical reasons. It is difficult to explain to children or infants that further examination or imaging is necessary to get more information about the disease. Often they are also physically simply unable to do so - any anesthesia or additional medical burden should be avoided at all costs [78].

Nephroblastoma or Wilms' tumor is responsible for 5% of all childhood cancers and is also the most common malignant kidney tumor in childhood [137]. Since this disease is rare and almost exclusively affects children - about 75% of all patients are younger than five years with a peak between two and three years [47, 91] - it has received little or no attention in the field of medical image processing.
In recent decades, however, a consortium of many hospitals has managed to partially solve the problem of data quantity in Wilms' tumors. The studies of the International Society of Pediatric Oncology (SIOP) allowed the data collection of more than thousand patients with nephroblastoma [67, 88].

In the past, these studies primarily addressed the objective of optimizing drug therapy and therapy outcome. This included the modulation of different chemotherapeutic agents, duration and intensity of chemotherapy, and optimization of prognosis for different subtypes that may develop during the course of therapy. During this research, clinical information such as therapy progressions, clinical patterns and general patient information as well as imaging data were collected.
In particular, the imaging was only partially used: The only uses consisted of visual observation of the course of therapy, surgical planning with regard to the localization of the tumor and a rough estimation of the tumor volume.

In order to address the aforementioned issues, this work deals with nephroblastoma in childhood and addresses the current practice in Europe for the treatment of this type of cancer from a medical image processing perspective: Based on the collected data, we analyze the fundamental aspects: from tumor identification, to diagnosing and differentiating relevant diseases, up to prediction of tumor development.

1.1 Scope and Contributions

This work is dedicated to three overarching objectives: We want to improve therapy planning, facilitate differential diagnosis between nephroblastoma and its precursor lesion nephroblastomatosis, and enable an assessment of tumor development during chemotherapy. For each of these goals we proceed according to a similar scheme. First we analyze the current clinical practice, then we evaluate it for possible weaknesses and in a final step we point out possibilities for improvement.

Improvement of Therapy Planning Nephroblastoma usually changes its appearance and volume due to preoperative chemotherapy: some areas degenerate, others change their composition and yet others are resistant to the given chemotherapeutic agents. These processes influence the tumor volume, which to a certain extent reflects the response to therapy.
The tumor extension is currently determined by a reference radiologist using a specific MR image - depth, width and height are measured to approximate an elliptical shape. This information has an important role in the planning of the course of therapy: volume and the local stage determine whether and which further chemotherapy or radiation is necessary. A precise assessment is obviously essential for therapy planning. We therefore evaluate the following questions:

1. How large is the inter-rater variability? Is the measured tumor volume dependent on the radiologist/human expert? Is it generally valid?

2. Is the approximated elliptical shape correct? Are there deviations from a pixel-based measurement with the exact volume values? Is the approximated volume larger or smaller? Does it therefore indicate more or less additional therapies?

3. Is there a way to determine this volume automatically? Are there limitations?

In a first step, it is therefore necessary to create a data set that allows to address these questions. We have compiled a patient cohort whose MRI sequences have been annotated by human experts. From this information we calculated a consensus truth (or ground truth) and evaluated the quality of the individual annotators, the main sources of error and their deviations from each other.
Based on this ground truth, we were also able to evaluate the accuracy in clinical practice, i.e. the approximation of the tumor with an elliptical shape. This heterogeneous and diverse benchmark data set also allows us to evaluate the segmentation performance of fully automated algorithms and to define a baseline. Finally, we also designed and evaluated a semi-automated approach specifically adapted to Wilms' tumors, which is robust to the user's different prior knowledge.

Reduction of False Diagnoses The Wilms tumor is just one of several diseases and abnormalities that can affect the kidneys. Some can be clearly distinguished visually, while others are difficult to differentiate. In particular, nephroblastomatosis, a precursor

lesion of nephroblastoma, has a very similar appearance.
Despite their visual similarity, they differ fundamentally: While the nephroblastoma is malignant and has invasive tendencies, the nephroblastomatosis is neither. Chemotherapy is not necessary in the situation of nephroblastomatosis and the object can be removed directly without further treatment. In the case of a nephroblastoma, however, the renunciation of chemotherapy can be fatal, as the probability of a tumor rupture during surgery increases.
A precise distinction between these two diseases is of immense importance. In current clinical practice it is assumed that nephroblastomatosis is a rather small and homogeneous mass. We put these criteria to the test and answer the following questions:

1. Are the assumptions of size and homogeneity correct to the extent that they allow a reliable and accurate classification?

2. Can we identify properties of nephroblastomatoses that facilitate differentiation?

3. Is there a way to automate this classification?

To answer these questions, we have compiled an extensive data set. It contains a comprehensive database of a more than 202 kidneys before chemotherapy, of which 148 have a nephroblastoma and the remaining 52 have a nephroblastomatosis. It contains 7 of the 9 occurring subtypes (the remaining two are very rare and imaging was unfortunately not available) and is therefore an almost complete visual representation of all possible entities of nephroblastoma.
We have annotated all these affected kidneys with our semi-automatic segmentation approach. This enables us for the first time to make mathematically valid statements about the appearance of nephroblastomatosis in MR images. First, we analyze the current methodology of visual differential diagnosis between Wilms' tumor and nephroblastomatosis. In addition, we propose a way to automate this differentiation while providing a significant improvement in classification accuracy and noise robustness. This allows a reliable differential diagnosis and can thus ensure that the necessary therapy steps are taken for both diseases.

Prediction of the Course of the Disease In the last part of this thesis we deal with the prediction of subtype development. Nephroblastoma is a solid tumor consisting mainly of three tissues: blastema, epithelium and stroma [182]. One of the most important aspects of the treatment protocol for nephroblastoma is a preoperative chemotherapy. During this therapy, the tumor tissue changes and a total of nine different subtypes can develop. Depending on this and the local stage, the patient is divided into one of the risk groups (low, medium or high risk patients) and the further therapy is adapted accordingly.
Of course, it would be of crucial importance for therapy and treatment planning to determine the appropriate subtype as early as possible. In view of analysing this problem, we have addressed the following questions:

1. What is the current approach to predict this tumor development?

2. Can we make any statements about this development with the help of the available data?

3. What would be necessary to make accurate statements? How far can we get with the current data situation? Can trends be identified?

The prediction of tumor development is a complex investigation. Each person responds differently to chemotherapy and many and partly unknown factors contribute to how intensively a patient responds to a therapy. Accordingly, the development of the tumor under the influence of chemotherapeutic agents is a complex process.
We have used the comprehensive data set on differential diagnostics that we created for this research. As it contains almost all tumor entities, it also allows us first attempts to identify differences between subtypes and to differentiate between groups.
We start with a comprehensive analysis of the existing imaging with respect to simple texture patterns. Since these do not contain clear patterns that allow differentiation, we build a more complex model and represent each tumor as a visual model of an average subtype of each class, making remarkable observations. We can show that different entities can be distinguished from each other already at diagnosis, i.e. before any therapy. We evaluate these observations and show a way to improve the prediction in the future.

1.2 Organization

The structure of this work begins with the reasons for our research and the goals we want to achieve.
In the second chapter we introduce the notations and definitions we use in this work and discuss the basics about Wilms' tumors and image processing that are necessary to follow our investigations. We then summarize the related work in Chapter 3.

We have divided our contributions into individual chapters that can be considered independently of each other. Nevertheless, they build on each other and are interdependent. Chapter 4 presents the data sets we use. This includes a data set for the segmentation of nephroblastoma, as well as comprehensive data sets for differential diagnosis of nephroblastomatosis, and the subtype classification of Wilms' tumor after chemotherapy.
Subsequently, in Chapter 5, we analyze both the variability between individual human experts and the quality of the measure currently used to determine the extension of a Wilms' tumor.

In Chapter 6, we present our interactive method, which allows accurate and intuitive annotation of tumor tissue, and evaluate it against the previously presented data set. Additionally, we introduce a baseline algorithm for Wilms' tumor segmentation and show performances of several fully automatic approaches. Immediately afterwards, we highlight weaknesses of deep neural networks for medical image segmentation. In addition, we also suggest a simple scheme to robustify their generalization performance. Furthermore, we show that our semi-automatic segmentation approach is an easy-to-apply post-processing step to improve their segmentation accuracy.

Chapter 8 deals with the differential diagnosis between nephroblastoma and nephroblastomatosis. We analyze the current clinical practice and demonstrate how to improve the actual situation.

In Chapter 9, we then evaluate the extent to which we can predict the development of a Wilms' tumor under chemotherapy. We close this work at the end with a summary and further research directions in Chapter 10.

2 Foundations

"Begin at the beginning", the King said, very gravely, "and go on till you come to the end: then stop."

– Lewis Carroll, Alice in Wonderland

Contents

This chapter describes the basics of our work. As it contains both mathematical and medical aspects, we first explain the mathematical and natural scientific basics and then the medical ones. Therefore, at the beginning we illustrate the representation of discrete images as continuous functions. Subsequently we explain medical images and MRI images in particular.
Since we often face inverse problems and convex optimization during our work, we afterwards introduce the basic concepts of these two topics in Section 2.3. First we define the nature of inverse problems and explain in detail how to tackle their inherent difficulties. Next, we present their minimization by means of convex optimization strategies.
We close the mathematical foundations of this chapter with error measures in Section 2.4.We use these metrics to evaluate several methods of segmentation and classification in subsequent chapters.
Finally we give a comprehensive explanation of Wilms' tumors and their treatment in Section 2.5.

2.1 Definitions and Notations

First, we define the mathematical formulations we will need in the further course of this work. For simplicity, we list our conventions concerning the spelling of scalars, vectors, matrices or tensors and functions in tabular form to ease understanding, see Tab. 2.1.

Table 2.1: Our conventions of mathematical formulations.

Type	Convention	Example
scalar	-	$\lambda \in \mathbb{R}$
vector	bold	$\boldsymbol{x} \in \mathbb{R}^n$
matrix/tensor	bold, capitalized	$\mathbf{A} \in \mathbb{R}^{n \times n}$
matrix/tensor entries	-	$a_{1,1} \in \mathbb{R}$
function	bold, italic	$\boldsymbol{f} : \Omega \to \mathbb{R}$
Set	capitalized	$\Omega \in \mathbb{R}^3$

Mostly we want to approximate a continuous function on a discrete voxel grid. We define the gradient operator applied to a sufficiently often differentiable function $u \in \mathbb{R}^N$ as

$$\nabla := \begin{pmatrix} \partial_x \\ \partial_y \\ \partial_z \end{pmatrix}$$

$$\begin{aligned}
(\partial_x u)_{i,j,k} &:= \begin{cases} \frac{u(i+1,j,k)-u(i,j,k)}{h_x} & : i \in \{1 \dots n_x - 1\}, j \in \{1 \dots n_y\}, k \in \{1 \dots n_z\} \\ 0 & : i = n_x, j \in \{1 \dots n_y\}, k \in \{1 \dots n_z\} \end{cases} \\
(\partial_y u)_{i,j,k} &:= \begin{cases} \frac{u(i,j+1,k)-u(i,j,k)}{h_y} & : i \in \{1 \dots n_x\}, j \in \{1 \dots n_y - 1\}, k \in \{1 \dots n_z\} \\ 0 & : i \in \{1 \dots n_x\}, j = n_y, k \in \{1 \dots n_z\} \end{cases} \\
(\partial_z u)_{i,j,k} &:= \begin{cases} \frac{u(i,j,k+1)-u(i,j,k)}{h_z} & : i \in \{1 \dots n_x\}, j \in \{1 \dots n_y\}, k \in \{1 \dots n_z - 1\} \\ 0 & : i \in \{1 \dots n_x\}, j \in \{1 \dots n_y\}, k = n_z \end{cases}
\end{aligned} \tag{2.1}$$

where we assume reflecting Neumann boundary conditions, i.e. the gradient vanishes at image boundaries. Hence, we define the discrete gradient operator ∇ for a discrete image $u \in R^N$ as

$$(\nabla u)_{i,j,k} = \begin{pmatrix} (\partial_x u)_{i,j,k} \\ (\partial_y u)_{i,j,k} \\ (\partial_z u)_{i,j,k} \end{pmatrix} \quad , \forall i \in \{1 \dots n_x\}, j \in \{1 \dots n_y\}, k \in \{1 \dots n_z\}. \tag{2.2}$$

In order to improve readability, we typically write u_x instead of $\partial_x u$.
In some of our image analysis tasks, we also utilize the gradient direction. For a two dimensional image, it is given by

$$\boldsymbol{g}_{\text{dir}} = \arctan \frac{\boldsymbol{u}_x}{\boldsymbol{u}_y}. \tag{2.3}$$

Since we work in Euclidean vector spaces, i.e. $\mathbb{R}^N$, the inner product is defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{N} x_i y_i \quad , \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N. \tag{2.4}$$

This induces also the (Euclidean) norm:

$$\|\boldsymbol{x}\|_2 := \sqrt{\langle \boldsymbol{x}\,, \boldsymbol{x} \rangle} = \sqrt{\sum_{i=1}^{N} x_i^2}. \tag{2.5}$$

In the following, we make often use of the *gradient magnitude*. It can be defined with the Euclidean norm as

$$\|\nabla \boldsymbol{u}\|_2 := \sqrt{\boldsymbol{u}_x^2 + \boldsymbol{u}_y^2 + \boldsymbol{u}_z^2}. \tag{2.6}$$

On the basis of the Euclidean norm, we can also define the *Frobenius norm* as

$$\|A\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}. \tag{2.7}$$

Using the inner product, we define the *divergence*, i.e. the negative adjoint operator of the gradient [33]

$$\text{div } \boldsymbol{u} = \nabla * \boldsymbol{u} = \sum_{i=1}^{N} \partial_i \boldsymbol{u}. \tag{2.8}$$

The *Jacobian matrix* of $\boldsymbol{f} : R^3 \to \mathbb{R}^m$ is then the compound of all first-order derivatives

$$\mathbf{J}(\mathbf{f}) = \begin{bmatrix} \partial_x \boldsymbol{f} & \partial_y \boldsymbol{f} & \partial_z \boldsymbol{f} \end{bmatrix} = \begin{bmatrix} \partial_x \boldsymbol{f}_1 & \partial_y \boldsymbol{f}_1 & \partial_z \boldsymbol{f}_1 \\ \vdots & \vdots & \vdots \\ \partial_x \boldsymbol{f}_m & \partial_y \boldsymbol{f}_m & \partial_z \boldsymbol{f}_m \end{bmatrix}. \tag{2.9}$$

In some of our experiments, we add so-called *Gaussian noise* to our data [106]. Its probability density function is equivalent to that of the normal distribution - which is therefore also named as Gaussian distribution. Given a Gaussian random variable z it is defined as

$$\mathcal{P}_G(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}, \tag{2.10}$$

where z is the intensity value, μ the mean and σ the standard deviation, respectively. Intuitively, adding Gaussian noise means that the noise follows the Gaussian distribution, i.e. the values are Gaussian-distributed.

In the context of continuous convex optimization, we heavily rely on several mathematical principles, sketched in the following. For a detailed explanation, we refer to [145].

Let $\mathcal{X} \subseteq \mathbb{R}$ be a vector space, and $V \subseteq \mathcal{X}$. We call V a *convex set*, if and only if

$$t\boldsymbol{x} + (1-t)\boldsymbol{y} \in V, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in V, \forall t \in [0,1]. \tag{2.11}$$

A function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}$ within a convex set V is convex if

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X, \forall t \in [0,1], \tag{2.12}$$

and strictly convex if

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) < tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X, \forall t \in [0,1]. \tag{2.13}$$

The main difference between convex and strictly convex functions is their amount of global minima: While a strictly convex function has a unique global optimum, a convex one can have multiple minima. Convexity is preserved under several transformations:

- The sum of convex functions $\boldsymbol{f}_i$ is a convex function, i.e. $\boldsymbol{g} = \sum_i \boldsymbol{f}_i$ is convex.
- The maximum over convex functions is again a convex function, i.e. $\boldsymbol{g} = \max\{\boldsymbol{f}_1 \dots \boldsymbol{f}_n\}$ is convex.
- Convexity is invariant under affine mappings, i.e. $\boldsymbol{g} = f(\mathbf{A}\boldsymbol{x} + b)$ is convex.
- the pointwise supremum of the collection of all affine functions $\boldsymbol{g}$ such that $\boldsymbol{g} \leq \boldsymbol{f}$ is a closed convex function $\boldsymbol{f}$ [145].

A function $\boldsymbol{f}$ is upper (lower) semi-continuous for a point $\boldsymbol{x}_0$ if for small $\boldsymbol{\epsilon}$, $f(\boldsymbol{x}_0 - \boldsymbol{\epsilon})$ is either close or less (greater) than $f(\boldsymbol{x}_0)$; see Fig. 2.1.

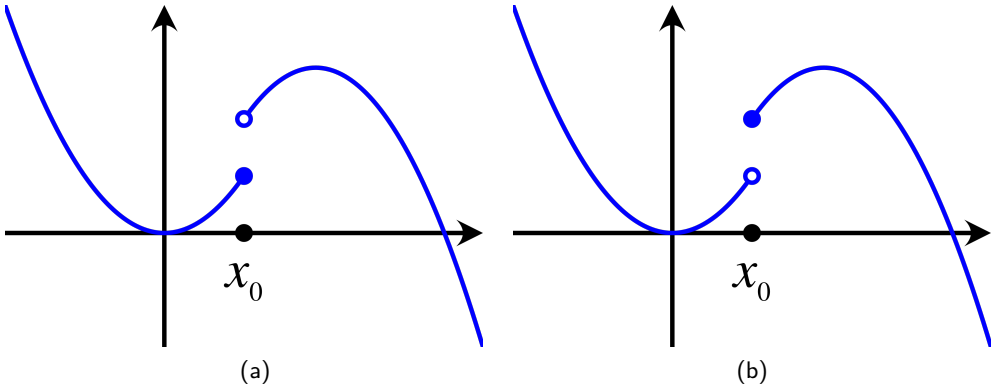


Figure 2.1: Lower and upper semi-continuous functions. (a) lower semi-continuous function, (b) upper semi-continuous function. Image courtesy of Wikipedia [117].

Let $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}$ be a lower semi-continuous function, where $\mathcal{X} \in \mathbb{R}^n$ is a nonempty convex subset of $\mathbb{R}^n$. The *convex envelope* over $\mathcal{X}$ is a function

$$\mathbf{g} : \mathcal{X} \to \mathbb{R}, \quad g(\boldsymbol{x}) \leq f(\boldsymbol{x}) \; \forall \; \boldsymbol{x} \in \mathcal{X} \tag{2.14}$$

such that $\boldsymbol{g}$ is a convex function defined over the set $\mathcal{X}$. If $\boldsymbol{h}$ is any other convex function such that $h(\boldsymbol{x}) \leq f(\boldsymbol{x}) \; \forall \; \boldsymbol{x} \in \mathcal{X}$ then $h(\boldsymbol{x}) \leq g(\boldsymbol{x}) \; \forall \; \boldsymbol{x} \in \mathcal{X}$.
Hence, $\boldsymbol{g}$ is uniquely determined and the pointwise supremum among any convex underestimators of $\boldsymbol{f}$ over $\mathcal{X}$.
In the context of primal-dual optimization, see Sec. 2.3.2, we also need the Legendre-Fenchel *convex conjugate* [145]. It is defined in terms of the supremum for a function $\boldsymbol{f} : \mathbb{R}^{d \times k} \to \mathbb{R}$ as

$$f^*(\boldsymbol{y}) := \sup \left\{ \langle \boldsymbol{y}, \boldsymbol{x} \rangle - f(\boldsymbol{x}) | \, \boldsymbol{x} \in \mathbb{R}^{d \times k} \right\}. \tag{2.15}$$

For the convex conjugate, the following properties hold:

- The convex conjugate is always a closed convex function.
- Convex conjugation inverts the ordering: $\boldsymbol{f} \leq \boldsymbol{g} \Rightarrow \boldsymbol{f}^* \geq \boldsymbol{g}^*$.
- Slopes in $\boldsymbol{f}$ correspond to points in $\boldsymbol{f}^*$ and vice versa.
- The double convex conjugation $\boldsymbol{f}^{**}$ is the convex envelope, i.e. if $\boldsymbol{f}$ is convex then $\boldsymbol{f}^{**} = \boldsymbol{f}$.

- The convex conjugate is always lower semi-continuous.

The *proximal operator* [118,146] is another essential building block of convex optimization approaches. It is defined as

$$\text{prox}_{\boldsymbol{f}}(\boldsymbol{v}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \left(f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 \right) \tag{2.16}$$

where $\|.\|_2$ denotes the Euclidean norm. Please note that solving for the proximal operator corresponds to a minimization of $\boldsymbol{f}$ in the neighborhood of $\boldsymbol{x}$. Thus it involves the evaluation of a convex optimization problem where $\boldsymbol{f}$ can be non-smooth. For this transformation, the following properties hold [14]:

- Let $\boldsymbol{f}$ be closed and convex. Then $\text{prox}_{\boldsymbol{f}}$ exists and is unique.
- $\text{prox}_{\lambda \boldsymbol{f}^*}(\boldsymbol{x}) = \boldsymbol{x} - \lambda\ \text{prox}_{\lambda^{-1}\boldsymbol{f}}(\lambda^{-1}\boldsymbol{x})$,
- Let $\boldsymbol{y} = \text{prox}_{\boldsymbol{f}}(\boldsymbol{y})$, i.e. $\boldsymbol{y}$ is a fixed point. Then $\boldsymbol{y}$ minimizes $\boldsymbol{f}$.

2.2 Images

In our work we deal with volumetric three dimensional gray scale MRI images where each point in a spatial 3D coordinate system is assigned a gray value. This can also be represented as a mapping of a continuous function $\boldsymbol{f}$ with

$$\boldsymbol{f} : \Omega \to R. \tag{2.17}$$

Here, $\Omega \subset \mathbb{R}^3$ is the cubic image domain and $R \subset \mathbb{R}$ denotes the range of intensity values. In practice, the images captured by a digital recorder are always discrete, i.e. data is only available at cuboid grid points within the image domain Ω. Thus each cell represents a single image element - in a 2D image these are referred to as pixels, in 3D as voxels. The co-domain R is also quantized analogously, since each acquisition system has only a limited number of sensors.
Usually, an image is defined on a regular Cartesian grid. Suppose h_x, h_y and h_z represent the grid sizes (or spacings) in horizontal, vertical and depth directions. If we further assume that n_x, n_y and n_z indicate the number of voxels in each of these directions, then the value of each voxel (i,j,k) with $i = 1 \dots n_x$, $j = 1 \dots n_y$, $k = 1 \dots n_z$ can be determined by sampling the function $\boldsymbol{f}$ at position $\boldsymbol{f}_{i,j,k} = (h_x(i-\frac{1}{2}), h_y(j-\frac{1}{2}), h_z(k-\frac{1}{2}))$, see Fig. 2.2.

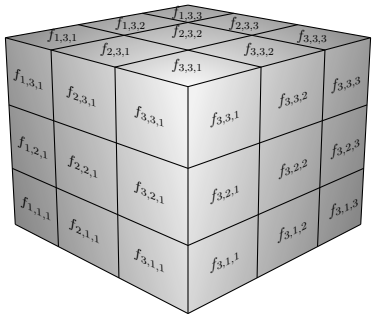


Figure 2.2: Exemplary image cube of $3 \times 3 \times 3$ of a continuous function $\boldsymbol{f}$.

In our notation, images are formally represented as tensors. However, each **n**-dimensional image can be transformed to a one dimensional vector by concatenation of the dimensions. In our case, we first append in depth $k = 1$ all rows successively, then repeat the procedure for the successive depth image $(k + 1)$. We work with both representations without mentioning it explicitly - it will be clear from the context: whenever the index is a triplet, the grid is defined as mentioned above.
The extension of this approach from a single 3D image, e.g. a MRI sequence, to a collection of several images or sequences is straightforward: Instead of scalar-valued functions f, we consider vector-valued functions $\boldsymbol{f} : \Omega \to R^n$ where R^n represents the range of intensity values within each image or sequence $c \in \{1 \dots n\}$. The notation follows directly, since we concatenate the vectors representing the different sequences.
Commonly, an image is stored with 8 bits, i.e. the gray values lie in the interval $\{0, 255\}$.

Unfortunately, this assumption is not always valid for MRI images, as the range depends on the respective parameter settings. For simplicity, we therefore linearly rescale all sequences to this interval in the following.

2.2.1 Medical Images

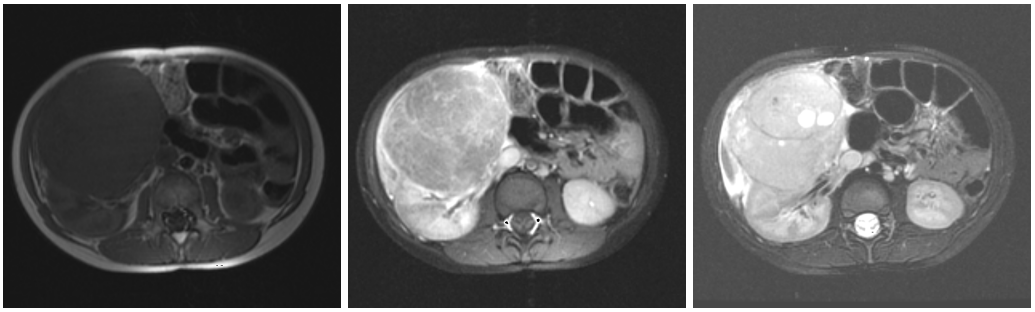
Medical images differ from photographs in many aspects. The analysis of a simple photograph can, for example, deduct to identify the objects on it, restore 3D information, separate light effects from object appearances, process partially hidden objects or track objects over time - to name just a few, see. Fig. 2.3.

(a) standard image (b) medical image

Figure 2.3: Comparison of a standard photography and a medical image. (a) standard image, (b) medical image. While the photography displays effects of light reflections, the medical image (MRI) is based on magnetic fields and radio waves. Image courtesy of It's interesting [155].

Medical images are special. They vary from ordinary images in representing distributions of different physical characteristics measured on the human body and show attributes that are otherwise not accessible. In addition, the analysis of such images is based on very specific expectations that led to the images being taken. This affects the nature of the analysis and the requirements for algorithms that perform some or all of the analysis.

Consider the right image in Fig. 2.3. The appearance of the object is not created by light reflection, but by strong magnetic fields and radio waves. Although the detection



(a) T_1 (b) T_2+ contrast (c) T_2

Figure 2.4: Examples of different MRI sequences. (a) T_1 sequence, (b) T_1 sequence with contrast agent, (c) T_2 sequence. Each kind of MRI sequence visualizes the same phyiscal properties in different ways.

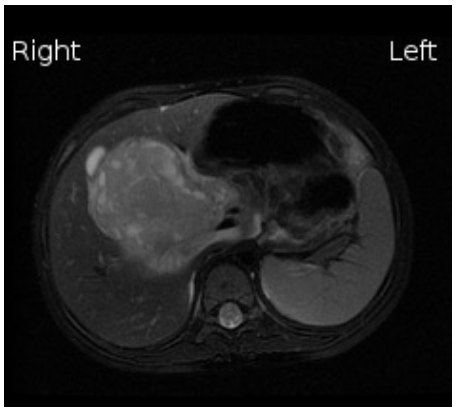


Figure 2.5: Exemplary abdominal MRI scan (axial) with a right-sided kidney tumor.

of a structure can be the goal of the analysis, the exact description of the object and its substructures can be the first task. Different imaging techniques or even variations in the method used produce images of several physical properties in different ways that can be the subject of inspection (see Fig. 2.4). However, it is difficult to compare this information with reality because there are few or no non-invasive methods to verify the information obtained from the images. This causes a fundamental problem: there is no measurable ground truth - it must be approximated.

The first step in reviewing medical or radiology images is the knowledge of the spatial arrangement. Basically, the image is always arranged as if the viewer were standing in front of the patient. This results in the images always being reversed. The same convention applies to the inspection of an axial or coronal image with the patient's feet pointing towards the viewer. Fig. 2.5 shows an exemplary magnetic resonance image in a

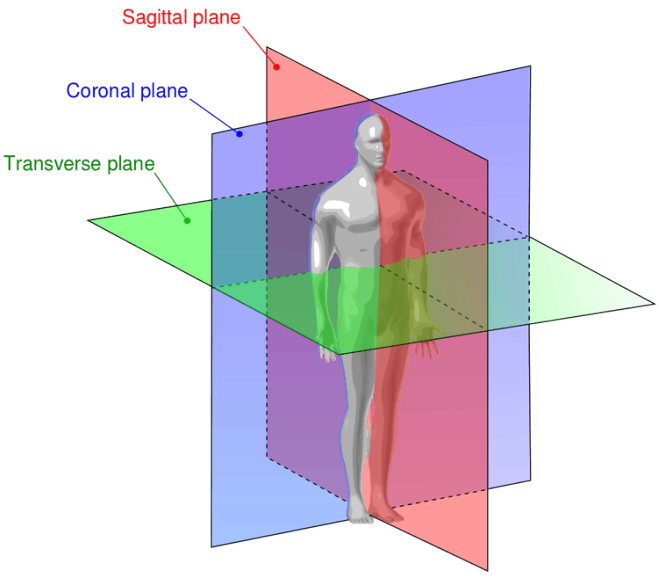


Figure 2.6: Anatomical imaging planes. Image courtesy of Wikimedia [120].

standard fashion visualizing a right-sided kidney tumor. In total, there are three primary imaging planes that are utilized in renal medical imaging and neuroimaging, see Fig. 2.6:

- Axial or transverse plane: Transverse images represent "slices" of the body
- Sagittal plane: Images taken perpendicular to the axial plane separating the left and right sides.
- Coronal plane: Images taken perpendicular to the sagittal plane separating the front from the back.

In our work, we have mainly considered axial MRI sequences and refer to this imaging plane when we mention imaging slices. Of course, all our methods can be directly tranfered to other directions of MR acquisition.

2.2.2 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a commonly used form of medical imaging. Since it is non-invasive, it is often used in radiology to represent internal organs such as the liver, kidney or brain. This technique is based on the fact that different types of tissue behave differently under the influence of a strong magnetic field. It was originally developed independently by P. Lauterbur [95] and P. Mansfield, who derived it from the Nuclear Magnetic Resonance Imaging discovered by I. Rabi [141] and further developed by Bloch and Purcell [22].
The two most important advantages of this method are on the one hand its very good soft tissue contrast and on the other hand the fact that no ionising radiation is used, such as in computer tomography. In addition, images can be taken from any direction.

How does it work?

MRI exploits the principle that atomic nuclei with an odd number of protons or neutrons have an intrinsic angular moment - this so-called spin transforms these nuclei into small magnets themselves. Hydrogen nuclei are particularly important as they are most abundant in the human body.
Under natural conditions, the magnetic orientation of the hydrogen nuclei (protons) is completely random. As soon as they are exposed to a strong magnetic field B_0, however, they align themselves parallel or anti-parallel to the field direction and perform a gyroscopic movement around the field lines (precessional movement). The frequency of this movement, the Larmor frequency, is related to the strength of the magnetic field.
The alignment of the nuclear spins alone would not be sufficient to create an image representation: This is accomplished by a short high frequency pulse emitted perpendicular to the direction of the magnetic field. The frequency of this pulse, the resonance frequency, corresponds to the Larmor frequency of the hydrogen protons.
Firstly, this results in a deflection of the nuclear spins aligned with the static outer magnetic field. Secondly, the precessional movement of all atomic nuclei is briefly synchronized (excitation), i.e. phase-coherent. This leads to a transverse magnetization perpendicular to the field lines of the external magnetic field.

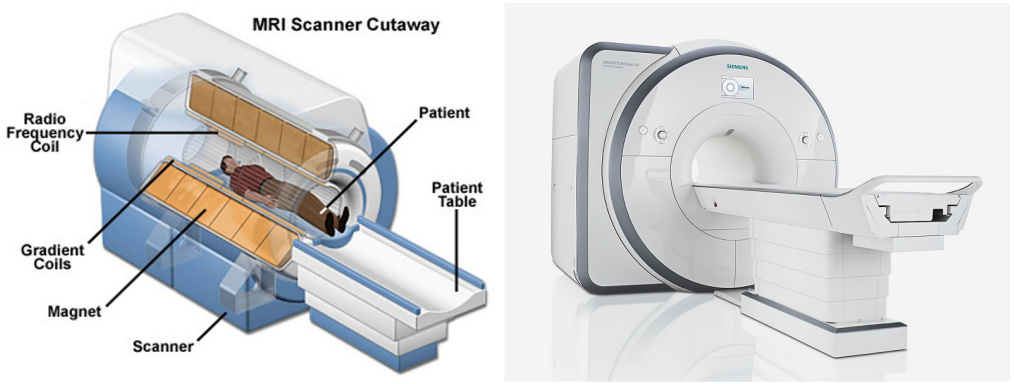


Figure 2.7: Exemplary MR Scanner. Left: Schematic representation of a MR machine. Right: Exemplary MR scanner. Image courtesy Siemens healthcare [158].

After the impulse, the spins again align themselves with the magnetic field (spin-lattice relaxation) and emit exponentially decaying radio waves, the spin-echo, in terms of thermal energy at larmor frequency to the environment.
This process of reestablishing longitudinal magnetization is called T_1 relaxation. Since it depends essentially on the thermal conductivity of the tissue, tissues with rapid heat transfer, e.g. fat, appear bright in T_1-weighted images, while poorly conducting tissues such as water appear dark.
Almost immediately after the pulse, the nuclear spins lose coherence, as some spin a little faster than the others (*dephasing*). This loss of coherence of the spin system attenuates the signal with a time constant, the so-called transverse relaxation time (T_2 relaxation).
The spatial encoding of the MR signal is achieved with the support of small magnetic fields, the gradients. These disturb the main magnetic field and cause hydrogen protons to precess at different locations at a slightly different rate. The gradient coil section and the associated electric field perpendicular to the main magnetic field cause a force (Lorentz force) on the coils. The gradients are switched on and off very quickly so that they oscillate and generate most of the noise associated with the MRI environment, even though they are usually embedded in epoxy resin.
While the protons relax, the change in the local magnetic fields generates currents in the receiving coils. These currents can be detected as voltage changes. The signal must then be sampled, the analog signal converted into a digital one and then stored for processing. Afterwards, the MRI signal is resolved and spatially localized to produce images.
An MR scanner therefore consists of several components, see Fig. 2.7:

- a strong magnet to align the nuclei. The strength of this field is measured in teslas (T). The majority of MRI systems in clinical use are 1.5T or 3T machines.
- a radiofrequency (RF) system to emit the radio pulses,
- gradient coils to generate the field gradients and spatially encode the resonance signals,
- receiver coils for registering the generated resonance signal,
- shim coils, to provide localised auxiliary magnetic fields and field homogeneity,
- and a computer to control the entire system, and to calculate the images.

In clinical procedures, several types of MRI sequences are usually performed one after the other in order to map different properties of the present tissue types. The two most important parameters are the echo time (TE) and the repetition time (TR). The first one is responsible for how much T_2 relaxation is present in the imaging. It refers to the time between the application of the radiofrequency pulse to excitation and the peak of the signal induced in the coil, measured in milliseconds.
The repetition time, also measured in milliseconds, is the time between the application of excitation pulses and determines how much longitudinal magnetization can form again between the pulses.
In summary, T_1 relaxation is the recovery of magnetization along the *longitudinal* axis and T_2 relaxation is the decay of magnetization along the *transverse* axis while each proton has unique T_1 and T_2 relaxation times. Both relaxation times are independent and occur in parallel.

Signal to Noise Ratio and Image Contrast

Tissue properties as well as T_1 and T_2 relaxations significantly influence the contrast and signal-to-noise-ratio (SNR) ratio of a MR image. T_1 relaxation is a measure of how fast the net magnetization vector recovers to its ground state in the direction of the outer magnetic field. The decline of the excited atomic nuclei from the high-energy state to the low-energy or ground state is associated with a loss of energy to the surrounding atomic nuclei. This is an exponential process [22], where the length of the net magnetization vector is given by

$$M_z = M_{\max}(1 - \exp\left(\frac{-t}{T_1}\right)) \tag{2.18}$$

where M_z is the magnetization at time = t (the time after the 90° pulse), $M_{\max}$ is the maximum magnetization at full recovery. At a time = T_1, the signal will recover to 63% of its initial value after the RF pulse has been applied, see Fig. 2.8. After two T_1 times, the magnetization is at 86% of its original length, while three T_1 times gives 95%. Spins are considered completely relaxed after 3-5 T_1 times. In biological material, it can range from a few tenths of a second to several seconds.

Consequently, improvements in the SNR occur particularly when the repetition time (TR) is significantly increased by 3 to 5 times the T_1 time - the longitudinal magnetization has time to recover. At the same time a change of the TR time influences the T_1 weighting of an image but also the acquisition time: Short TR spin echo sequences have a stronger T_1 weighting since the longitudinal magnetization can only be restored incompletely. The effect of TR on SNR can be shown graphically by a T_1 relaxation curve as illustrated in the exponential growth curve in Fig. 2.8.
The recovery of net magnetization vectors is fastest when the protons' motion (rotations and translations) is according to the Larmor frequency, induced by the outer magnetic field. As a result, stronger magnetic fields are associated with longer T_1 times.
Wile T_1 refers to the recovery of the excited system to the state of thermal equilibrium, T_2 represents the decay of the spin synchronization and is nearly independent of the outer magnetic field [23] following an exponential decay, see Fig. 2.8:

$$M_{xy} = M_{max} \exp\left(\frac{-t}{T_2}\right) \tag{2.19}$$

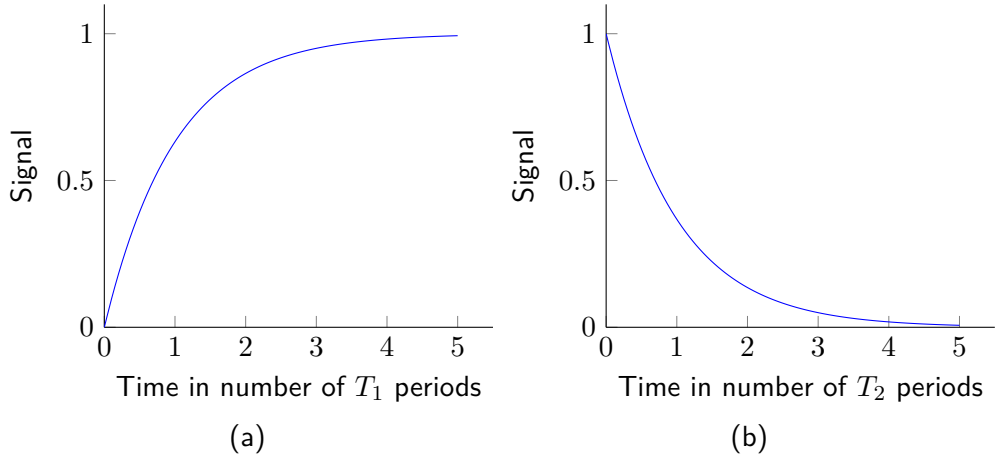


Figure 2.8: Exemplary relaxation curves. (a) T_1 Relaxation, (b) T_2 Relaxation.

Increasing the echo time TE results in a decreased SNR: there is more time for dephasing to occur and therefore lower signal intensities. Nevertheless, TE is deliberately increased to improve the contrast, i.e. the T_2 weighting of an image.

Sequences

The easiest way to classify the multitude of available sequences of modern MRI scanners is to distinguish them according to the visual appearance of tissue. This results in a distinction between T1- and T2-weighted, diffusion-weighted, proton density (PD) weighted, flow-sensitive and "other". Since we only use different variants of T_1 and T_2 weighted images in our work, we do not go into the remaining sequence types here, and refer to [121] and the references therein. In addition, there are a number of optional "plugins", such as fat or fluid attenuation or contrast enhancement.

In the description of most MRI sequences, we refer to the grey value of tissue or fluid, which leads to the following absolute terms:

- high signal intensity = bright
- intermediate signal intensity = intermediate-bright
- low signal intensity = dark

Often clinicians refer to the appearance by relative terms in comparison to a neighbouring object. In this way, hyperintense means brighter, isointense equal brightness, and hypointense means darker than the comparing tissue.
Unfortunately, these relative terms are applied without reference to the tissue being used for comparison. In many situations this does not result in a problem, e.g. a hyperintensive lesion in a kidney is clearly hyperintensive compared to the surrounding renal parenchyma. In other scenarios, however, it can be confusing: Assuming a lesion in the ventricles of the brain is to be visualized. The corresponding image shows the brain, the lesion and the cerebrospinal fluid. If the lesion is now called hypointense, it is ambiguous whether it is darker than the brain or the liquor.

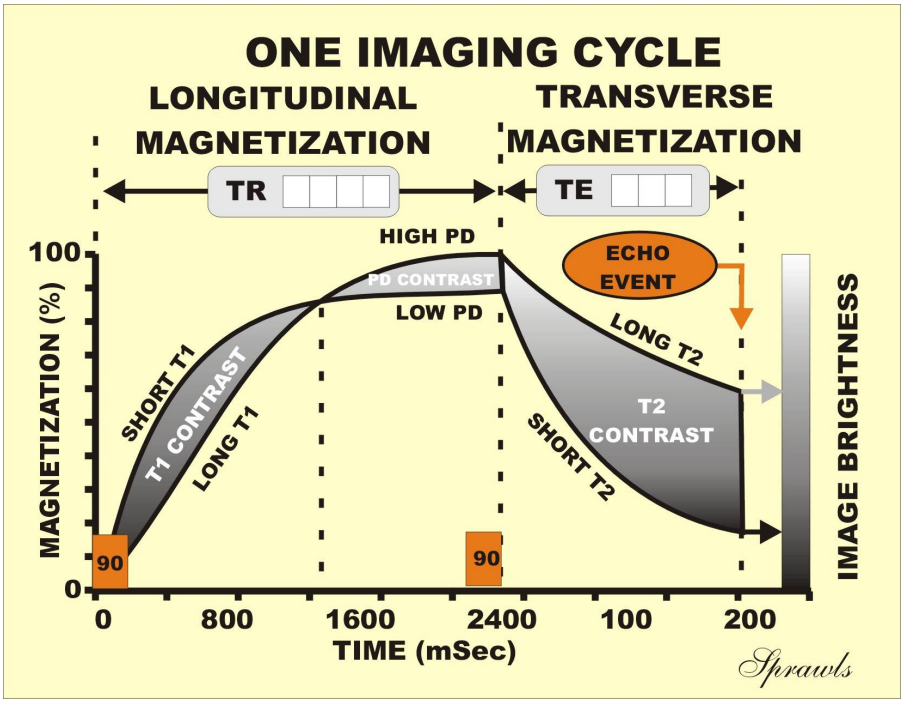


Figure 2.9: Sketch of an imaging cycle of T_1 weighted, proton density (PD), and T_2 weighted acquisition. Image courtesy Perry Sprawls [162].

For this reason, we prefer absolute terminology and, in the few cases where we use relative terms, indicate the tissue to be compared.
In order to adequately evaluate a tissue, several sequences are usually required, and the combination of sequences is called the MRI protocol.

Spin echo sequences Spin echo pulse sequences are among the earliest developed MRI pulse sequences, but are still widely used in their fast spin echo form. The timing of the pulse sequences can be varied to generate both T_1 and T_2 weighted and proton density weighted images, see Fig. 2.9.
The two important parameters are echo and repetition time. As a result, T_1 and T_2 weighted images are the most common sequences. While the former show differences in T_1 relaxation, the latter exhibit variations in T_2 relaxation.

Table 2.2: Properties of T_1 and T_2 weighted images.

	T_1 weighted image	T_2 weighted image
Repetition time	short	long
Echo time	short	long
Fat	bright	intermediate-bright
Fluid	dark	bright

T_1 sequences usually have short repetition and echo times, see. Tab. 2.2. Assuming the repetition time would be long, then all protons would have enough time to restore their alignment to the external magnetic field and the image would have an equal intensity. By a repetition time, which is lower than the regeneration time of the corresponding tissue, a distinction, i.e. a tissue contrast, is possible. For example, fat realigns quickly to the external magnetic field as it has a low T_1 relaxation time. In contrast, water aligns very slowly along this field and has a lower signal intensity.
T_1-weighted sequences also provide the best image contrast for paramagnetic contrast agents (e.g. gadolinium-containing compounds).

Similarly, T_2 weighted images normally have long echo and repetition times, see. Tab. 2.2. In contrast to T_1 weighted images, paramagnetic contrast agents do not result in a similar bright tissue contrast: Gadolinium shortens T_2 relaxation times and causes a low (hypointense) signal intensity. If the echo time is extended tremendously, only tissues with a very long T_2 relaxation time will continue to emit a signal.

Inversion recovery pulse sequences Inversion recovery pulse sequences are a type of MRI sequence used to selectively eliminate the signal for certain tissues (e.g. fat or fluid). However, inversion recovery can also produce highly T_1-weighted images and was originally developed for this purpose.

In principle, an inversion recovery pulse sequence is a spin echo pulse sequence preceded by a 180° RF pulse. This preparatory pulse inverts the longitudinal magnetization, i.e. it turns it to its negative value. Tissues recover their longitudinal magnetization at different longitudinal relaxation rates, which are characterized by their T_1 relaxation times.
The 90° readout pulse of the spin echo is applied exactly when the longitudinal magnetization reaches zero for the tissue to be eliminated. The time between the preparatory 180° pulse and the 90° readout pulse is called Time to Inversion (TI).

By selecting the adequate TI, the oppression of different tissues is possible: Short Tau Inversion Recovery (STIR) eliminates the signal from fatty tissue, Fluid Attenuated Inversion Recovery (FLAIR) or Double Inversion Recovery (DIR) suppress signal from fluids.

2.3 Continuous Convex Optimization

In this work we mainly face problems for which we have to infer the original conditions on the basis of measurements: If we want to segment a nephroblastoma to calculate its volume, our source of information is imaging data: We have to solve inverse problems. In the following we first define this kind of problems and their characteristics. Then we explain the basic scheme of our problem descriptions and the optimization methodology.

2.3.1 Inverse Problems

There are many approaches to describe inverse problems, but one of the oldest descriptions goes back to Plato [96], written around 380 BC. In the seventh book of "The Republic" the Greek philosopher discusses in a Socratic dialogue our limitation in understanding the world:

Figure 2.10: Allegory of the cave. Image courtesy Mejia-Foster's AP Language and Composition [114].

"Socrates: Behold! human beings living in a underground den, which has a mouth open towards the light and reaching all along the den; here they have been from their childhood, and have their legs and necks chained so that they cannot move, and can only see before them, being prevented by the chains from turning round their heads. Above and behind them a fire is blazing at a distance, and between the fire and the prisoners there is a raised way;and you will see, if you look, a low wall built along the way, like the screen which marionette players have in front of them, over which they show the puppets.

Glaucon: I see.

Socrates: And do you see, I said, men passing along the wall carrying all sorts of vessels, and statues and figures of animals made of wood and stone and various materials, which appear over the wall? Some of them are talking, others silent.

Glaucon: You have shown me a strange image, and they are strange prisoners.

Socrates: Like ourselves, I replied; and they see only their own shadows, or the shadows of one another, which the fire throws on the opposite wall of the cave?"

Plato's problem can be understood as a description of a problem in which the conditions and parameters of the physical systems are unknown [54]. This leads to an inverse problem where the parameter values describing the system are derived from indirect measurements of an object or function; see Fig 2.10. These observations often have errors such as the fire in the cave, which produces a projected, distorted and blurred image. In contrast to the complementary forward problem, the inverse problem usually has no clear solution and is according to Hadamard [70, 71] an ill-posed problem, i.e. a problem that is not well-posed:

Definition 2.1 (Well-posed problem)**.** *A problem is called well-posed if and only if*

1. *there exists a solution to the problem (existence),*
2. *there is at most one solution to the problem (uniqueness),*
3. *and the solution depends continuously on the data (stability).*

In daily life we are often confronted with inverse problems. In a soccer game, for example, a player has to use his visual and auditory observations to assess where the ball is going, how his teammates and opponents are moving, and how to choose the right path. Similar problems can also be found in the field of medical image processing: a MRI is the visual representation of a physical condition (forward problem). However, it is almost impossible to know all the circumstances, e.g. the noise distribution or fluctuations in the magnetic field, that have influenced the result in order to evaluate the corresponding physical situation. In addition, imaging is always digitized and thus also quantized, such that information is lost - the inverse problem cannot be solved unambiguously and is therefore ill-posed.
In other cases there may not be a solution anymore as soon as information is disturbed. Typically, the major problem is stability - the smallest changes are amplified and can lead to massive errors: even if a problem can be described exactly, it does not automatically lead to it being stable.
A common approach to transform an ill-posed problem into a well-posed one is to make certain assumptions about the solution and to use a priori information [169, 174]. Intuitively we impose a certain regularity on the solution, as discussed in the following.

2.3.2 Convex Optimization

Many of the fundamental optimization problems in image processing can be defined as

$$\min_{\boldsymbol{u}\in\mathbb{R}^n} E(\boldsymbol{u}) = \min_{\boldsymbol{u}\in\mathbb{R}^n} \{D(\boldsymbol{u}, \boldsymbol{f}) + \lambda R(\mathcal{L}\boldsymbol{u})\}. \tag{2.20}$$

where $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ is a continuous linear operator and $\mathcal{X}, \mathcal{Y}$ two (finite-dimensional) real vector spaces. The first term $D(\boldsymbol{u}, \boldsymbol{f})$ is a data or fidelity term that represents the similarity and relationship between the original measurements $\boldsymbol{f}$ and the solution $\boldsymbol{u}$. The

second term $R(\boldsymbol{u})$, the regularizer or smoothness term, imposes regularity on the solution. The fixed parameter $\lambda \in \mathbb{R}$ steers the tradeoff between the fidelity to the original measurements and their regularisation.
First order methods to optimize (2.20) are based on the first derivative. Here, the direction of the gradient gives an indication where to find the minimal solution and the magnitude of the gradient indicates the steepness of the local slope. The problem is optimized when the Euler-Lagrange equation is satisfied, i.e. the gradient is zero.
In general, gradient decent methods converge to the globally optimal solution as long as the problem is convex. Unfortunately, these optimization tends to be slow when the gradient magnitude is low, i.e. in flat regions. Besides, they often do not move directly towards the globally optimal solution.

Primal-Dual Algorithms

In recent years, primal-dual methods became popular in the area of image processing to solve optimization problems [35, 64, 132]. One major advantage is their ability to yield highly efficient splitting optimization schemes, solving the original so-called primal problem as well as the usually simpler dual problem in parallel. In this way, they can handle both differentiable and nondifferentiable terms, i.e. by using explicit steps (gradient operators) or implicit steps with a proximal operator.
In this context, Equation 2.20 is typically referred to as the *primal problem*

$$E_{\boldsymbol{x}} = \min_{\boldsymbol{x}\in\mathcal{X}} \{D(\boldsymbol{x}) + R(\mathcal{L}\boldsymbol{x})\}. \tag{2.21}$$

with $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ is a continuous linear operator and $\mathcal{X}, \mathcal{Y}$ two (finite-dimensional) real vector spaces, and D, R are convex and lower-semicontinuous functions; see Fig. 2.1. In the following we attempt to give an intuitive geometrical interpretation of duality and refer the interested reader to [145] for further theoretical evidence.
In addition to the variant already shown in (2.11), there is a second possibility to represent a convex set, namely as a set of half spaces containing the set: A closed convex set $\mathcal{X}$ can be restored with its supporting hyperplanes by taking the intersection of all closed half spaces containing the closed convex set. The set of all supporting hyperplanes on $\mathcal{X}$ is then the dual representation.
Since the epigraph (the set of points above or on the function) of a convex function $\boldsymbol{f}$ is a convex set [145], we can also imagine $\boldsymbol{f}$ as a set of affine functions $\langle \boldsymbol{y}, \boldsymbol{x}\rangle - \alpha$ that fulfill the following constraint:

$$f(\boldsymbol{x}) \geq \langle \boldsymbol{y}, \boldsymbol{x}\rangle - \alpha \quad \forall \boldsymbol{x} \in \mathcal{X} \tag{2.22}$$

$$\iff \alpha \geq \sup_{\boldsymbol{x}\in\mathcal{X}} \langle \boldsymbol{y}, \boldsymbol{x}\rangle - f(\boldsymbol{x}). \tag{2.23}$$

For a given slope $\boldsymbol{y} \in \mathcal{V}$ the convex conjugate (see (2.15))

$$f^*(\boldsymbol{x}) = \sup_{\boldsymbol{x}\in\mathcal{X}} \langle \boldsymbol{y}, \boldsymbol{x}\rangle - f(\boldsymbol{x}) \tag{2.24}$$

is an optimal choice for α. The double convex conjugation is the convex envelope and $\boldsymbol{f}^{**} = \boldsymbol{f}$, when $\boldsymbol{f}$ is convex; see (2.15). Hence, the dual problem is simply to optimize $\boldsymbol{f}^{**}$ and the corresponding *dual problem* to (2.21) is then [145]

$$E_{\boldsymbol{y}} = \min_{\boldsymbol{y}\in\mathcal{Y}} \{D^*(-\mathcal{L}^*\boldsymbol{y}) + R^*(\boldsymbol{y})\}. \tag{2.25}$$

Algorithm 1 General algorithm to solve the convex saddle-point problem

Initialization: $\tau, \sigma > 0$, $\theta \in [0, 1]$, $\boldsymbol{x}^0 \in \mathcal{X}$, $\boldsymbol{y}^0 \in \mathcal{Y}$, and $\bar{\boldsymbol{x}}^0 = \boldsymbol{x}$
for $(n = 0;\ \ n < N;\ \ n{+}{+})$ **do**
$\boldsymbol{y}^{n+1} = (\mathcal{I} + \sigma \delta R^*)^{-1}(\boldsymbol{y}^n + \sigma \mathcal{L} \bar{\boldsymbol{x}}^n)$
$\boldsymbol{x}^{n+1} = (\mathcal{I} + \tau \delta D)^{-1}(\boldsymbol{x}^n - \tau \mathcal{L}^* \bar{\boldsymbol{y}}^{n+1})$
$\bar{\boldsymbol{x}}^{n+1} = \boldsymbol{x}^{n+1} + \theta(\boldsymbol{x}^{n+1} - \boldsymbol{x}^n)$

Obviously, the difference between primal and dual optimization, the primal-dual gap, is an inherent meaningful convergence measure: in case of $E_{\boldsymbol{x}} = E_{\boldsymbol{y}}$, the optimization converged to the global optimum [35, 145]. Please note, that it does not decrease continuously.
However, we can further reformulate (2.21) with (2.15) to

$$\min_{\boldsymbol{u} \in \mathcal{X}} \max_{\boldsymbol{y} \in \mathcal{Y}} \ \{D(\boldsymbol{x}) + \langle \mathcal{L}\boldsymbol{x}, \boldsymbol{y} \rangle - R^*(\boldsymbol{y})\} \tag{2.26}$$

where $\mathcal{X}, \mathcal{Y}$ are two (finite-dimensional) real vector spaces, $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ is a continuous linear operator, and $D : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, $R^* : \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ are two proper, convex, and lower semi-continuous functions. This is a so-called non-smooth convex saddle point problem. Chambolle and Pock [35] proposed a fast first order primal-dual algorithm for these kind of problems. Their convex relaxation framework is used for example in [127, 128, 130, 175, 192] for the task of interactive segmentation. In [165] it has been extended to a non-metric prior, in [164] to generalized ordering constraints, which constraints labels appearing adjacent to each other in a certain direction, in [50] to RGB-D data, and in [20] to the context of semantic segmentation. As we also make heavy use of their flexible formulation within our work, we explain it in detail in the following.
Algorithm 1 states their general method to optimize the saddle point problem in (2.26). It defines an iterative approach where the dual variable $\boldsymbol{y}$ is updated with a gradient ascend step and a resolvent operator. Concurrently, the primal variable $\boldsymbol{x}$ changes based on a gradient decent step (similar to the classical approach of gradient decent optimization), also in combination with a resolvent operator.

Algorithm 2 Accelerated algorithm to solve the convex saddle-point problem

Initialization: $\tau_0, \sigma_0 > 0$, $\tau_0 \sigma_0 L^2 \leq 1$, $\boldsymbol{x}^0 \in \mathcal{X}$, $\boldsymbol{y}^0 \in \mathcal{Y}$, and $\bar{\boldsymbol{x}}^0 = \boldsymbol{x}$
for $(n = 0;\ \ n < N;\ \ n{+}{+})$ **do**
$\boldsymbol{y}^{n+1} = (\mathcal{I} + \sigma_n \delta R^*)^{-1}(\boldsymbol{y}^n + \sigma_n \mathcal{L} \bar{\boldsymbol{x}}^n)$
$\boldsymbol{x}^{n+1} = (\mathcal{I} + \tau_n \delta D)^{-1}(\boldsymbol{x}^n - \tau_n \mathcal{L}^* \bar{\boldsymbol{y}}^{n+1})$
$\bar{\boldsymbol{x}}^{n+1} = \boldsymbol{x}^{n+1} + \theta_n(\boldsymbol{x}^{n+1} - \boldsymbol{x}^n)$
$\theta_n = \frac{1}{\sqrt{1+2\gamma\tau_n}}$, $\tau_{n+1} = \theta_n \tau_n$, $\sigma_{n+1} = \frac{\sigma_n}{\theta_n}$

Additionally, the extrapolation $\bar{\boldsymbol{x}}$ of the primal variable $\boldsymbol{x}$ allows faster convergence, i.e. $O(\frac{1}{N})$ [35, 126]. Please note that without this extrapolation, i.e. $\theta = 0$, Alg. 1 corresponds to the semi-implicit Arrow-Hurwicz method [8] and has a convergence rate of $O(\frac{1}{\sqrt{N}})$ [35].

The resolvent operators are defined as

$$\boldsymbol{x} = (\mathcal{I} + \tau\delta D)^{-1}(\tilde{\boldsymbol{x}}) = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ \frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2}{2\tau} + D(\boldsymbol{x}) \right\} \tag{2.27}$$

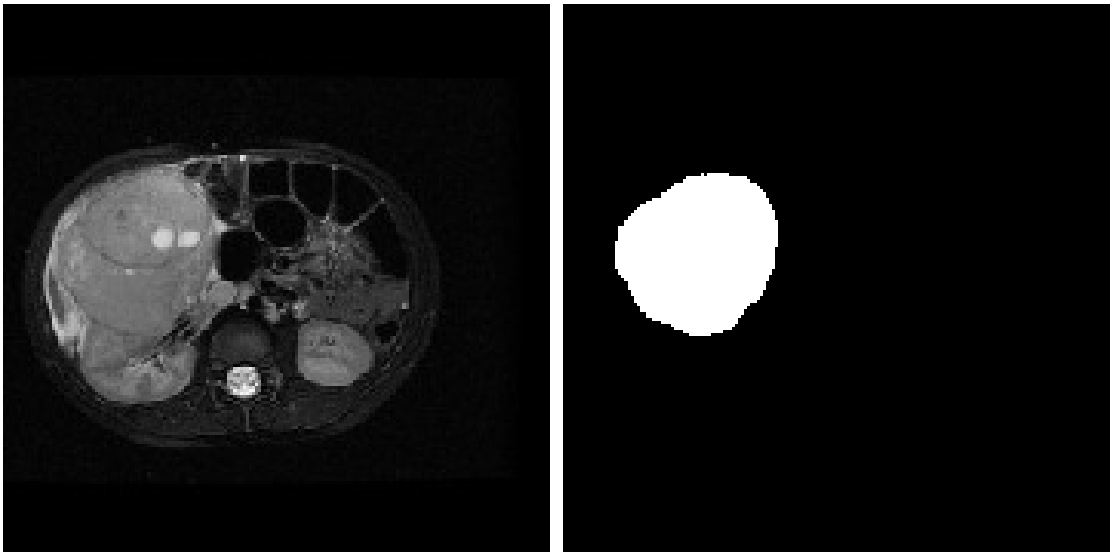
$$\boldsymbol{y} = (\mathcal{I} + \sigma\delta R^*)^{-1}(\tilde{\boldsymbol{y}}) = \underset{\boldsymbol{y}}{\operatorname{argmin}} \left\{ \frac{\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|^2}{2\sigma} + R^*(\boldsymbol{y}) \right\} \tag{2.28}$$

where Chambolle and Pock assume that closed-form representations exist or can be approximated efficiently [35].
However, a large group of functions is not only convex, but uniformly convex (has a Lipschitz continuous gradient). If D or R fulfill this property, it is possible to use the accelerated approach of Chambolle and Pock with a convergence rate of $O(\frac{1}{N^2})$; see Alg. 2. The main difference to their general approach is a modification of $\theta, \tau,$ and σ depending on the number of iterations [35, 174]. Fortunately, all energy formulations we use have a uniformly convex data term and we can apply the second algorithm.

2.4 Error Measures

Error measurements are indispensable to evaluate the quality of the solution of a method. In this work we consider two different kinds of problems, i.e. segmentation and classification.

The result of a segmentation algorithm is the partitioning of an input image into clusters or segments, representing the different classes, see Fig. 2.11. Consequently, it is intuitive



(a) (b)

Figure 2.11: Exemplary binary partitioning. (a) Input image, (b) partitioning into 2 clusters (tumor and non-tumor).

to evaluate segmentation results using two different questions [127, 128, 131, 153]:

1. How much of the labeled cluster actually matches the ground truth for this class, so how *precise* is the selection?
2. How many of the voxels belonging to this class were hit by the segmentation, i.e. how is the *recall* factor?

Therefore, these are the most common error measures. We define *Precision* as

$$P_{\hat{\Omega}_i,\Omega_i} := \frac{|\hat{\Omega}_i \cap \Omega_i|}{|\Omega_i|}\,, \tag{2.29}$$

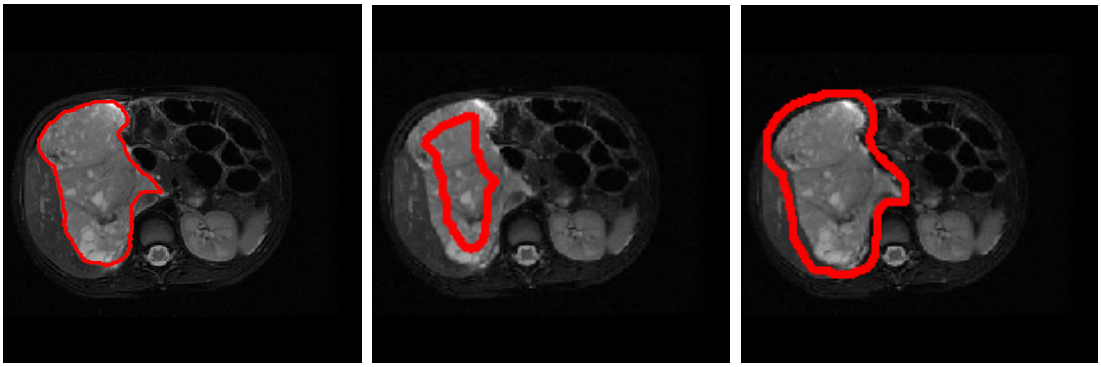
where $\hat{\Omega}_i$ is the ground truth, and Ω_i the algorithmic prediction for label i. Precision or positive predictive value, describes the fraction of the correct selections within all selected voxels. In addition, *Recall* (or true positive rate)

$$R_{\hat{\Omega}_i,\Omega_i} := \frac{|\hat{\Omega}_i \cap \Omega_i|}{|\hat{\Omega}_i|}\,, \tag{2.30}$$

describes the fractional number of all relevant instances selected over the total number of relevant instances.
However, these measures are only informative if both are given. For example, it is possible to get a precision $P_{\hat{\Omega}_i,\Omega_i} = 1$ if only a single randomly correct pixel is selected. At the same time recall is always perfect, i.e. $R_{\hat{\Omega}_i,\Omega_i} = 1$ if all voxels of an image are selected.
For this reason it is common practice to calculate the *Dice score*, defined as the harmonic mean between precision and recall

$$F_{\hat{\Omega}_i,\Omega_i} = \frac{2P_{\hat{\Omega}_i,\Omega_i}R_{\hat{\Omega}_i,\Omega_i}}{P_{\hat{\Omega}_i,\Omega_i} + R_{\hat{\Omega}_i,\Omega_i}}. \tag{2.31}$$

It relativizes the area of a cluster to its overlap with the ground truth. Finally, the average Dice score over an entire data set determines the overall segmentation accuracy. To give an intuition for these measures, we show in Fig. 2.12 some example evaluations. If we compare Fig. 2.12b with Fig. 2.12a, all selected areas are clearly correctly classified



consensus truth $P = 1,\ R = 0.45,\ F = 0.62$ $P = 0.69,\ R = 1,\ F = 0.82$

Figure 2.12: Exemplary image with different annotations, outlined in red.

as tumor, i.e. $P = 1$. However, since not the entire region is hit, the recall $R = 0.45$ is low. In Fig. 2.12c it is inversely: The complete region is marked ($R = 1$), but not all pixels are correct, i.e. $P = 0.69$.
Some of the chapters deal with binary classification. As it is very similar to binary segmentation, the same error measures can be used. The only other error measure we consider is *Accuracy*

$$A_{\hat{\Omega}_i,\Omega_i} := \frac{|\hat{\Omega}_i \cap \Omega_i|}{2}. \tag{2.32}$$

This concludes the mathematical introduction. In the following we now describe the Wilms' tumor including its origin, appearance and treatment.

2.5 Wilms' Tumor

Wilms' tumor, named after the German surgeon Max Wilms, or nephroblastoma is the most common malignant renal[1] tumor in childhood [137]. About 75% of all patients are younger than five years - with a peak between two and three years [47,91]. Although the first descriptions have been attributed to Thomas F. Rance in 1814 [142], the first known specimen was collected by the British surgeon John Hunter between 1763 and 1793 [16]. Its frequency varies between races, i.e. it is more frequent in African than in Caucasians but rarest in East Asian populations.

2.5.1 Associated Syndroms and Precursor Lesions

In 10%-15% of patients with Wilms' tumor, the cause is contemplated to be an epigenetic[2] alteration during embryogenesis[3] or a germline[4] pathogenic variant[5]. These may be correlated with known congenital malformation syndromes [157]. The most common ones are listed in Tab. 2.3.

Table 2.3: Most common syndromes and anomalies associated with Wilms' tumor.

Stage	Description
WAGR	**W**ilms' tumor, **A**niridia (lack of the iris of the eyes), **G**enitourinary tract abnormalities, and mental **R**etardation. Patients show a deletion in region 11p13 of chromosome 11, including the Wilms' tumor gene 1 (WT1) [31]. The incidence of a bilateral (both kidneys are affected) Wilms' tumor is about 15% [28].
Beckwith-Wiedemann	This syndrome is caused by a malformation in region 11p15 of chromosome 11, involving the Wilms' tumor gene 2 (WT2). Approximately 10% of patients develop a Wilms' tumor [151].
Denys-Drash	This syndrome is caused by germline missense mutations in WT1 and patients have a risk of Wilms' tumor about 90% [138].
Hemihypertrophy	This anomaly where one side of the body is (partly) larger than the other often occurs in conjunction with congenital syndromes including the Beckwith-Wiedemann syndrome [151].
Urogenital deformities	Urogenital abnormalities occur in about 4% of all Wilms' tumors [53].

Wilms' tumor is assumed to originate from anomalies in renal histogenesis (the formation and development of body tissues), correlated with the presence of nephrogenic rests - in

[1] pertaining to the kidney
[2] development of an organism from an undifferentiated cell for formation of organs and parts
[3] phase of prenatal development involved in establishment of the characteristic configuration of the embryonic body
[4] cell line from which egg or sperm cells are derived
[5] alteration in a gene associated with a increased disease risk or an abnormal phenotype

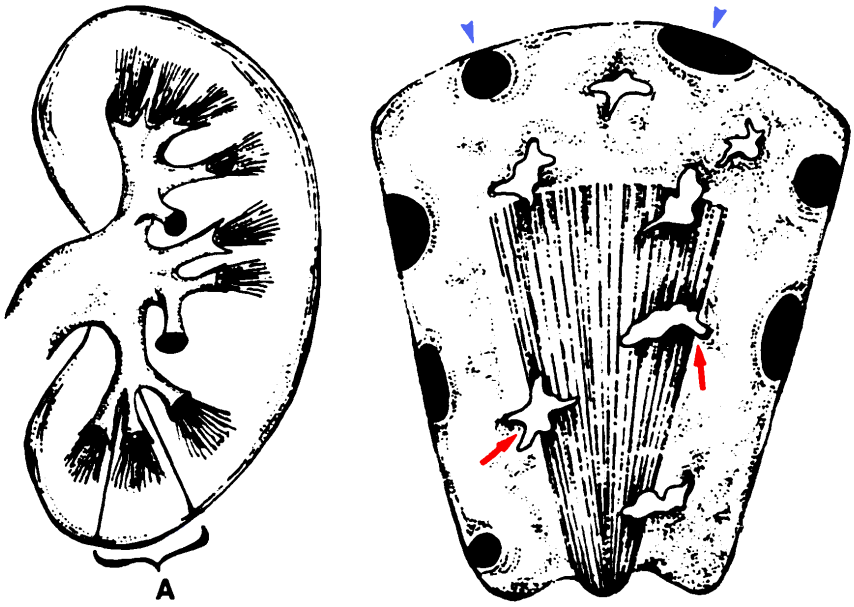


Figure 2.13: Illustration of a renal lobe (**A**) with perilobar nephrogenic rests (blue) and randomly distributed, intralobar nephrogenic rests (red). Adapted from Lonergan et al. [101].

case of an diffuse or multifocal appearance also called nephroblastomatosis. This precursor lesion, a persistent metanephric (embryological structure that give rise to the kidney) tissue, is found in about 40% of unilateral cases, while in patients with bilateral tumors the incidence of nephrogenic rests is 99% [19, 140]. Similarly, these lesions are prevalent in individuals with syndromes associated with Wilms' tumors [17, 18], see Tab. 2.4.
There are two main categories of nephrogenic rests - perilobar (PLNR) and intralobar (ILNR), distinguished by their position within the renal lob [17]. The former is located at the periphery of the renal lobules and the latter in the central part of the lobe, see Fig. 2.13. ILNR is believed to arise earlier in the development when compared with PLNR, which may explain the higher frequency of heterologous elements in ILNR, such as striated muscle, fat, cartilage and bone. Depending on the stage of their development, both ILNR and PLNR might present with different morphological patterns.
The type and percentage of nephrogenic rests vary in patients with unilateral or bilateral disease. Patients with bilateral Wilms' tumor have a higher proportion of perilobar rests (52%) than of intralobar or combined rests (32%) and higher relative proportions of rests, compared with patients with unilateral tumors (18% perilobar and 20% intralobar or both) [94].

2.5.2 Clinical Presentation

Typically, nephroblastoma presents clinically as an increase in the abdominal girth - rarely in combination with abdominal pain, loss of appetite or a feeling of weakness. This initial suspicion is then finally confirmed by the histology of the extracted prepara-

[6] consecutive development of tumors

Table 2.4: Correlation between nephrogenic rests and accociated syndroms. Adapted from Beckwith et al. [17]. HH: Hemihypertrophy.

Population	PLNR (%)	ILNR (%)
Infant autopsies	1	0.01
Unilateral Wilms' tumor	25	15
Bilateral Wilms' tumor (synchronous)	74 – 79	34 – 41
Bilateral Wilms' tumor (metachronous[6])	42	63 – 75
Beckwith-Wiedemann/HH and Wilms' tumor	70 – 77	47 – 57
Aniridia and Wilms' tumor	12 – 20	84 – 100
Denys-Drash and Wilms' tumor	11	78

tion. Nevertheless, imaging techniques can support the diagnosis. Often an abdominal sonography is performed at the beginning because it is neither cost-intensive nor time-consuming. Although this method is non-invasive and radiation-free, the quality of the images is often insufficient and strongly dependent on the operator. Therefore, with the help of a CT scanner or a MRI, higher-quality cross-sectional images of the patient are produced in the further course of the diagnostic. MRI images are generally preferred in everyday clinical practice because they exhibit a high soft tissue contrast and do not cause radiation exposure. As these images generally take a relatively long time, children are usually sedated to avoid movement artifacts.
In order to rule out pulmonary metastases, an additional x-ray of the lungs is usually taken or, if necessary, a computer tomography. It is problematic to base the diagnosis of a nephroblastoma exclusively on imaging and clinical appearance: In the case of benign tumors such as cystic nephroma or low malignant tumors such as nephroblastomatosis, there is a risk of excessive or even unnecessary chemotherapy.
Due to this uncertainty, information on previous illnesses, see Sec. 2.5.1, and laboratory findings also play an important role in diagnostics. This additional information allows for a better differential diagnosis compared to other malignant and benign tumors, as well as conclusions about the course of chemotherapy and whether the kidney function is impaired. Especially the neuroblastom is an important differential diagnosis to the nephroblastoma - even if it does not originate directly from the kidney, the origin is usually very close and the age of onset and visual appearance are nearly identical. One of the clearest indications of this is an increase of catecholamines in the urine. There are more malignant neoplasias, e.g. renal cell carcinoma or rhabdomyosarcoma, but since they are no part of this work we refer to the corresponding literature [107].

2.5.3 Treatment

Patient's treatment is established upon evidence obtained from three cooperating organizations, i.e. the International Society of Pediatric Oncology (SIOP), the National Tumor Study Group, and the United Kingdom Children's Cancer Study Group. In Europe, diagnosis and therapy follow the guideline of SIOP [67,88], promoting the use of a preoperative chemotherapy to downstage the tumor [183]. After diagnosis of a nephroblastoma, the

Table 2.5: Staging system of the Societe Internationale d'Oncologie Pediatrique (SIOP) based on findings after preoperative chemotherapy [55].

Stage	Description
I	• Tumor is limited to kidney or surrounded with a fibrous pseudocapsule and infiltration does not reach the perirenal tissue. The tumor is completely resected and excision marginsclear of tumor. • Tumor may be protruding into the pelvic system and dipping into the ureter, but is not infiltrating the walls. • Vessels of renal sinus are clear of tumor. Intrarenal vessels may be involved. • Fine-needle aspiration is allowed.
II	• Tumor extension beyond kidney or renal pseudocapsule but is completely resected with clear resection margins. • Infiltration of renal sinus or vessels outside the renal parenchyma (essential or functional elements of the organ) but is completely resected with clear resection margins. • Tumor infiltrates adjacent organs or vena cava (one of two major veins of the blood circulatory system) but is completely resected.
III	• Incomplete excision of the tumor which extends beyond resection margins • Invasion of abdominal lymph nodes • Preoperative or intraoperative tumor rupture • Tumor implants found on peritoneal surface • Tumor thrombi present at resection margins of vessels or ureter • Open biopsy prior to start of treatment
IV	Haematogenous metastases or lymph node metastases outside the abdominopelvic region.
V	Bilateral renal tumors at the time of diagnosis.

typical course of therapy based on the recommendation of the SIOP studies is as follows:

1. Chemotherapy to reduce risk of tumor-rupture during surgery
2. Resection of the tumor
3. Possible follow-up treatment (chemotherapy/irradiation), see Tab. 2.7

One of the main risks during surgery is a tumor rupture and the resulting possible formation of metastases - for the same reasons, open biopsy is typically not performed. The

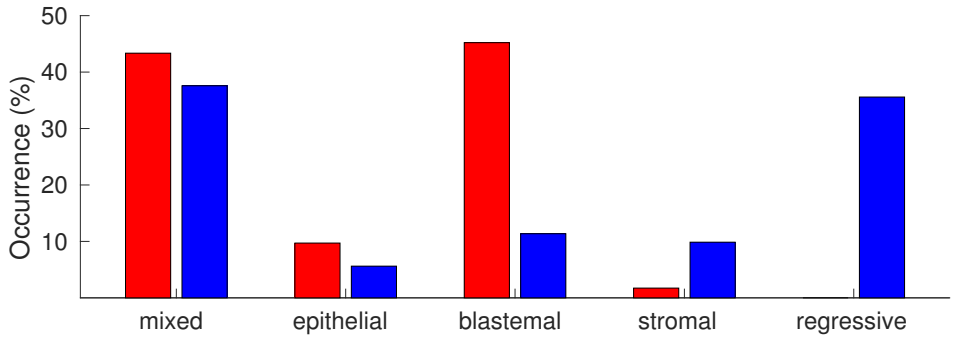


Figure 2.14: Subtype distribution of the most common ones without (red) and with (blue) pre-operative chemotherapy. These distributions indicate a change in tumor structure during chemotherapy.

chemotherapy reduces this risk and has positive effects on metastasis formation and tumor growth - usually the tumor shrinks during the administration of chemotherapeutics. Hence, in a first step, the patient is classified based on the criteria in Tab. 2.5 and Tab. 2.6. This classification determines post-operative treatment, see Tab. 2.7: The chemotherapies differ in duration and drug administration: AV (before surgery) and AV2 (after surgery) vary in duration, but both contain only the drugs actinomycin-D and vincristin. AVD, on the other hand, contains another chemotherapeutic agent, doxorubicin. Finally, there is a high-risk chemotherapy, which consists of four drugs (carboplatin, etoposid, doxorubicin, cyclophosphamid) and is highly toxic. Possible side effects - just to mention a few of them - reversible bone marrow aplasia that can cause anemia, leukopenia and thrombocytopenia. Also the cells of the intestinal epithelium divide very fast and can be affected leading to diseases like diarrhoea up to a paralytic ileus. The liver, as a detoxifying organ, can be harmed, so monitoring liver enzymes is important. There are several other side effects caused by chemotherapeutica. Doxorubicin in particular has a cardiotoxic effect, which can lead to various secondary diseases such as cardiomyopathy with heart insufficiency even years later. Vincristin, on the other hand, has a major effect on the nervous system, which can manifest itself as muscle weakness or nerve paralysis. Also the other chemotherapeutics have various side effects - we refer therefore to [134] for further details.
Obviously, a chemotherapy is a great physical and psychological burden for the patient, and it is essential to select the appropriate therapy and duration. Nonetheless, it has also an influence on the composition of tumor tissue, see Fig 2.14. Wilms' tumor generally consists of three types of tissue [182]: blastemal, epithelium and stromal, see Fig. 2.15. If

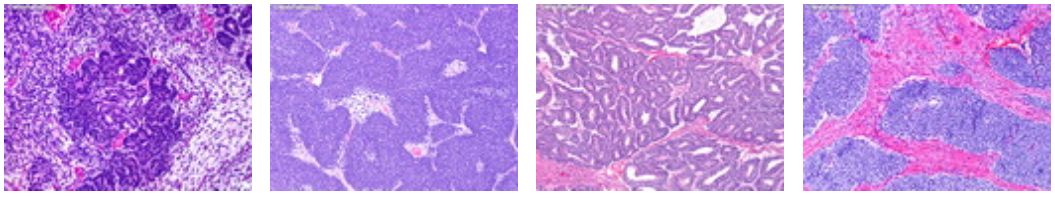


Figure 2.15: Histological patterns of Wilms' tumors: (a) Triphasic, (b) blastemal pattern, (c) epithelial pattern, (d) stromal pattern. Image courtesy of WebPathology [188].

all are present, one refers to a triphasic, otherwise to a bi- or monophasic tumor. During chemotherapy, the composition might change and the proportion and degree of differentiation can vary considerably from patient to patient. Prior to chemotherapy, patients with blastemal dominant tumors are very common, whereas the number of the regressive subtype increases significantly after this therapy step.

All in all, this leads to a number of histological manifestations with different outcomes, see Tab. 2.6:

- *cystic partially differentiated nephroblastoma* (CPDN) : Tumor composed entirely of cysts, lined with epithelium, and thin septa, the only solid parts of the tumor. The tumor forms a well demarcated tumor mass that stands out from the non-cystically transformed adjacent kidney tissue. In the septa small blastemal islands, partly mixed by stromal cells and epithelial structures, are detectable.
- *completely necrotic nephroblastoma* : No vital tumor tissue is detectable.
- *regressive type*: Classical changes occur after preoperative chemotherapy, with more than two-thirds of the tumor have to be regressed by definition. The remaining vital tumor elements may contain the three different differentiation forms of nephroblastoma, such as blastemal, epithelial and stromal cells.
- *mixed type*: The regressive changes in the tumor account for less than two thirds. The vital tumor tissue consists of blastemal and/or epithelial and/or stromal tumor cells. None of these components make up more than two thirds of the vital tumor.
- *epithelial dominant type*: The regressive changes in the tumor account for less than two thirds. At least two thirds of the vital tumor tissue consists of epithelial cells. Small lesions of blastemal cells, making up less than 10% of the vital tumor tissue, can occur. Stromal cells can occur to a variable extent.
- *stromal dominant type*: The regressive changes in the tumor account for less than two thirds. At least two thirds of the vital tumor tissue consists of stromal cells. Small lesions of blastemal cells, making up less than 10% of the vital tumor tissue, can occur. Epithelial cells can occur to a variable extent.
- *blastemal dominant type*: The regressive changes in the tumor account for less than two thirds. At least two thirds of the vital tumor tissue consists of blastemal cells. Other cell types can occur to a variable extent.
- *focal anaplastic tumor*: Detection of enlarged, atypical tri- or multipolar mitoses. Pronounced nucleus enlargement, at least three times larger than the surrounding tumor cell nuclei. Pronounced hyperchromasia of the cell nuclei. The tumor contains either one, or at most two sharply defined tumor nuclei that contain cells with these criteria. The surrounding tumor tissue does not show these changes. In case of possible extrarenal tumor expansion, no anaplastic cells may be detected there.
- *diffuse anaplastic tumor*: Detection of enlarged, atypical tri- or multipolar mitoses. Pronounced nucleus enlargement, at least three times larger than the surrounding tumor cell nuclei. Pronounced hyperchromasia of the cell nuclei. The above mentioned anaplastic changes are present at different sites of the tumor or outside the

Table 2.6: Current SIOP classification of Wilms' tumors.

Risk category	Pre-treated tumors	Primary nephrectomy tumors
Low	Cystic partially differentiated Completely necrotic	Cystic partially differentiated –
Intermediate	epithelial dominant stromal dominant mixed regressive focal anaplasia	Non-anaplastic and its variants – – – focal anaplasia
High	blastemal dominant diffuse anaplasia	– diffuse anaplasia

tumor capsule. Anaplastic cells are found in intrarenal or extrarenal vessels, the renal sinus, or in metastases. Although the actual criteria for anaplasia are only focally realized (one to two foci), the surrounding tumor tissue exhibits pronounced mitotic activities, strong nuclear enlargement and polymorphism (so-called "nuclear unrest"). Although there is only one anaplasia focus, it is not sharply separated from the surrounding tumor tissue.

After surgery, the patient receives further chemotherapy or radiation of the preoperatively affected tissue area, depending on the local stage (sometimes with up-staging due to the outcome of the surgery), histological subtype and tumor volume. Patients with a tumor volume above 500 ml are treated more intensively, see. Tab. 2.7. In case of a bilateral Wilms' tumor, this decision is always based on the worse stage and histological subtype. Obviously, tumor volume is an important characteristic, that should be determined exactly.

Table 2.7: Post-operative treatment. Chemotherapeutic Agents: A, actinomycin D; D, doxorubicin; HR-1: etoposide, carboplatin, cyclophosphamide and doxorubicin; V, vincristine [177].

Disease	Tumour volume after preoperative chemotherapy	Treatment		
		Stage I	Stage II	Stage III
Low-risk	All	None	AV (27 weeks)	AV (27 weeks)
Intermediate-risk, all subtypes	$< 500ml$	AV (4 weeks)	AV (27 weeks)	AV (27 weeks) + flank radiotherapy
Intermediate-risk, stromal or epithelial-type	$\geq 500ml$	AV (4 weeks)	AV (27 weeks)	AV (27 weeks) + flank radiotherapy
Intermediate-risk, nonstromal, nonepithelial	$\geq 500ml$	AV (4 weeks)	AVD (27 weeks)	AVD (27 weeks) + flank radiotherapy
High-risk, blastemal type	All	AVD (27 weeks)	HR-1 (34 weeks)	HR-1 (34 weeks) + flank radiotherapy
High-risk, diffuse anaplasia	All	AVD (27 weeks)	HR-1 (34 weeks) + flank radiotherapy	HR-1 (34 weeks) + flank radiotherapy

2.5.4 Summary

In this section, we illustrated Wilms' tumors, their origin and development from the benign nephroblastomatosis, their appearance, as well as their therapy protocol. Obviously, treatment is mainly dependent on the subtype, local stage and tumor volume, respectively. Thus, we investigate these aspects further in the remaining part of this work.

3 Related Work

"A people without the knowledge of their past history, origin and culture is like a tree without roots."

– Marcus Garvey

Contents

In this chapter we describe previous publications that have influenced our work. Since segmentation is one of the most fundamental cornerstones of computer vision, the first section deals with the history of segmentation and current work in image processing and medical image analysis in particular. In this way, we also provide an overview of the research area and provide the context of our work. Although most methods were initially designed for two dimensional segmentation, the extension to the three dimensional case is mostly straight forward.

Section 3.2 introduces classification and explains its history from its beginnings to the present day. As before, we show the literature that is close to our and give an overview of different approaches.

3.1 Segmentation

Image segmentation is one of the most fundamental problems in the field of image processing and has been one of the most important research areas for decades. The basic idea is to partition an image into meaningful segments based on some prior knowledge. Typically, segmentation is one of the first steps of image analysis and its results are processed further to detect or track objects, or classify images according to their content.
In the following we list a wide range of methods: from the simplest approach of thresholding and region merging, to super-pixel approaches, up to machine learning and the area of deep learning, to continuous formulations with an energy functional. Especially energy formulations form the foundations for our own segmentation methods that we present in chapters 6 and 7.

Probably the simplest and most intuitive segmentation approach is thresholding. These approaches divide the gray values of the image directly into classes, distinguishing between global and local thresholding. While the first applies a single threshold for the entire image, the latter uses several values depending on the local environment. Otsu's thresholding [133], which maximize the variance between the classes, is probably the most famous method. However, there are many other approaches - as this huge topic is beyond our scope, we refer the interested reader to the broad introduction to classic concepts of Rogowska [147]. The main advantage of thresholding methods is their simple calculation and the low computational costs. Unfortunately, spatial relationships are neglected and they only work as long as the classes differ clearly in their gray values.
Another intuitive approach is region merging [58, 202]. Here we assume that neighboring pixels with similar characteristics can be merged to a new, larger region. Thus, spatial information is taken into account and the segmentation is guaranteed to have close contours. The disandvantages are unfortunately severe: The computational cost is large [5], and the size of objects is ignored, such that noise can lead to over-segmentations and objects might merge with the background if the local contrast is too low.
Obviously, a good segmentation result is only reachable, when sufficient prior knowledge is included in the partitioning. A straight forward idea is to group pixels in a first step on the basis of homogeneous areas and uniform texture, which drastically reduces the calculation costs. At the same time, these groups, or super-pixels, form a good starting point for feature extraction [7].
In the last years, many methods to generate such super-pixels were published [1, 99, 167]. One of the most intuitive methods is probably the watershed transformation [181]. The basic idea is to interpret the gray values of the image as heights in a relief. Starting from the valleys of this relief, the super-pixels are then defined by their basins. Various extensions of this approach have been published, ranging from marker-based [21] to weighting strategies of the generated boundaries [6] to entropy based super-pixel generation [99] up to normalized cuts algorithms [189]. Unfortunately, it is very difficult to determine an appropriate value for the number of super-pixels, the most important parameter for these kind of methods.
A completely different way to include prior knowledge in the segmentation method is to infer representations from given data. This broad class of machine learning approaches assumes that there is a sufficient amount of data available to learn the distribution of the different classes. In this way, it is possible to include massive prior knowledge in the segmentation procedure. The simplest methods are based on distances to decision

planes [24, 27] or use a clustering of areas in a high-dimensional feature space [100]. The more advanced variants of these approaches, so-called deep learning methods, not only extract the class distributions from the data, but also learn the associated features independently [74, 102, 149]. Since these methods are very powerful when a sufficient amount of data is available, they dominate the field of image segmentation. Unfortunately, their strength is also their biggest weakness: These approaches are able to segment a learned distribution exactly - but only the distribution present in the training data. Hence, the basic assumption that the existing distribution in the training data does not deviate too much from the image to be segmented is their Achilles' heel. Even small deviations, such as a different noise distribution, lead to their breakdown [62]; see Chapter 7. Another massive problem is the amount of necessary training data. Especially for rare diseases, like Wilms' tumors, it is almost impossible to train one of these approaches properly.
Hence it is essential to find a compromise, such that on the one hand previous knowledge can be integrated, and on the other hand the method is still robust against deviations in the data or disturbances like noise.
The wide range of variational approaches allows this balancing. As a rule, all these methods can be described in the same way: A data term that models the similarity between the solution of the segmentation and the assumed properties and a regularizing term allowing a deviation from these and ensuring smoothness; see Sec. 2.3.2. This balancing paves the way for this type of method to be very robust to noise and other disturbances. A major advantage is their ability to directly model prior knowledge such as shape constraints [29], relationships within regions [38, 127], texture [128], or even information about the ordering of objects [44].
One of the most important and powerful approaches in this class of segmentation methods is the Mumford-Shah functional [123]. In the following, we briefly sketch this famous energy formulation as well as its most common simplification. Afterwards, we explain the approach of Nieuwenhuis and Cremers [127] that forms the basis for our segmentation method for Wilms' tumors.

3.1.1 The Mumford-Shah Functional

In their proposed energy formulation for image segmentation and denoising, Mumford and Shah assumed that the image domain $\Omega \subset \mathbb{R}^n$ can be split into non-overlapping segments separated by a contour C [123]. Their most general formulation

$$E(\boldsymbol{u}, C) = \int_{\Omega} \|\boldsymbol{u} - \boldsymbol{f}\|^2 d\boldsymbol{x} + \mu \int_{\Omega \setminus C} \|\nabla \boldsymbol{u}\|^2 d\boldsymbol{x} + \nu\, \ell(C). \tag{3.1}$$

is minimal when the solution varies smoothly within the segments and has strong variations between them. Here, the unknown function $\boldsymbol{u}$ denotes a piecewise smooth approximation of $\boldsymbol{f} : \Omega \to \mathbb{R}^m$ and the segment boundaries C have a (Hausdorff) length of $\ell(C)$. The first term is a data term (see Sec. 2.3.2) penalizing differences between the segments and the original image, the second integral ensures smoothness within the segments and the last term favors short segment boundaries. The parameters $\mu, \nu > 0$ allow the weighting of their corresponding integrals with respect to inhomogeneities within each segment. Obviously, the choice of μ and ν is of crucial importance: The higher the value for ν, the less segments are contained in the final result. Likewise, if the value for μ increases or in the extreme case reaches infinity, the smoother the segments become until they are

constant at some point. This special case of (3.1) is referred to as the Mumford-Shah cartoon model and defined as

$$\begin{aligned} E(\boldsymbol{u}, C) &= \sum_i \int_{\Omega_i} \|\boldsymbol{u} - \boldsymbol{f}\|^2 d\boldsymbol{x} + \nu\, \ell(C)\,, \\ \text{s.t.} \quad \Omega &= \bigcup_{i=1}^{n} \Omega_i, \quad \Omega_i \cap \Omega_j = \emptyset, \quad \forall\, i \neq j \end{aligned} \tag{3.2}$$

Here, the a priori unknwn number of segments Ω_i partition the data domain Ω and $\boldsymbol{u}$ is a piecewise constant approximation of $\boldsymbol{f}$. One main advantage of the formulation in (3.2) is the guarantee of closed segment boundaries as the enforced constancy within each segment forbids smooth transitions between regions.

The Mumford Shah functional [123] has been studied extensively for decades. However, it took a very long time for its general problem - the high non-convexity - to find a feasible solution. The probably oldest approach to simplify the problem goes back to Köpfler et al. [92], who suggested an approximation by region merging. Other methods try to directly approximate this non-convex problem [179] or simpler elliptic variational problems [4]. In recent years, however, many researchers have attempted to approach this method using convex relaxation [97,198]. The basic idea is to move the problem into a higher-dimensional space and then make a simple binary decision. Unfortunately, the computation time increases quadratically for each additional color channel, i.e. in our application for each additional MRI sequence. More recently, however, Strekalovskiy and Cremers [166] proposed a primal-dual approach that allows to approximate a high quality solution of the functional in real time. We use this approach in our iterative segmentation method in Chap. 7 to show that good prior knowledge is sufficient to provide competitive results for high-grade brain tumor segmentation even with a classical approach such as the Mumford Shah functional - without training or extensive feature extraction.

3.1.2 Spatially Varying Color Distributions

A different possibility to incorporate prior knowledge in the segmentation procedure is to include user interaction. These semi-automatic or interactive methods start with a seed for the object that is to be segmented in the image. In the context of interactive segmentation, these labels are often called scribbles and are set by a user. However, labels can also be generated automatically by an algorithmic method.
A successful idea is the estimation of statistical features from given scribbles in order to define likelihoods for each label, e.g. mean value [38], color/intensity histograms [25,59,127,150], or texture [6,128,153,154]. As a regularizer usually the boundary length in the image metric is minimized.
A flexible framework for multi-label segmentation is the formulation as a minimal partitioning problem [63,123], which can be solved via convex relaxation methods; see Sec. 2.3.2. This framework, which is sometimes referred to as Potts multi-label segmentation model, is very flexible, since all the above data likelihoods can be incorporated. The boundary length is represented using the total variation of the label indicator functions and is easily adjusted to a modification of the image metric.
Let us consider a minimal partitioning problem of the image domain $\Omega \subset \mathbb{R}^m$ into

$\Omega_1, \dots \Omega_n$ non-overlapping regions [34]. Then the generic variational problem is

$$\begin{aligned} \min_{\Omega_1,\dots,\Omega_n \subset \Omega} \; & \frac{1}{2} \sum_{i=1}^{n} \mathrm{Per}(\Omega_i; \Omega) + \sum_{i=1}^{n} \int_{\Omega_i} h_i(\boldsymbol{x}) \, d\boldsymbol{x} \,, \\ \text{s.t.} \quad & \Omega = \bigcup_{i=1}^{n} \Omega_i, \quad \Omega_i \cap \Omega_j = \emptyset, \quad \forall \; i \neq j \end{aligned} \tag{3.3}$$

where $\mathrm{Per}(\Omega_i; \Omega)$ denotes the perimeter of region Ω_i inside Ω, and $h_i \colon \mathbb{R} \to \mathbb{R}_+$ are potential functions reflecting the cost for each pixel being assigned to a certain label $i = 1, \dots, n$. To align image and region boundaries, the perimeter is commonly measured in a metric induced by the underlying image $\boldsymbol{f} : \Omega \to \mathbb{R}^m$.

Assume the user provides a measurable set of scribbles $S_i \subset \Omega$ for each label i. Nieuwenhuis and Cremers [127] suggest to define the potential function $h_i(\boldsymbol{x})$ in 3.3 as the negative logarithm of the linearly to $[0, 1]$-scaled function $\tilde{h}(\boldsymbol{x})$ of

$$\frac{1}{|\mathcal{S}_i|} \int_{\mathcal{S}_i} k_{\rho_i(\boldsymbol{x})}(\boldsymbol{x} - \boldsymbol{y}) \; k_\sigma(f(\boldsymbol{x}) - f(\boldsymbol{y})) \, d\boldsymbol{y}. \tag{3.4}$$

Here $|\mathcal{S}_i|$ is the area occupied by the ith label, and k_σ and k_{ρ_i} are Gaussians with standard deviation σ in intensity space and adaptive standard deviation $\rho_i(\boldsymbol{x}) = \alpha \inf_{\boldsymbol{y} \in \mathcal{S}_i} |\boldsymbol{x} - \boldsymbol{y}|$ in the spatial domain, respectively. The spatially adaptive standard deviation attenuates the influence of the intensity distribution from scribbles that are far away proportionally to the distance of $\boldsymbol{x}$ to the closest scribble location.
The regions $1 \dots n$, as well as their non-overlapping and covering criterion can be easily represented by label indicator functions, whose ranges get relaxed to $[0, 1]$. In this representation, the perimeter of the regions is measured by the weighted total variation. Using the dual definition of the total variation makes the application of the efficient primal-dual algorithm presented in Sec. 2.3.2 is straightforward [127]; see Alg. 3.

Algorithm 3 Algorithm to solve for spatially varying color distributions (3.3) and (3.4)

Initialization: $\tau, \sigma > 0$,$\boldsymbol{x}^0 \in \mathcal{X}$, $\boldsymbol{y}^0 \in \mathcal{Y}$, and $\bar{\boldsymbol{x}}^0 = \boldsymbol{x}$
for $(t = 0; \;\; t < T; \;\; t{+}{+})$ **do**
$\boldsymbol{y}_i^{t+1} = \Pi_{\mathcal{K}_g}(\boldsymbol{y}_i^t + \sigma \nabla \bar{\boldsymbol{x}}_i^t)$
$\boldsymbol{x}_i^{t+1} = \Pi_B(\boldsymbol{x}_i^t - \tau(\mathrm{div}\; \boldsymbol{y}^{t+1} - f_i))$
$\bar{\boldsymbol{x}}_i^{t+1} = 2\boldsymbol{x}_i^{t+1} - \boldsymbol{x}_i^t$

Here, the resolvent operator with respect to the primal problem Π_B is given by the projection onto the simplex [127]. The dual variables $\boldsymbol{y}_i$ are projected to

$$\mathcal{K}_g = \left\{ \boldsymbol{y}_i \in C_c^1(\Omega, \mathbb{R}^m) | \; |y_i(\boldsymbol{x})| \leq \frac{g(\boldsymbol{x})}{2}, \boldsymbol{x} \in \Omega \right\} \tag{3.5}$$

where C_c^1 is the space of smooth functions with compact support, and $g : \Omega \to \mathbb{R}^+$ is defined as

$$g(\boldsymbol{x}) = \mathrm{e}^{-\gamma |\nabla f(\boldsymbol{x})|} \,. \tag{3.6}$$

The advantages of this approach are manifold. First, the segmentation is quite robust to initializations. Second, in the case of binary segmentation, the global optimal segmentation (w.r.t. the scribbles) is obtained. And last but not least, the algorithm can be efficiently implemented on the GPU, resulting in real time performance.

3.2 Classification

Another essential research area in computer vision is devoted to image classification tasks. The basic problem is to decide for a given input image whether a specific class is displayed. Please note that every segmentation problem is obviously a classification task in pixel space. In general, it can be split into single and multi object class/category classification. Traditionally, these tasks are solved with machine learning algorithms.
Nowadays, most classification approaches heavily rely on deep learning methods with millions of parameters to be tuned during network optimization [144, 196]. Similarly to the segmentation approaches, these methods need a large amount of data for training imposing a major problem in the case of Wilms' tumors and its classification task: It is a rare disease with few data available. Thus, we do not consider deep learning approaches in the following.
However, most shallow machine learning methods are based on two essential cornerstones. On the one hand, they all include a classifier that decides which class is matched and on the other, a feature extraction step is performed. Actually, these two parts are interdependent: Obviously, a perfect classifier would not need expressive features and perfect features would be suitable for all classifiers. Hence, we start with an overview about possible classifiers and continue with approaches to feature extraction. Afterwards we show how to combine these approaches within bag of visual words models.

3.2.1 Classifiers

The k-nearest neighbor classifier is probably the easiest and most intuitive approach. The basic idea is to store all samples of the training set and to classify afterwards a new image depending on its k-nearest neighbors in this sample space [3]. Clearly, the implementation is trivial and no training time is needed. However, classifying a new example enforces a comparison with all samples in the training data and is therefore not applicable.
Adaptive boosting or Adaboost [60] is a iterative learning method to create a better classifier based on an ensemble of weak classifiers. At every iteration, the best performing weak classifier is selected.
An widely used approach are support vector machines (SVM) [24], one of the most known methods for image classification. Basically, it is designed to find the optimal hyperplane in a highdimensional space, separating between sets of objects having different classes [85]. Unfortunately, these approaches fail when classes are imbalanced, i.e. one class has a higher occurrence. Besides, support vector machines are designed for binary classification and do not naturally support multi-class problems.
In contrast, random forests [27] are inherently multi-class approaches. These methods combine a group of decision trees and employ a finite set of different learning algorithms to get better predictive performance than using a single learning algorithm. We make heavy use of random forests in our classification of nephroblastomatosis in Chapter 8. In order to understand our procedures, an in-depth understanding of this technique is essential. Thus, we explain it in the following in detail.

Random Forests

Random forests [27] are ensembles of decision trees. Intuitively, this building block is a hierarchical structure following a simple sequence of yes/no decision about the input data until the leaves of each tree are reached. Let us start with the simple problem in

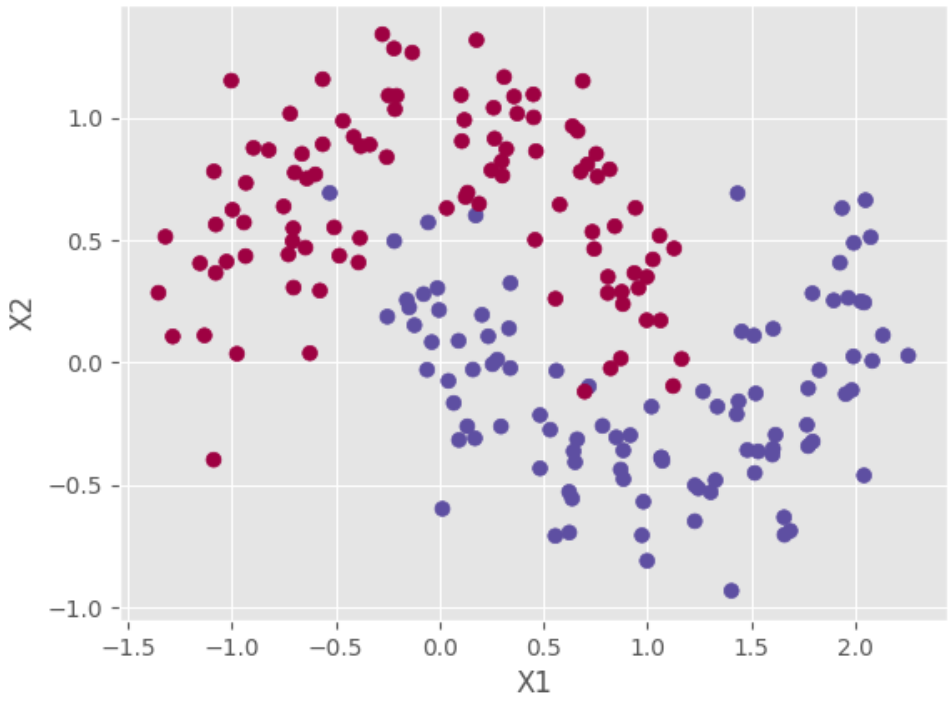


Figure 3.1: Example of linearly non-separable data. No single straight line can separate the classes.

Fig. 3.1 where our input consists of the two variables $X1$ and $X2$. Although the problem is visually simple, it is not linearly separable - no single straight line is able to separate the red and blue classes.
The basic idea of a decision tree is now to build a series of linear boundaries to partition the data into boxes - these nodes then represent a non-linear model. The best intuition in this case is to imagine that at each node a decision is made on which side of the current linear boundary the data point lies based on the given features. Assuming that the tree can have an infinite number of branches (i.e. linear boundaries), a single tree can learn every distribution - leading to a massive overfitting of the training data. For this reason, individual trees are extremely sensitive to noise.
Typically, this problem is addressed by limiting the flexibility of decision trees, i.e. a limited number of linear boundaries. In addition, ensembles of decision trees, namely random forests, are even more robust against disturbances. Training all decision trees of a random forest on the same training data would result in strongly correlated trees. Bagging (bootstrap aggregation) [26] generates new training sets by sampling from the original training set uniformly and with replacement. In this way, decision trees are decorrelated by using different training data. Additionally, random forests use feature bagging, i.e. features are randomly sampled for each decision tree [77].

3.2.2 Feature Extraction

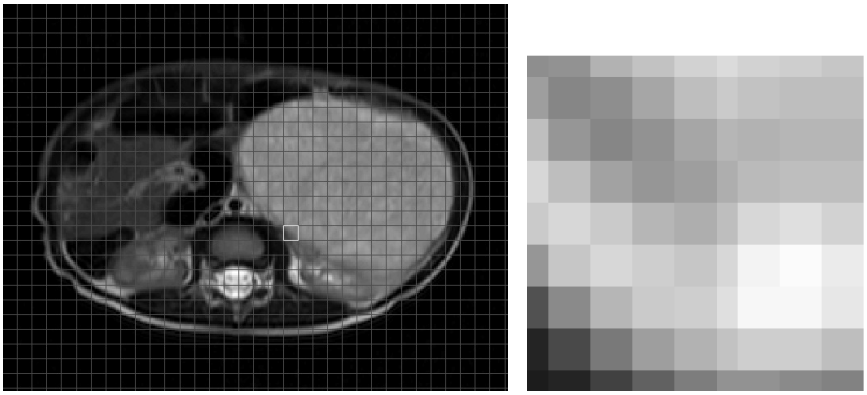
In general, feature extraction is defined as the problem of selecting a subset of a classifiers' input to enforce consideration of only important information. Images often contain not only the object to be classified, but also noise or additional (but unimportant) information. Hence, feature selection tries to find a subset that maximizes the learners classification performance, i.e. some scoring function. Intuitively, this steps performs a dimension reduction and maps input images to the feature space.
Here, two main aspects are important:

- Where to extract the feature?
- Which feature should be extracted?

Feature selection is an important step and there are numerous publications in this field. Probably the easiest approach is to collect features from edge maps [32]. Unfortunately, these information (although very important) are usually not representative enough for classification tasks. Other approaches are based on scale-space pyramids, or orientation assignments. However, probably the most common approaches are based on the SIFT (scale invariant feature transform) operator [104] to extract key points. This operator extracts local extrema over space and scale based on differences of Gaussians, i.e. scale space extrema.
However, after we found the points of interest, it is maybe more important to decide which features to extract. Obviously, color or gray value information is essential, but the range of possible features is huge: from color (or intensity) features [39], to moments [79] texture features [73,156,186], or histograms of oriented gradients [45]. Extracting features on different scales, or computing transformations are also widely used methods.
In general, feature selection and extraction is a challenging and time consuming topic and has to be adopted to each task separately.

3.2.3 Histograms of Oriented Gradients

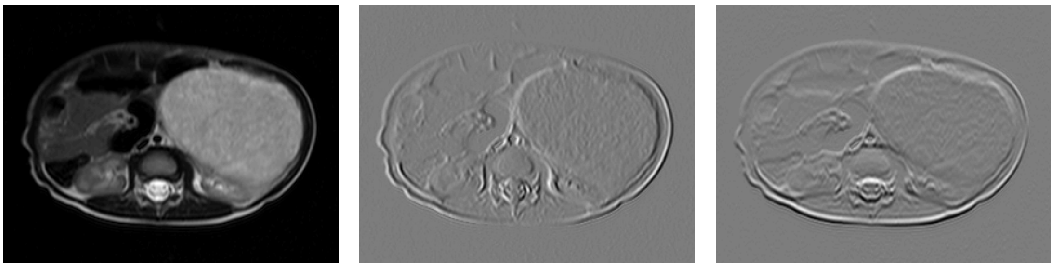
A *histogram of oriented gradients* (HOG) is a special type of feature descriptor [45]. Typically, such a descriptor denotes an useful image or patch representation, that extracts important information and discards the rest. In most cases, the output of such descriptors is vectorized to a vector of length n. In our work, we often make use of image patches, i.e. a splitting of the image into smaller parts, so called *patches*. The patches usually have a square shape of dimension $n_1 \times n_2$, but are not limited to this - typically, they are as large as necessary to capture important features. Fig. 3.2 demonstrates the procedure: On the left, an image with the overlayed patch structure is displayed. The right image corresponds to the highlighted patch in the left one.
Edges belong to the most valuable information for image analysis procedures as they carry details about the shape of an object as well as the local structures in an image. These components are described by the directional derivatives, i.e. the gradients in x and y direction. Figure 3.3 displays an exemplary image as well as its directional gradients in x and y direction. Figure 3.3(b) visualizes changes of the intensities in vertical direction, while Fig. 3.3(c) shows the same information for horizontal alterations.
Typically, the gradient magnitude (see (2.6)) is large at edges while low gradient magni-



(a) (b)

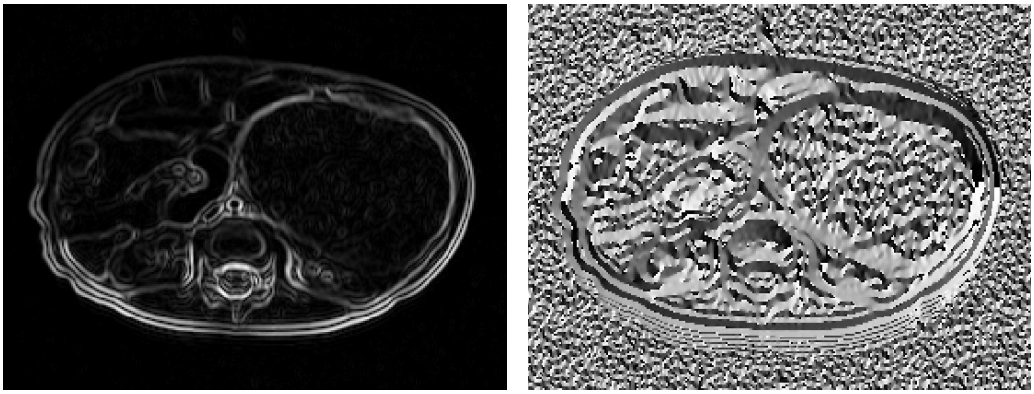
Figure 3.2: An exemplary image with indicated patch structure. The extracted patch is marked with a white border. (a) Input Image with patch structure, (b) Patch corresponding to marked area.

tudes indicate homogeneous regions as shown in Fig. 3.4(a). This information is complemented by the direction of the gradient; see Fig. 3.4(b).
An straightforward approach is now to use the histogram of oriented gradients (directional derivatives) as a feature descriptor of a local patch. The gradient information of each pixel in a patch consists of magnitude and direction, resulting in $2 \times n_1 \times n_2$ data points per patch.
Transforming these information to a histogram of gradients has one major advantage: it gets robust to noise. Individual gradients might easily be disturbed while this is rather unlikely for a complete patch. In a first step, the histogram is split in a specific number of bins - typically 9. The gradient orientation then defines the bin while the gradient magnitude contributes the value to be added. Let us for example assume, that we are given 5 pairs of gradient orientation and magnitude that we want to sort into a 9-bin histogram; see Fig. 3.5. Now, the first pair of $(20, 2)$ results in adding a value of 2 to the $20°$ bin. Please note that orientations in between histogram entries, e.g. $110°$ contribute proportionally to different positions.
Typically, HOG features are mainly used in a classification or segmentation scenario. In the first application setting, all histograms of all patches are typically concatenated to generate one huge feature vector. In the latter one, single patches and their HOG feature are mostly used to decide whether their central pixel belongs to a certain class.



(a) (b) (c)

Figure 3.3: An exemplary image and its gradients in x and y direction. (a) Input image, (b) gradients in x direction, (c) gradients in y direction. Gradient images are rescaled to [0,255] for visibility.



(a) (b)

Figure 3.4: An exemplary image, its gradient magnitude and its gradient direction. (a) Gradient magnitude, (b) Gradient direction. Image of gradient directions is rescaled to [0,255] for visibility.

In Chapter 6, we use HOG features as input for several segmentation methods and can show, that they contain important information also for Wilms' tumor segmentation.

3.2.4 Speeded-Up Robust Features

Speeded-up robust features (SURF) are an extension of the scale invariant feature transform (SIFT) [15, 104, 105]. SIFT is a powerful keypoint detector and descriptor, whose quantitative information are presumed to be invariant with respect to image scaling and small changes, rotation, and global uniform illumination changes.
The basic idea of SIFT features are to generate a scale space of the analyzed image. In a first step, the input dimensions are doubled via bilinear interpolation [104]. Afterwards, the new image is iteratively blurred with Gaussian blur and different standard deviations [105].
After several convolutions, the image is down-sampled by a factor 2 to a lower resolution and the process is repeated - each of these iterations is called an octave [104, 190]. Intuitively, the more prominent a feature is, the longer it takes to vanish in this pyramid: The feature is scale invariant. In this way, we can identify important features on different scales. Fig. 3.6 sketches this procedure: The input image in the top row is blurred

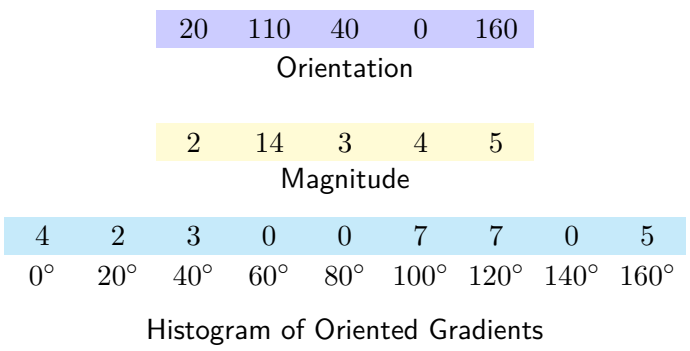


Figure 3.5: Calculation of histograms of oriented gradients.

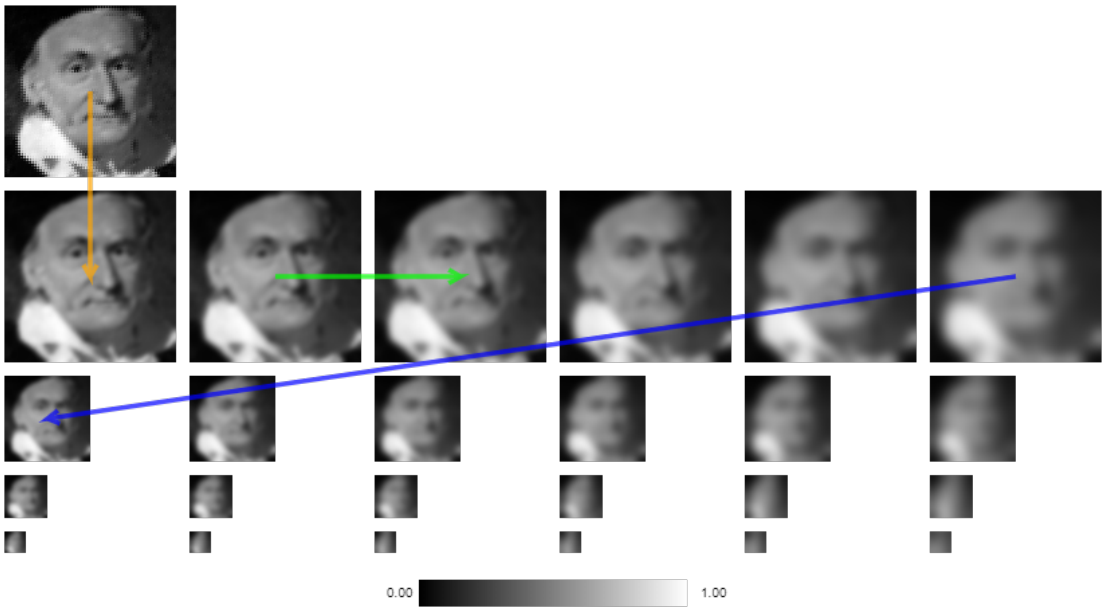


Figure 3.6: Generation of a scale space to extract SIFT features. Image courtesy of Edmund Weitz [190]

(orange arrow) and consecutively further processed via convolutions with Gaussian kernels of different standard deviation; indicated by the green arrow. After the octave, the process is repeated (blue arrow), until the image is no longer large enough anymore for convolutions.
The application of the Laplacian operator in a continuous space allows to identify points of interest: Important image information are mostly contained in high frequency areas. Unfortunately, we cannot generate such a space and have to approximate this operator. SIFT now approximates the Laplacian using a Laplacian pyramid, i.e. the differences between different levels of a Gaussian pyramid; see Fig. 3.7. Here, each image shows the differences in intensity for all horizontally neighboring images. The keypoints are then

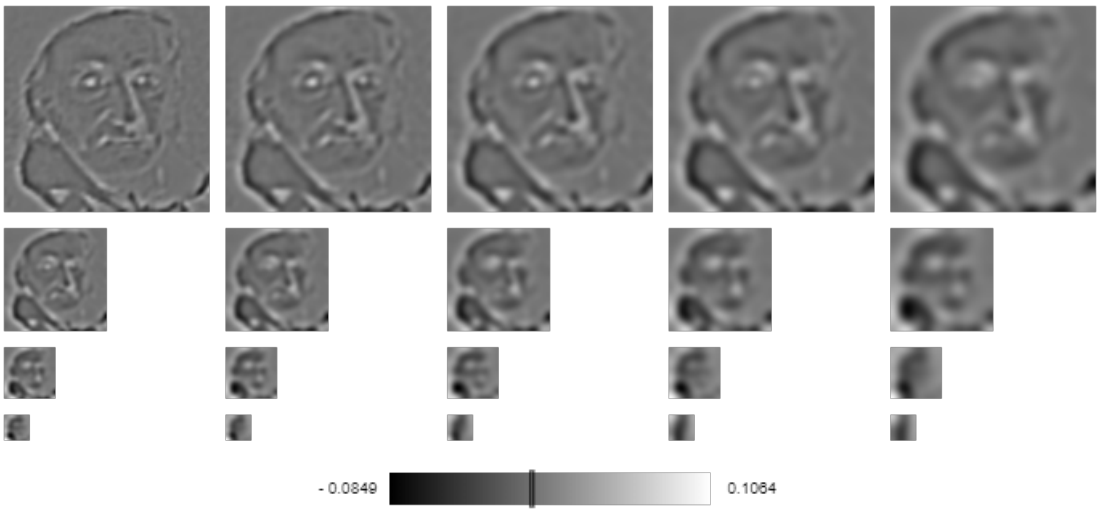


Figure 3.7: Generation of a Laplacian pyramid to approximate the Laplacian operator [190].

easily computed as extreme points of the approximated Laplacian.
Similar to HOG-features, a patch is extracted to generate an oriented histogram. In a first step, Gaussian weighted gradients are computed and accumulated to so-called sub-patches. The histograms of these sub-areas are now contain accumulated information and state the SIFT feature descriptor [105].
The SURF keypoint detector and feature descriptor is a derivative of this approach. Instead of differences of Gaussians, the Laplacian operator is approximated via box filters [15, 87]. To further speed up the generation of the needed scale space, SURF uses different scales of Gaussians instead of downsampling the image. However, the main difference of these two approaches is the feature descriptor: While SIFT uses an oriented histogram similar to HOG features, SURF extracts the sum of Haar wavelet responses around the central point of each patch. The huge topic of wavelets is beyond this thesis and we refer to [37] for more information.
Nevertheless, SURF and SIFT follow the same intuition. We use these kind of feature descriptors to generate a bag of visual words model that we use for subtype prediction in Chapter 9.

3.2.5 Bag of Visual Words Models

The idea of bag of visual words models is to represent images as an unordered set of important key points and their descriptors ("visual words") where spatial information is neglected [197]. Here, key points are the characteristic points of an image, which remain unchanged even if the image is rotated, reduced or expanded. At the beginning, the local neighborhood of each key point, the so-called local patch is extracted.
The corresponding descriptor to each key point is then given by this local patch. The

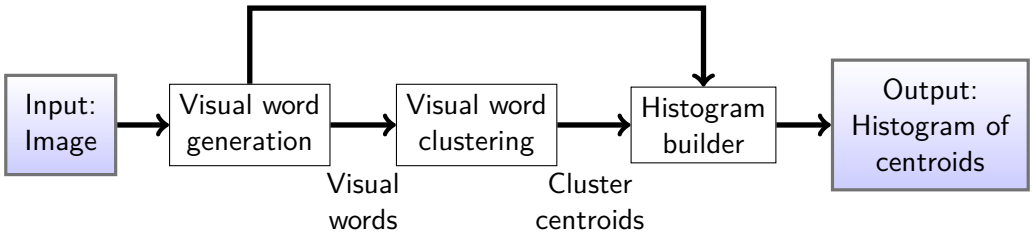


Figure 3.8: Schematic view of bag of visual words models.

basic idea is to use these pairs to create vocabularies that can be used to represent each image as a frequency histogram of the features present in the image; see Fig. 3.8. In a first step, all keypoints are clustered (usually with k-means [100]) to construct a visual vocabulary. The resulting cluster centroids are averaged representations of their class and define the visual words, where the sum of all visual words is the visual dictionary. Now, it is straight forward to represent any image using the dictionary of visual words:

1. Extract key points and their descriptors
2. Compute nearest neighbor in the dictionary
3. Count frequencies for each visual word

Hence, we can compute the histograms of discriminative visual words for all images in our data set and apply traditional machine learning approaches to classify the images based on their visual content.

4 Data Sets

"It is a capital mistake to theorize before one has data."

– Arthur Conan Doyle

Contents

In the last decades, it has become good scientific practice in the image processing community to evaluate algorithms for a specific problem on a common and established data set. This behavior triggered scientific progress from motion analysis [12, 13, 61] over optimization algorithms [86] to segmentation [112] and classification methods [48, 93]. In medical image processing, too, this habit has become more influential in the last years. Thus, more researchers get access to representative benchmark data sets [108, 115, 184]. If no such data set exists or is not open to the public, several things may happen:

1. Each data set has an inherent bias - this leads to the situation that methods, which are evaluated on different data sets, are not comparable. One cannot estimate how good or bad the approach behaves on another data set.
2. It is possible that weaknesses in one's own method are not detected as they do not occur with this particular data. Typically, it is also possible to adapt the data or its bias to the own problem.
3. Especially for data which are not easily accessible or available, the general research interest decreases. It is simply not possible to advance research into a rare disease without access to its data.

We want to address these issues and have therefore created two benchmark data sets for Wilms' tumors and made them publicly available. In the first part of this chapter

(Sec. 4.3) we discuss a segmentation data set that for the first time allows a valid and comprehensible comparability of different methods.
We then introduce our second data set, which can be used to differentiate nephroblastoma from its precursor lesion and to provide a starting point for predicting the development of the tumor under chemotherapy; see Sec. 4.4.

4.1 File Format and Data Access

Typically, MRI devices produce DICOM ("Digital Imaging and COMmunications") data, de facto the gold standard for medical imaging in clinical routine. However, since this data format contains highly sensitive patient data and a complete anonymization is difficult, we converted all MR scans to the NRRD file format [168]. NRRD stands for "Nearly Raw Raster Data" and is a standard file format for storing medical image data, fully anonymized and without sensitive patient information. The data sets are available at

- www.mia.uni-saarland.de/wilms-benchmark (segmentation data set),
- www.mia.uni-saarland.de/nephroblastomatosis (classification data set).

4.2 Research Ethics of the Study

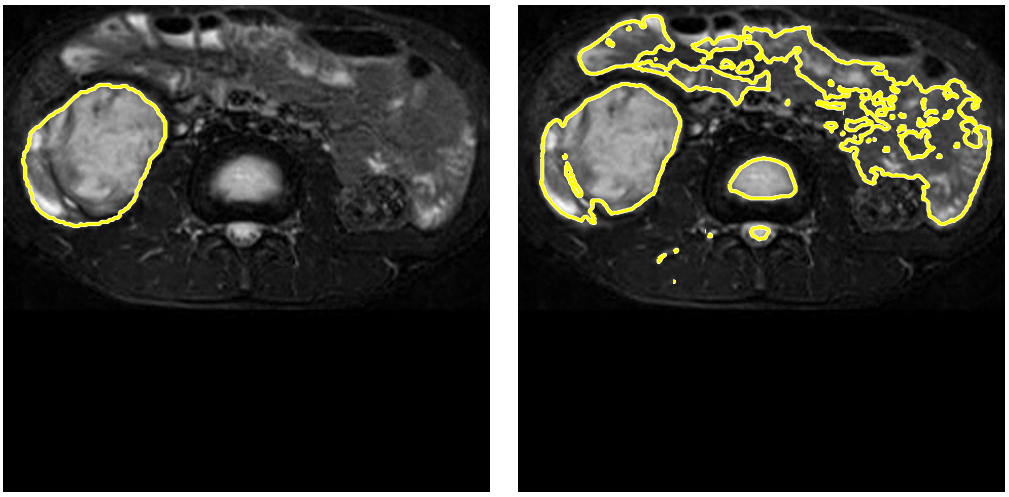
In our work, we deal mainly with volumetric images received as part of the SIOP 2001 prospective clinical trial. This trial received ethical approval from 'Ärztekammer des Saarlandes', Germany, No.: 248/13. Within this study, the parents or legal guardians gave informed consent of all enrolled children with nephroblastoma. In addition, all DICOM files were fully anonymized before analysis.

4.3 A Benchmark for Wilms' Tumor Segmentation

Image segmentation is one of the most essential building blocks in image processing. In simple terms, segmentation divides an image into related areas or regions that belong to different objects or parts of objects.
This crucial step leads us from viewing each pixel as an observation unit to classes consisting of a set of individual pixels. If this step does not work optimally, subsequent steps such as classification or tracking become much more complicated or even impossible. Therefore, any good segmentation typically has two important criteria:

- Neighboring pixels belonging to different classes differ in a property to be reproduced.
- Pixels in the same category are similar in a certain way and form a connected region.

Figure 4.1 shows examples of high and low quality segmentations. While one of them overlaps with ground truth annotations, the other one does not form a clear region and clutter dominates the result. However, it is only so obvious that the left segmentation



(a) (b)

Figure 4.1: Examples of high and low quality segmentations. (a) Segmentation overlaps with ground truth where pixels belonging to tumor region form a connected cluster. (b) Low quality segmentation. Many single pixels are labeled as tumor, no clear connected and unique region is visible.

surpasses the other because they are evaluated on the same data set: The comparison is fair.
In order to ensure this fairness also for the evaluation of different segmentation methods and their results on Wilms' tumors, we compiled a benchmark data set. Analogous to other medical data sets we include several different MRI sequences [108, 115]: From T_1 images that represent the general structure of the abdomen (fat is light, water dark), to T_2 sequences that better represent water-rich regions, to T_{1c} images with contrast agents that improve the visibility of body and tumor structures; see Section 2.2.2.
We included in total 28 of these multi-sequence MRI scans from 17 Wilms' tumor patients (5 male, 12 female), out of which 15 have been acquired from intermediate risk tumor

Table 4.1: Detailed overview of patients included in our benchmark data set. Age is stated in month. Mixed: mixed histology, regressive: regressive subtype, stromal: stromal predominant, blastemal: blastemal predominant.

ID	Subtype	Age	Pre-Chemotherapy	Post-Chemotherapy
Subject 1	mixed	1	✓	✓
Subject 2	regressive	3	✓	✓
Subject 3	blastemal	4	✓	✓
Subject 4	stromal	5	✓	✓
Subject 5	regressive	8	✓	✓
Subject 6	regressive	3	✓	✗
Subject 7	mixed	5	✓	✓
Subject 8	regressive	4	✓	✓
Subject 9	mixed	5	✗	✓
Subject 10	blastemal	3	✓	✓
Subject 11	regressive	2	✓	✓
Subject 12	mixed	4	✓	✗
Subject 13	regressive	3	✗	✓
Subject 14	mixed	1	✓	✓
Subject 15	regressive	3	✓	✗
Subject 16	mixed	3	✓	✓
Subject 17	stromal	7	✗	✓

(histological diagnosis: stromal predominant (2), mixed histology (6) or regressive type (7)) and 2 from high risk tumor types (histological diagnosis: blastemal predominant); see Section 2.5. For eleven patients, we have both data before and after chemotherapy. The remaining ones are missing either data before or after chemotherapy. Table 4.1 shows a detailed overview of these characteristics. In addition, Fig. 4.2 highlights the age distribution of the children.

We made sure that only patients with histologically confirmed Wilms' tumors were eligible for inclusion. The MRI sequences before and after chemotherapy for one of these patients are shown in Fig. 4.3.

Since it is difficult to obtain a comprehensive and representative set of Wilms' tumor data, the images have been acquired at different centers over the course of several years, using MR scanners from different manufacturers, varying field strength (1.5T and 3T) and implementations of the imaging sequences. However, we ensured that imaging sequences were as similar as possible.

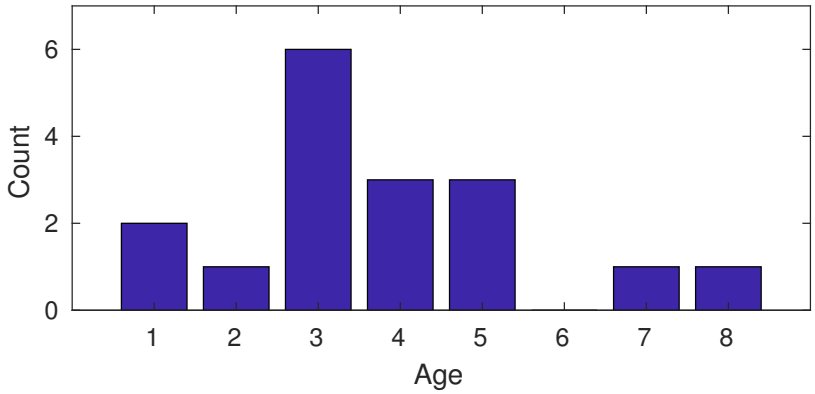


Figure 4.2: Age distribution of patients whose images are made available anonymously.

In total, the data sets used in our benchmark share the following three MRI settings; see Sec. 2.2.2:

- T_2: T_2-weighted images, axial 2D acquisition with 3.6 mm to 9.1 mm slice thickness and inslice-sampling ranging from 0.3 mm to 1.4 mm.
- T_1: T_1-weighted images, native image, axial 2D acquisition with 2.5 mm to 9.1 mm slice thickness and inslice-sampling ranging from 0.5 mm to 1.6 mm.
- T_{1c}: T_1-weighted and contrast enhanced (Gadolinium) images, axial 2D acquisition with 1.8 mm to 7.7 mm slice thickness and inslice-sampling ranging from 0.5 mm to 1.6 mm.

Originally, the images were spatially not registered. In order to align the different MRI images, we therefore manually co-registered the sequences on the T_2 sequence using a rigid transformation. Besides, we balanced the number of slices with tumor areas before and after chemotherapy; see Tab. 4.2. Unfortunately - due to limited amount of data - it was not possible to also balance the subtypes among the data sets.

Table 4.2: Image properties before and after chemotherapy. The values in brackets indicate the average occurrence.

	Training Set		Test Set	
	Slices	Tumor	Slices	Tumor
Pre-Chemo	$19-55$ (31)	$9-25$ (15)	$26-50$ (35)	$11-28$ (18)
Post-Chemo	$19-44$ (30)	$6-26$ (12)	$29-70$ (54)	$6-23$ (13)

4.3.1 Annotations by Human Experts

Since the MRI sequences are recorded non-invasively, it is necessary either to have the data annotated by human experts or to use the results of one or more segmentation

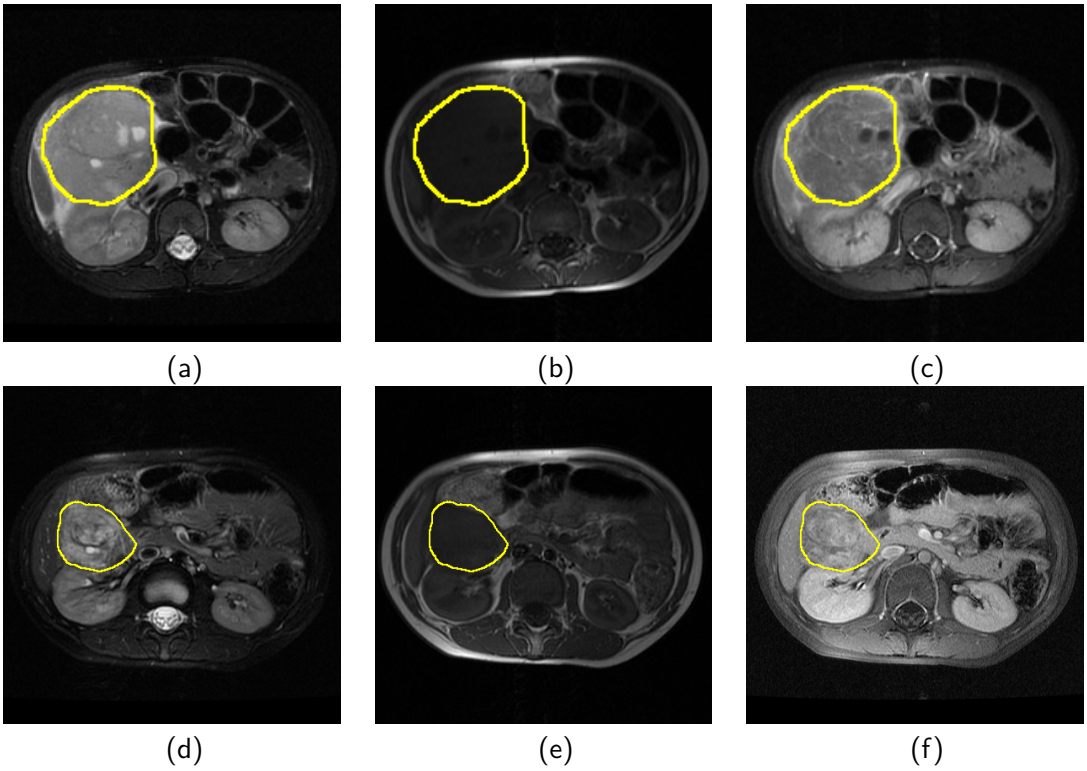


Figure 4.3: Example of Wilms' tumor (training data) before ((a)-(c)) and after ((d)-(f)) chemotherapy with experts' consensus truth. From left to right: T2, T1, T1c.

methods. A major drawback of no direct measurement is the fact that the ground truth suffers from an inherent bias regarding these particular methods.
For this reason, we decided to have the data annotated by humans. In order to guarantee a valid annotation that does not contain any preferences of a particular expert, we had a total of 5 different specialists to annotate the data.
Rater-1 and rater-4 are experienced radiologists with several years of experience in Wilms' tumor analysis. Rater-2 is a physician familiar with Wilms' tumors. Rater-3 is a M.D. student previously trained in MRI imaging with advanced experience in the field. Rater-5 is an experienced oncologist with decades of practice in Wilms' tumor exploration.
All the human experts were given instructions on how to make the annotations in advance with the given MITK software (www.mitk.org) and no one had seen the data before. Afterwards the raters outlined tumor structures in the data sets on the T_2 sequences in every axial slice.

4.3.2 Ground Truth Generation

Since the generation of error-free ground truth information for medical images is usually not possible, we rely on expert votes to approximate the tumor area. Majority voting for each voxel has been shown to be useful in several contexts [76, 143]. Unfortunately, this simple approach neither regards variability in quality or performance amongst the human raters nor does it provide guidance as to how many experts should agree before

a voxel is labeled as tumor. Hence, we decide to use the STAPLE framework [187] to produce consensus segmentations.
The STAPLE algorithm uses expectation maximization. Let Ω_i, $i = 1, \ldots, n$ be the expert decisions and $\hat{\Omega}$ the true consensus segmentation. The performance of each expert is rated on the basis of the *sensitivity* or true positive rate (see Sec. 2.4)

$$p_j = \frac{|\Omega_i \cap \hat{\Omega}|}{|\hat{\Omega}|}$$

and the *specificity* or true negative rate

$$q_j = \frac{|\Omega_i \cap \hat{\Omega}|}{|\mathbf{1} - \hat{\Omega}|}.$$

It iterates between estimating the conditional probability of $\hat{\Omega}$ in relation to the expert decisions and previous estimates of the performance parameters and estimation of updated reliability parameters.
Before chemotherapy, convergence is on average reached with less than 33 iterations. After chemotherapy, the algorithm converged on average after 52 iterations. The estimated quality parameters of each expert are shown in Tab. 4.3 and indicate high inter-rater variability. There is obviously also a noticeable difference between the radiologists. Please

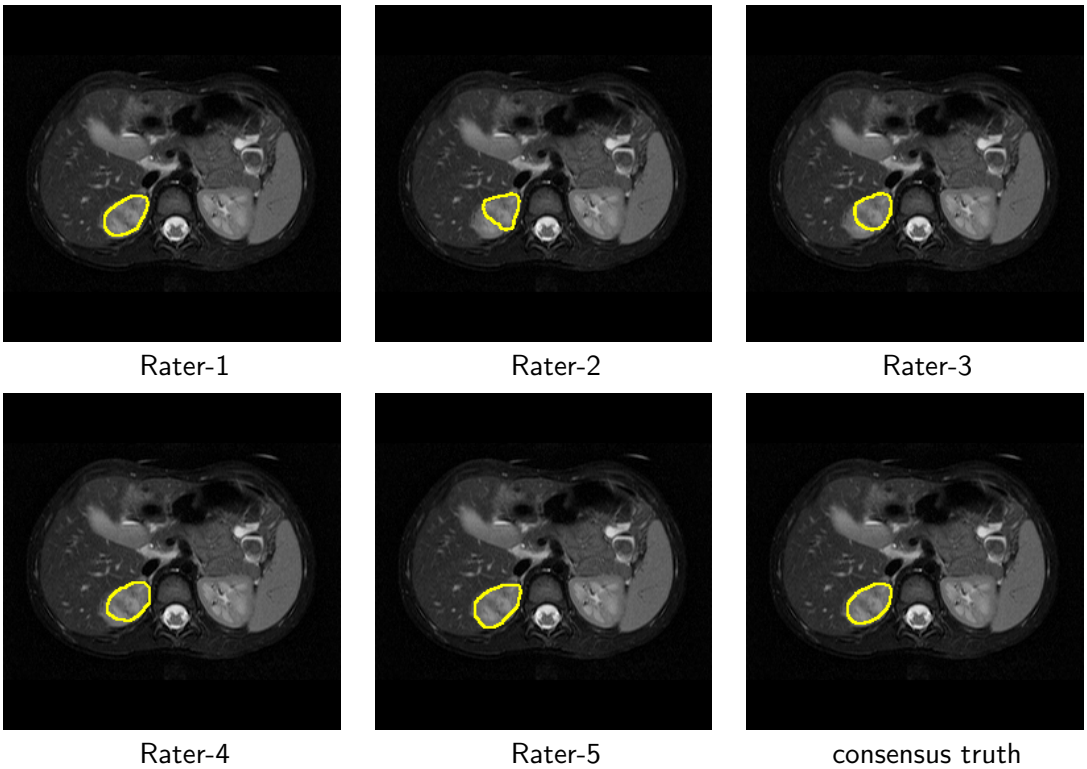


Figure 4.4: Example annotations by human expert raters. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

consider the results before chemotherapy: While rater-1 shows a high tendency to mark the complete tumor (sensitivity: 0.76) but also risks to label more non-tumor parts (specificity: 0.65), rater-4 is more hesitant and tends to indentify less areas as tumors where

he is confident (specificity: 0.82) - and thus indirectly accepts to mark too few areas (sensitivity: 0.65). We can observe this for the imaging after chemotherapy, too.
These tendencies are also visually reflected in the annotations: Fig. 4.4 shows tumor

Table 4.3: Estimated quality parameters of each expert before and after chemotherapy. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4	Rater-5
Pre-Chemotherapy					
Sensitivity	0.76	0.71	0.80	0.65	0.58
Specificity	0.65	0.75	0.73	0.82	0.74
Post-Chemotherapy					
Sensitivity	0.78	0.60	0.77	0.70	0.67
Specificity	0.72	0.57	0.75	0.85	0.82

outlines from all five human experts and the final ground truth approximation. While rater-1 labels the complete region, the area outlined by rater-4 is smaller, but coincides with the consensus truth approximation for each pixel marked as tumor. We discuss the differences between the human expert annotations in detail in Chapter 5.

4.3.3 Summary

In this section, we introduced the first heterogenous multi-sequence MRI benchmark data set for Wilms' tumor segmentation. It allows for a fair and reproducible comparison of segmentation methods, and training of human experts to improve their annotation accuracy. In Chapter 6, we extensively evaluate several out of the box methods for fully automatic segmentation. Furthermore, we introduce a semi-automatic baseline algorithm that all new approaches for Wilms' tumor segmentation should be compared with in Chap. 6. This segmentation approach is also used in the next section to generate a Wilms' tumor classification benchmark.

4.4 Differential Diagnostics and Subtype Determination

Image interpretation or classification, although often effordless for human observers, remains a major challenge in computer vision. Typically, the goal is to build or approximate a model describing objects in an image. This approximation is then used to decide whether an image belongs to a certain class, or not. In case of medical applications, a certain object or anomaly is to be detected to support physicians in their daily clinical routine. In general, these computer assisted decision systems help to minimize the human error. Thus, their classification models should satisfy the following important criteria:

- Objects that belong to the same class have at least one property in common.
- Global disturbances (e.g. noise, blur, etc) have none to little influence on the final result - the method is robust.

However, in some cases it is difficult to find appropriate properties or *features* to distinguish objects not belonging to the same class. Fig. 4.5 shows three different MR images containing tumors and abdominal masses: While nephroblastomatosis is a benign

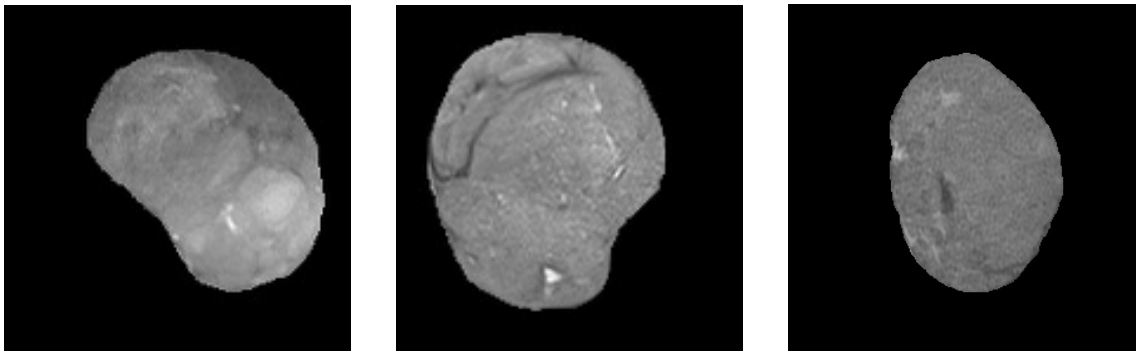


Figure 4.5: Examplary images to be classified. Although each of them represents a different class of Wilms' tumor subtype, their visual appearance is highly similar.

lesion, the others are subtypes of the malignant renal tumor nephroblastoma; see Section 2.5. Obviously, it is not easily possible to distinguish these types of disease. Although clinicians assume nephroblastomatosis to be a homogeneous and small object, there is neither a clear threshold at which a mass is considered to be a nephroblastomatosis rather than a nephroblastoma, nor a reproducible and mathematical sound evaluation if these assumptions hold.
Typically, the treatment of nephroblastoma begins with a chemotherapy, followed by a surgery and continues with chemotherapy or irradiation of the tumor bed (except in rare cases when the specimen is completely necrotic); see Section 2.5.3. In order to adopt the therapy for the patient as soon as possible, it would be essential to distinguish the subtypes of Wilms' tumor based on their imaging data before chemotherapy - a biopsy before extraction is unfortunately not possible.
Hence, there is massive need for a data set that meets the following requirements:

- It allows for a scrutinization of current clinical assumptions regarding the visual appearance of nephroblastomatosis in comparison to nephroblastoma.

- It enables the development and examination of classification methods for nephroblastomatosis and nephroblastoma.
- The data set provides the opportunity to evaluate approaches for Wilms' tumor subtype prediction.

With a focus on this dilemma, we compiled a data set of 202 patients, out of which 148 suffer from a nephroblastoma and 52 are affected by nephroblastomatosis; see Tab. 4.4.

We included only patients with histologically confirmed diagnoses and excluded patients with kidneys affected by both nephroblastoma and nephroblastomatosis. Hence, all contained MR sequences fulfill the following criteria:

- Histology was confirmed by a pathologist.
- In case of bilateral tumors or renal masses, both kidneys are affected by the same subtype.
- All patients with nephroblastomatosis did not suffer from a Wilms' tumor.

In this way, we can minimize human error mixing up left and right kidney and can provide a representative data set where subtypes are not biased towards single patients.
The gold standard in MR sequences for Wilms' tumor segmentation are T_2 images. Typically, radiologists approximate tumor's outline on this modality. Hence, T_2 images are always acquired during the therapy protocol. In order to allow for applicability to daily clinical routine, we therefore decided to restrict ourselfs to these kind of data. In order to reduce parameter noise due to different imaging settings in MR machines, we made sure that the main parameter settings of the T_2 sequences are as similar as possible - this has drastically reduced the amount of imaging data available.
Nevertheless, we have compiled so far the largest and most complete compilation of nephroblastoma and nephroblastomatosis. All data sets are T_2-weighted images (axial 2D acquisition) with 3.6 mm to 9.1 mm slice thickness and inslice-sampling ranging from 0.3 mm to 1.4 mm.

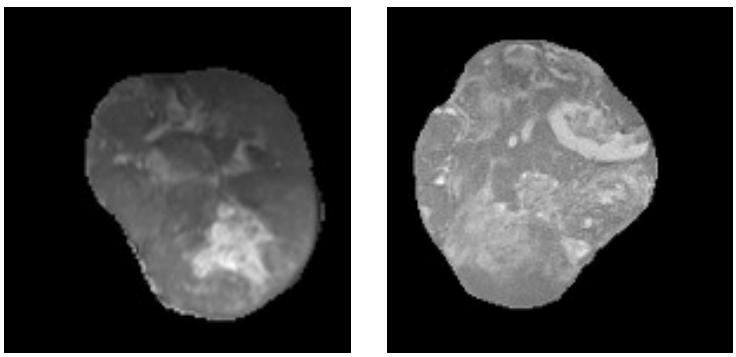


Figure 4.6: Two exemplary images from our data set with epithelial (left) and stromal dominant (right) subtypes. Tumor areas are extracted and embedded in a square shaped black background. No additional context is available.

Table 4.4: Detailed information about our data set.

Patient characteristics		
Age	range (month)	$1-153$
	average	34.3
Gender	female	50.9%
	male	49.1%
Metastasis (Wilms' Tumor)		22 (14.86%)
Tumor characteristics		
Nephroblastoma Subtypes	diffuse anaplastic	3
	blastemal	18
	regressive	50
	mixed	29
	stromal	28
	epithelial	17
	necrotic	3
	total	148
Nephroblastomatosis		54
Total		202

4.4.1 Segmentation

Classifiers try to separate objects based on their features. In some contexts, it might be useful to also consider the surrounding areas for this decision. It is for example much more likely that a tree stands on a pasture than on water. This information can then be included in the classification procedure. Unfortunately, data sets including context information are usually extremely large. In case of imaging data of Wilms' tumors, we cannot provide huge amounts of data and therefore have to limit the context.
MR images of the renal cortex contain several organs like liver, kidneys, intestine, or bladder. However, not all of these organs are visualized in each of the included images - while some show the liver but miss the bladder, others contain all of these. In order to minimize this kind of variance and to restrict the influence of the image context, we decided to segment Wilms' tumors before classification.
In a first step, we resampled all images to a grid size of one in x and y direction, but refrained from resampling in z direction as the interpolation error would be too high. Then, we rescaled image intensities for simplicity to the interval $[0, 1]$. In the end, a human expert with years of experience in the field of nephroblastoma annotated the tumor regions using our method that we describe in Chap. 6. We mask everything except the tumor areas and embed them in a square shaped image; see Fig. 4.6.

4.4.2 Preprocessing for Classification

In classification data sets, it is important to map the distribution of the appearance as comprehensively and completely as possible. Typically, researchers use data augmentation and rotate, flip or shear the available images [194]. In case of nephroblastoma, this does not lead to an information gain - its outer shape is heavily influenced by surrounding

tissue, and organs. This makes data augmentation by simple transformations difficult. We address this problem in a different way by using its inhomogeneity. Wilms' tumor is a very heterogeneous mass (also due to its three contributing tissues blastemal, epithelial, and stromal) such that even in one of these objects the differences are typically massive; see Fig 4.7. This applies not only to the tumor shape, but also to the inner areas. Bleeding and lesions can occur, some areas are better, others less well supplied with blood. For this reason, we have added three slices of each of the included patients to the data set: One from the upper third, the middle slice, and one from the lower third. In this way, we can naturally enlarge the data set.

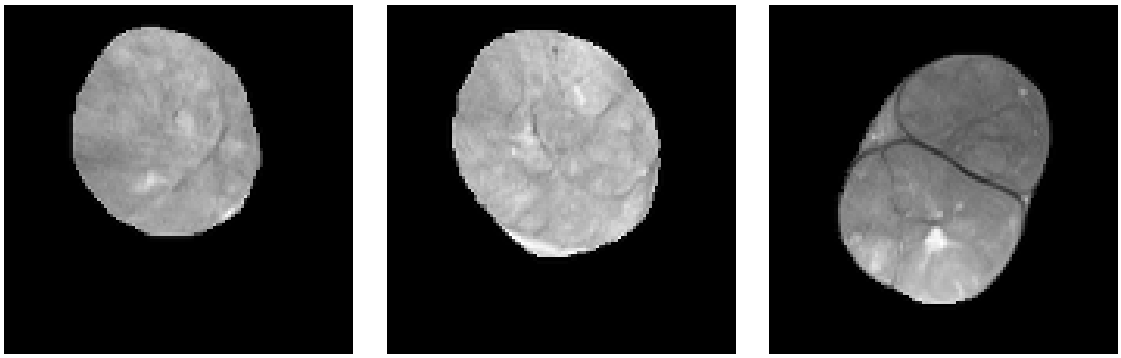


Figure 4.7: Exemplary images from our data set with top, middle, and bottom slice to best represent heterogeneities within each tumor.

4.4.3 Summary

Classification and especially subtype determination of Wilms' tumors is a challenging task. We provide for this purpose a large data set compiled of more than 200 abdominal masses. It addresses two main issues: First, we can evaluate the visual differences between nephroblastoma and its precursor lesion nephroblastomatosis in Chap. 8. Second, it allows us to make first attempts to distinguish the subtypes of Wilms' tumors at beginning of treatment in Chap. 9.

4.5 Conclusions

In this chapter, we introduced two data sets that allow us further investigation of Wilms' tumors and their imaging data. The first one, a multi-sequence segmentation benchmark addresses several problems. It provides us with the necessary information to evaluate human expert annotations. Furthermore, we can analyze the current clinical practice of approximating the tumor volume with an elliptic shape; see Chap. 5. And last but not least, it empowers us to evaluate out of the box methods for Wilms' tumor segmentation and to develop a semi-automatic method adjusted for these kind of disease in Chap. 6. Our second data set addresses a different issue in image processing for Wilms' tumors: the classification problem. Nephroblastoma are visually similar to their precursor lesion, the nephroblastomatosis. One main aspect of this data set is its ability to provide data to verify current assumptions about visual appearance and to identify sufficient features for

their distinction; see Chap. 8. In addition, this data set paves the way for first attempts of subtype determination before chemotherapy.
In the next chapter, we first evaluate inter-rater variability of human experts, i.e. how reliable an annotation of a single human is and how much humans vary in their decisions about tumor and non-tumor areas. We further investigate the precision of current clinical practice with respect to tumor volumes.

5 Human Expert Segmentations and Clinical Practice

"Any man can make mistakes, but only a idiot persists in his error."

– Marcus Tullius Cicero

Contents

Since centuries, it is clinical practice that human experts decide based on examinations, imaging data, their experience and intuition about the therapeutic process. In case of examinations and measured properties such as complete blood counts, decisions are reproducible and typically reasonable for external experts, not involved in the current decision process. However, intuition and experience are no measurable values making reproducibility a difficult to nearly impossible task.

Traditionally, radiologists assess tumor outlines and its development based on imaging data. Unfortunately, experience plays a major role in accuracy of tumor annotations marked by a domain expert - thus it is not possible to reproduce the decision process, or reasons for the actual annotation. Mazzara et al. [113] evaluated inter- and intra-rater variability in case of brain tumor annotation, i.e. variability between raters and the same rater at different points of time. Unfortunately, it turned out, that domain experts show high inter- and intra-rater variability: They neither agreed with other experts, nor with themselves exactly. This introduces a strong human bias, avoiding reproducibility and reliability.

Similarly, it is common practice in the therapy planning of nephroblastoma, that a single radiologist is delineating tumor areas. From this annotation the clinical volume is then approximated, which among other aspects characterizes the response of a patient to chemotherapy; see Sec. 2.5. Unfortunately, the reliability and consistent reproducibility of expert delineations of Wilms' tumors has not been investigated so far.

In addition, tumor volume is typically approximated by an ellipsoid shape. Here, a radiologist measures its expansion for all three axes, i.e. width, height, and depth [66]. Based

on this, the largest ellipsoid that fits into the corresponding cuboid is calculated. Unfortunately, there is no valid mathematical evaluation whether the assumption of an ellipsoidal shape actually applies. However, both variants - manual annotation and approximation - are performed on the basis of MR images. All in all, the important questions can be summarized as:

- Is it reasonable to base treatment planning on annotations and measures of a single radiologist?
- How reliable is the current gold standard in volume measurement for Wilms' tumors? Is the approximation by an ellipsoid shape valid? How accurate is this procedure?

We address these problems in the following and first investigate the inter-rater variability among human experts in Sec. 5.1. In addition, we analyze their individual discrepancies to the consensus truth of all human annotations. Afterwards, we evaluate the current gold standard of approximating the tumor volume with an ellipsoid in Sec. 5.2. In the end, we give a short summary of our findings.

5.1 Evaluation of Human Expert Segmentations

Since the last decades, it is gold standard in Wilms' tumor treatment planning, that a reference radiologist determines tumor extents for all axes based on MRI data. This information plays a major role in further planning of the course of therapy: volume and the local stage determine whether and which further chemotherapy or radiation is necessary; see Sec. 2.5. A precise assessment is obviously of crucial importance.
Unfortunately, Wilms' tumors have neither a discriminative texture (due to their heterogeneity), nor do they always show a clear border and can be directly attached to the remaining kidney. In addition, MR images can be of low quality - image sequence acquisitions cannot be repeated infinitely many times as anesthesia of toddlers is limited in its application [78].
Although the reference radiologist typically has decades of experience in Wilms' tumor analysis, these circumstances result in a possible human bias. Hence, the overall question if tumor annotation of a single radiologist is reliable enough for treatment planning, can be split into several steps:

- How large is the inter-rater variability, i.e. to which extent do domain experts agree on tumor outlines?
- How much do the exact pixel-based volumes differ?
- How strong is the impact on treatment planning?

In order to investigate the current situation, we first analyze inter-rater variability among human experts to access the reliability of single domain experts. Afterwards, we evaluate its effect on volume determination and treatment planning.

5.1.1 Inter-Operator Variability

In order to evaluate inter-rater (or inter-operator) variability, we make use of our benchmark data set for Wilms' tumor segmentation in Sec. 4.3. Here, five human experts outlined tumor areas for all data sets. We use these annotations to calculate the inter-operator variability using all 28 data sets of all 17 patients. In order to do so, we compute the pixelwise disagreement of the outlined volume marked by each physician with each volume outline prepared by each of the other four clinicians for the same data set.
This process was repeated for each patient to provide a data set comprising the average disagreement between the five contours for each data set. We also divide the data sets based on their acquisition time relative to chemotherapy, i.e. before and after chemotherapy. Tab. 5.1 shows the inter-operator variability in terms of Dice score before chemotherapy, and after chemotherapy, respectively. Before chemotherapy, the average Dice score between human experts shows their agreement on average with 0.87 ± 0.05 on tumor areas. After chemotherapy, when tumor tissues are barely visible, the average Dice score between human expert raters drops down to 0.78 ± 0.11 indicating a high inter-rater variability.
This coincides with the specificity and sensitivity measures determined by the STAPLE algorithm, that we use to generate the consensus approximation; see Tab. 4.3. We show a detailed analysis of the inter-rater variability for each domain expert in Tab. 5.2.

Table 5.1: Inter-operator variability before and after chemotherapy in terms of Dice score. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4
Pre-Chemotherapy				
Rater-2	0.85 ± 0.13			
Rater-3	0.89 ± 0.11	0.89 ± 0.08		
Rater-4	0.85 ± 0.13	0.90 ± 0.05	0.88 ± 0.08	
Rater-5	0.83 ± 0.13	0.89 ± 0.05	0.87 ± 0.07	0.89 ± 0.05
Post-Chemotherapy				
Rater-2	0.69 ± 0.37			
Rater-3	0.83 ± 0.24	0.70 ± 0.37		
Rater-4	0.84 ± 0.10	0.70 ± 0.36	0.80 ± 0.24	
Rater-5	0.84 ± 0.10	0.69 ± 0.35	0.80 ± 0.24	0.89 ± 0.05

This allows us to evaluate the single annotators: We investigate domain experts' behavior based on Dice score, precision and recall. Let us assume, we want to compare rater-1 and rater-2. We define the annotations by rater-2 as ground truth and compute our quality measures for rater-1. In this way, we can evaluate annotation behavior of each expert in relation to all others. Please note that precision and recall are exchanged in this setting for the respective reverse case.

Rater-1: Radiologist

Rater-1 is an experienced radiologist with years of experience in Wilms' tumor analysis. Before chemotherapy, he agreed on average with a Dice score of 0.86 with the other human experts on tumor data. However, his annotated tumor outlines are more generous than the others. Hence, his average precision of 0.79 is much lower than his mean recall rate of 0.96. Surprisingly, his precision with respect to the annotations of the other radiologist as well as the oncologist is even lower, while the recall rate is slightly higher.
After chemotherapy, his annotations overlap in average on 0.80 with tumor outlines of the other domain experts; see Tab. 5.3. Similarly, his precision and recall rate drop to 0.75 and 0.88 respectively. However, his agreement with the annotations of the second radiologist as well as the oncologist stays roughly the same.

Rater-2: Physician

Rater-2 is a physician familiar with Wilms' tumors. His agreement before chemotherapy with other domain experts on tumor areas is relatively high with a Dice score of 0.88; see Tab. 5.3. This indicates that tumor areas are easy to identify and their outlines are more or less obvious. However, after chemotherapy, when tumor regions are barely visible, his overlap with the other human experts decreases dramatically to 0.69. Both, precision as well as recall are diminished by 0.17 and 0.18 respectively, resulting in a

Table 5.2: Inter-operator variability before and after chemotherapy in terms of precision and recall. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist. The first line in each cell determines the average precision, while the second one is the average recall value.

	Rater-1	Rater-2	Rater-3	Rater-4	Rater-5
Pre-Chemotherapy					
Rater-1	- -	0.81 ± 0.19 0.94 ± 0.07	0.85 ± 0.15 0.95 ± 0.03	0.77 ± 0.18 0.98 ± 0.02	0.75 ± 0.18 0.97 ± 0.03
Rater-2	0.94 ± 0.07 0.81 ± 0.19	- -	0.91 ± 0.07 0.88 ± 0.13	0.86 ± 0.08 0.95 ± 0.04	0.85 ± 0.08 0.94 ± 0.04
Rater-3	0.95 ± 0.03 0.85 ± 0.15	0.88 ± 0.13 0.91 ± 0.07	- -	0.84 ± 0.13 0.95 ± 0.05	0.82 ± 0.11 0.95 ± 0.05
Rater-4	0.98 ± 0.02 0.77 ± 0.18	0.94 ± 0.04 0.86 ± 0.08	0.95 ± 0.05 0.84 ± 0.13	- -	0.89 ± 0.08 0.90 ± 0.06
Rater-5	0.97 ± 0.03 0.75 ± 0.18	0.94 ± 0.04 0.85 ± 0.08	0.95 ± 0.05 0.82 ± 0.11	0.90 ± 0.06 0.89 ± 0.08	- -
Post-Chemotherapy					
Rater-1	- -	0.68 ± 0.37 0.74 ± 0.33	0.81 ± 0.25 0.85 ± 0.25	0.75 ± 0.15 0.97 ± 0.02	0.75 ± 0.14 0.97 ± 0.02
Rater-2	0.74 ± 0.33 0.68 ± 0.37	- -	0.74 ± 0.33 0.70 ± 0.37	0.68 ± 0.31 0.76 ± 0.39	0.68 ± 0.30 0.75 ± 0.38
Rater-3	0.85 ± 0.25 0.81 ± 0.25	0.70 ± 0.37 0.74 ± 0.33	- -	0.74 ± 0.25 0.89 ± 0.26	0.74 ± 0.23 0.89 ± 0.26
Rater-4	0.97 ± 0.03 0.75 ± 0.15	0.76 ± 0.39 0.68 ± 0.31	0.89 ± 0.26 0.74 ± 0.25	- -	0.89 ± 0.05 0.89 ± 0.08
Rater-5	0.97 ± 0.02 0.75 ± 0.14	0.75 ± 0.38 0.68 ± 0.30	0.89 ± 0.26 0.74 ± 0.23	0.89 ± 0.08 0.89 ± 0.05	- -

quite poor annotation outcome: His Dice score drops to 0.69. In two situations rater-2 fails completely: In one case he does not detect the tumor at all, in the other he marks a different area.

Rater-3: M.D. Student

Rater-3 is a M.D. student with advanced experience in the field of Wilms' tumor annotation. Similar to rater-2, his tumor outlines before chemotherapy show a large overlap with these of the other domain experts, indicated by a Dice score of 0.88; see Tab. 5.3. However, while rater-2 has a slightly higher precision, rater-3 annotates more tumor areas, i.e. recall rate is higher; see Tab 5.2. After chemotherapy, the medical student seems to have not enough experience to always detect the tumor. Hence, he missed one specimen after chemotherapy, resulting in lower precision and recall values; see Tab. 5.5.

Rater-4: Radiologist

Rater-4 is also an experienced radiologist with profound experience in Wilms' tumor analysis. Before chemotherapy, he shows large overlap with the other domain experts. In

Table 5.3: Averaged quality measures of each expert in comparison to the others. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4	Rater-5
Pre-Chemotherapy					
Precision	0.79 ± 0.06	0.89 ± 0.05	0.87 ± 0.07	0.94 ± 0.04	0.94 ± 0.04
Recall	0.96 ± 0.03	0.89 ± 0.07	0.91 ± 0.06	0.84 ± 0.08	0.83 ± 0.08
Dice Score	0.86 ± 0.04	0.88 ± 0.04	0.88 ± 0.04	0.88 ± 0.05	0.87 ± 0.04
Post-Chemotherapy					
Precision	0.75 ± 0.11	0.71 ± 0.08	0.76 ± 0.09	0.90 ± 0.09	0.87 ± 0.10
Recall	0.88 ± 0.05	0.72 ± 0.04	0.83 ± 0.06	0.77 ± 0.09	0.77 ± 0.09
Dice Score	0.80 ± 0.08	0.69 ± 0.07	0.78 ± 0.05	0.81 ± 0.08	0.81 ± 0.08

direct comparison with the first radiologist, he is more hesitant in marking tumor regions. This results in a higher precision of 0.94 (average 0.89), while his recall rate of 0.84 is slightly below the average recall of 0.89; see Tab. 5.2 and Tab. 5.3. After chemotherapy, rater-4 shows the same annotation behavior and tends to mark only tumor regions where he is confident - his agreement on tumor areas coincides mostly with the other domain experts. Hence, his precision rate stays on nearly the same level of 0.90 (average 0.80). However, due to his cautious annotations, recall value drops to 0.77 (average 0.79).

Rater-5: Oncologist

Rater-5 is an experienced oncologist with decades of practice in Wilms' tumor exploration. Due to his extensive knowledge, his annotations might be qualitatively on the same level as those of experienced radiologists. Indeed, before chemotherapy, his annotations show a strong overlap with the remaining domain experts. Hence, his average precision and recall values of 0.94 and 0.83 respectively, were to be expected. However, after chemotherapy, when tumor outlines are barely visible and its structure vanishes, he shows a similar behavior to rater-4: His tumor outlines are cautious, resulting in a high precision of 0.87 remarkably above average. Analogously to rater-4, his recall reduces to 0.77.

All in all, it turns out that annotations of rater-4 and rater-5 are most reliable in direct comparison to our other domain experts, especially after chemotherapy. Surprisingly, tumor outlines of rater-5 (oncologist) outperform annotations of rater-1 (radiologist) with respect to precision.
However, before chemotherapy, all domain experts more or less agreed on tumor outlines. After chemotherapy, inter-rater variability increases dramatically and human bias towards preferences and intuition becomes more important. Hence, annotation of Wilms' tumors obviously becomes more challenging after chemotherapy.

5.1.2 Deviation from Consensus Truth

Before, we analyzed the discrepancies between different human expert annotations. Now, we investigate their variation from the generated consensus truth. This approximated ground truth reflects on the one hand reliability of the domain experts' tumor outlines and on the other their consensus about tumor regions.
The comparison of human annotators with the generated ground truth (Tab. 5.4) shows basically the same tendency as the previous analysis of the experts among themselves: Rater-1 has lower precision but higher recall rates before and after chemotherapy. Annotations of rater-2 are competitive in the beginning of treatment, but lose accuracy after chemotherapy. Similarly, tumor outlines of rater-3 reveal the same behavior as in the direct comparison between raters. Their overlap with consensus truth after chemotherapy drops slightly due to missed tumor regions. Rater-4 and rater-5 have similar performance on the approximated ground truth - both show a high precision before and after chemotherapy.
In this way, the consensus truth incorporates humans' knowledge about tumor regions and is more reliable than single domain experts annotations. The average Dice score before chemotherapy of human raters in comparison to ground truth is 0.92 ± 0.06. After chemotherapy, the contrast of tumor regions is usually lower and the tumor outlines are more ambiguous. Consequently, human experts agree less on tumor areas. The average Dice score decreases to 0.85, and variability increases dramatically to 0.16.

Table 5.4: Comparison of human domain experts and their consensus truth. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4	Rater-5
Pre-Chemotherapy					
Precision	0.83 ± 0.18	0.93 ± 0.06	0.91 ± 0.12	0.97 ± 0.02	0.96 ± 0.04
Recall	0.98 ± 0.02	0.95 ± 0.04	0.96 ± 0.05	0.90 ± 0.06	0.88 ± 0.08
Dice Score	0.89 ± 0.13	0.94 ± 0.04	0.93 ± 0.07	0.93 ± 0.03	0.92 ± 0.05
Post-Chemotherapy					
Precision	0.83 ± 0.19	0.74 ± 0.33	0.83 ± 0.26	0.95 ± 0.07	0.97 ± 0.04
Recall	0.97 ± 0.04	0.74 ± 0.38	0.89 ± 0.26	0.88 ± 0.11	0.89 ± 0.08
Dice Score	0.88 ± 0.13	0.72 ± 0.36	0.85 ± 0.25	0.91 ± 0.07	0.92 ± 0.04

Tumor volume is an important decision marker for follow up treatment after chemotherapy; see Sec. 2.5.3. The generated ground truth allows us to investigate the differences of human experts in comparison to the real volume; see Tab. 5.5. We computed the ranges of deviations for each domain expert. Here, *Min* indicates the most negative difference of the human rater in comparison to the consensus truth. Similarly, *Max* is the most positive discrepancy to experts' volume. *Median* denotes the median position in the sorted deviations for each rater. *Average* indicates the absolute value of the average difference. Before chemotherapy, volumes differ in average on 10%. However, the maximal negative discrepancy is 30%, i.e. for one of the data sets, rater-4 assumed the tumor volume to be much smaller than the approximated ground truth volume. His median of -8% also coincides with his lower recall values. Similarly, rater-1 shows a maximal positive difference

Table 5.5: Relative difference (in %) in tumor volume of each expert in comparison to consensus truth. Rater-1: Radiologist, Rater-2: Physician, Rater-3: M.D. student, Rater-4: Radiologist, Rater-5: Oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4	Rater-5
Pre-Chemotherapy					
Average (abs)	15%	6%	9%	9%	10%
Min	2%	−18%	−18%	−30%	−24%
Max	53%	20%	43%	3%	2%
Median	10%	2%	2%	−8%	−7%
Post-Chemotherapy					
Average (abs)	16%	39%	16%	16%	12%
Min	−49%	−100%	−100%	−23%	−13%
Max	17%	292%	10%	60%	37%
Median	5%	7%	6%	−10%	−10%

of 53%, assuming one of the tumors to be much larger. After chemotherapy, the average difference of domain experts' volumes and the consensus truth's volume increase to 20% - ignoring rater-2, the discrepancy still increases to 15%.
Unfortunately, these differences might have a huge impact on treatment decisions. Let us assume, rater-1 and rater-4 decide about the follow up treatment of two patients with an intermediate risk tumor and an extension of these beyond the kidney (Stage 2) with a true tumor volume of 500ml. Rater-1 has a tendency for more generous tumor outlines - his approximated volume is in average 16% larger, resulting in a volume of $\approx$ 580ml and a clear incidence for a strong follow up treatment. In contrast, rater-4 tends to cautiously annotate tumor areas and might miss regions. In average, his volume will be smaller, and consecutive therapy will be reduced. This indicates, that an exact tumor volume determination is of crucial importance.

5.1.3 Summary

In this section, we analyzed human expert annotations of Wilms' tumors. It turned out, that tumor outlines of domain experts are not as reliable as assumed before. Especially after chemotherapy (the time point of decision for follow up treatment), inter-rater variability of human annotators increases dramatically. This allows us several conclusions: First, a human bias is remarkably present and single human annotators are not reliable. Second, tumor delineation becomes more challenging during the curse of the therapy.
Additionally, we investigated the important decision marker of volume differences between single domain experts' volumes and their approximated consensus truth. It turned out that the discrepancies can be massive and influence the final decisions of follow up treatment. We cannot conclude whether these effects result in more side-effects and maybe an unnecessary medical burden. However, any decision for follow up treatment should be aware of these massive differences depending on the human expert.

5.2 Evaluation of Clinical Practice

Tumor expansion after preoperative chemotherapy is an important metric used to decide about follow up treatment and an accurate determination of tumor volume is critical. Nowadays, it is daily clinical practice that radiologists estimate tumor volume by measuring the three axes of its extension and assuming the nephroblastoma to have an ellipsoid shape [66].
In the last section, we saw that annotations of human experts suffer from a large inter-rater variability. Now, we investigate the applicability of the assumption of an ellipsoid shape itself. The most important criterion we want to examine is:

- How accurately does an ellipsoid shape approximate the tumor volume?

In the following, we investigate the current gold standard in clinical practice and compare the approximated volume with the one given by the consensus truth of our domain experts.

5.2.1 Volume Variability

In daily clinical routine, it is gold standard to estimate the volume of Wilms' tumors by an approximated ellipsoid shape. This *clinical volume* is then used in therapy and treatment planning and computed as is computed as $\texttt{width} \times \texttt{height} \times \texttt{depth} \times 0.524$ [66]. Here, `width`, `height` and `depth` of tumor denote the maximal expansion of tumor tissue on MR images. Please note, that the volume of the largest ellipsoid that fits in a cuboid is $\pi/6 \approx 0.524$ times the cuboid volume.
In the last section, we showed that the consensus truth reflects the agreement of several human expert raters on the tumor outlines. Hence, we assume that the true tumor volume is found through this approximated ground truth. This allows us to compare human expert annotations and clinical volumes in terms of percental volume differences in relation to the ground truth volume before and after chemotherapy, respectively. All information about clinical volumes, used in treatment planning were confirmed by the reference radiologist of the SIOP studies.
In a first step, we compare the ground truth volume to the clinical volumes used in therapy planning of the patients included in our data set. It turns out that they differ before chemotherapy on average by 22.62 ± 16.12 %; see Fig. 5.1. After chemotherapy, when tumor outlines are more difficult to determine, this deviation increases to 35.07 ± 41.01 % from the ground truth volumes.
However, this information is not free of a human bias towards the specific radiologist deciding about tumor outlines. We highlighted in the last section the strong inter-rater variability in Wilms' tumor annotation. In order to guarantee for a fair comparison of the two ways of approximating the ground truth volume, we also computed the tumor extents for all three axes based on our consensus truth.
Figure 5.1 highlights the differences of the estimated volumes in comparison to the pixel-wise computer ground truth volume before chemotherapy: In average, the ellipsoid shape deviates by $11.57 \pm 10.1\%$ with a median of 8.6795% difference. Obviously, nephroblastoma typically does not fit into an ellipsoid shape - with a volume difference of up to 30%, the approximation is of low quality. After chemotherapy, the shape of Wilms' tumors

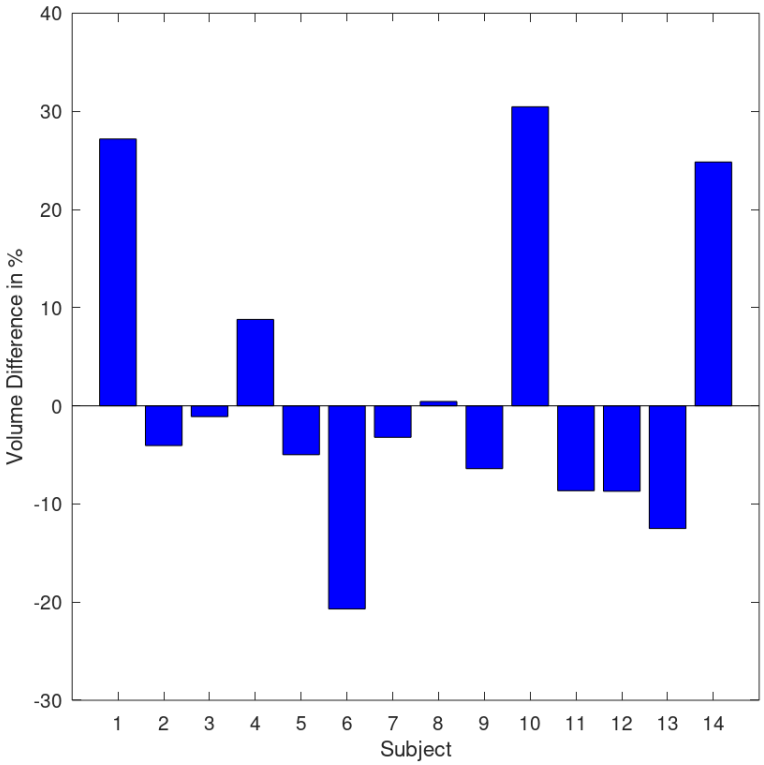


Figure 5.1: Comparison of consensus truth and ellipsoid volume before chemotherapy. Positive values indicate that the approximated ground truth volume is larger than the clinical volume.

moves even further away from the conformity with the assumption of an ellipsoid shape: The average difference in tumor volume increases to 20.46 ± 13 with a median of 17.895; see Fig. 5.2. In both cases, before and after chemotherapy, no clear tendency towards too large or too small volume estimations can be detected. Obviously, the assumption of an ellipsoid shape is an erroneous oversimplification.

5.2.2 Summary

The approximation of the tumor volume with an ellipsoid shape is current gold standard in treatment planning of Wilms' tumors. In this section, we investigated two situations. First, we analyzed the differences of the pixel-wise volume of our generated ground truth and the clinical volumes used in the treatment plan of the patients in our data sets. However, due to the large inter-rater variability, the data contains a bias towards the reference radiologist of the SIOP studies.
We therefore evaluated the deviations in a more strict setting and computed the volumes of the ellipsoid approximation based on our consensus truth. It turns out that even in this case, the approximation is erroneous and volumes of nephroblastoma should not be approximated in this way.

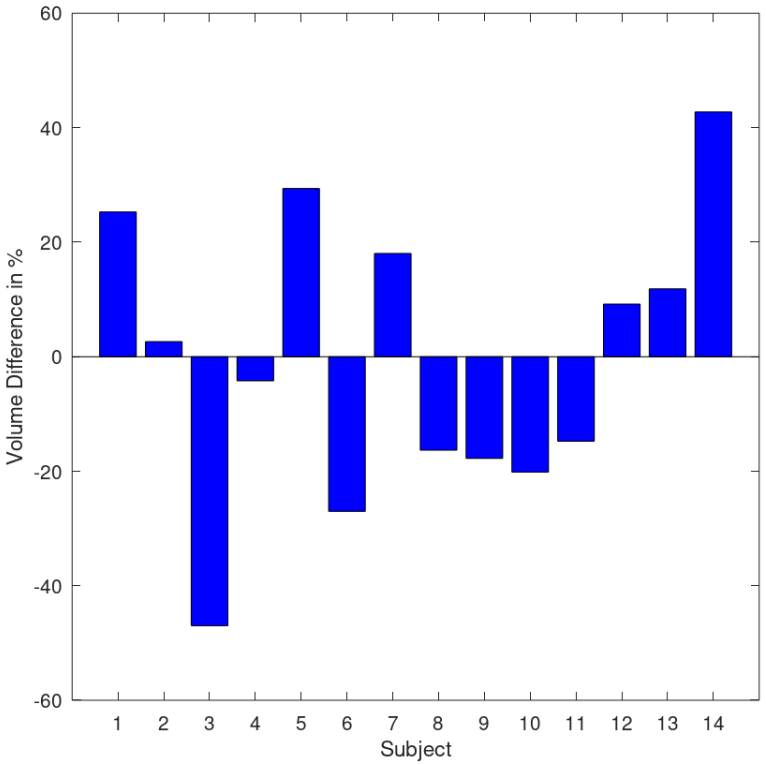


Figure 5.2: Comparison of consensus truth and ellipsoid volume after chemotherapy. Positive values indicate that the approximated ground truth volume is larger than the clinical volume.

5.3 Conclusions

In this chapter, we first investigated inter-rater variability of human expert annotations of Wilms' tumors. Unfortunately, we found that annotations of single domain experts are not reliable as previously assumed: While the differences between human raters are acceptable before chemotherapy, the situation changes after this milestone in therapy planning: inter-rater variability of human annotators increases dramatically after chemotherapy. We can observe that human bias increases and is remarkably present in their annotations in the later stages of treatment. In addition, tumor delineation becomes more challenging during the curse of the therapy.
Afterwards, we evaluated the differences in the determined volume of single raters and their consensus truth. It turned out that the deviations cannot be neglected and might influence the final decisions of follow up treatment.
Last but not least, we analyzed the approximation of the tumor volume with an ellipsoid shape. Unfortunately, it turns out that the gold standard of volume determination is erroneous and volumes of nephroblastoma should not be approximated by an ellipsoid shape.
We draw two conclusions from our analysis: First, no single domain expert should define tumor extensions after chemotherapy - the present variability even between radiologists is remarkable and can heavily influence treatment decisions. Second, nephroblastoma can-

not approximated by a simplistic shape. Its surface is complex and oversimplifications lead to large approximation errors. We believe therefore that a reliable and reproducible annotation of tumor outlines is essential for treatment planning.

6 Wilms' Tumor Segmentation

"Everything flows, and nothing abides, everything gives way, and nothing stays fixed."

– Heraclitus

Contents

The volume of a Wilms' tumor is an important decision criterion whether the follow up treatment after chemotherapy has to be more aggressive to ensure a full remission of tumor cells. The current gold standard of volume determination by approximation of an ellipsoid shape is prone to errors and the reproducibility is limited; see Sec. 5.2: Due to inter-rater variability and the oversimplification of tumors' shape, the differences between the determined and the real volume can be massive.
One obvious step to avoid at least the reproducibily problem is to replace human segmentations by automatic ones. Fully-automatic segmentation of Wilms' tumors is a challenging task as these tumors do not show a discriminative texture, might have intensities overlapping with the surrounding tissue, and can be directly attached to the remaining kidney. To the best of our knowledge, there is no method available so far that does not need massive user interaction. Moreover, the scientific literature on computer-based segmentation algorithms for Wilms' tumors is fairly limited. An initial idea for segmentation is to extend user marked seed points in the tumor by region growing based on intensity thresholding [46]. A refined approach is to initialize an active contour inside the tumor and to expand the segmentation according to image intensities and gradients [46]. In spite of the fact that segmentation is an active research field in image analysis for quite some decades, it is remarkable that many well-established classes of algorithms have not been evaluated in the context of Wilms' tumor segmentation. All in all, a high quality segmentation of nephroblastoma fulfills the following criteria:

- The segmentation forms clear clusters and does not show clutter.

- Pixels belonging to the same object are similar.

We close this gap and propose in Sec. 6.1 a flexible multi-label segmentation model with minimal user interaction that also reduces the variability of human experts by robustness to initialization. In spite of its ability to deal with uncertain seed points, it can be easily adopted to other segmentation scenarios. We demonstrate this capability on two completely different problems. Afterwards, we investigate the applicability and performance of several fully automatic methods to the problem of Wilms' tumor segmentation in Sec. 6.2.

6.1 Interactive Segmentation

Recent advances on convex relaxation methods allow for a flexible formulation of many interactive multi-label segmentation methods. The building blocks are a likelihood specified for each pixel and each label, and a penalty for the boundary length of each segment. While many sophisticated likelihood estimations based on various statistical measures have been investigated, the boundary length is usually measured in a metric induced by simple image gradients. A straightforward idea is to complement these methods with recent advances of edge detectors.

However, let us start with a basic interactive segmentation model. The "object" that is to be segmented in the image is seeded with labels. In the context of interactive segmentation these labels are often called scribbles and are set by a user. Nonetheless, labels can also be generated automatically by an algorithmic method. Please note that both scenarios are fundamentally different. While user scribbles are usually assumed to be correct, i.e., the segmentation method needs to fill in labels between the given ones, automatically generated labels can be erroneous and need to be corrected partially.

A successful idea is the estimation of statistical features from given scribbles in order to define likelihoods for each label, e.g. mean value [38], color/intensity histograms [25,59,127,150], or texture [6,128,153,154]. As a regularizer usually the boundary length in the image metric is minimized. Unger [175] penalized the boundary length in the metric induced by the image gradients, which aligns the segmentation boundaries to image edges.

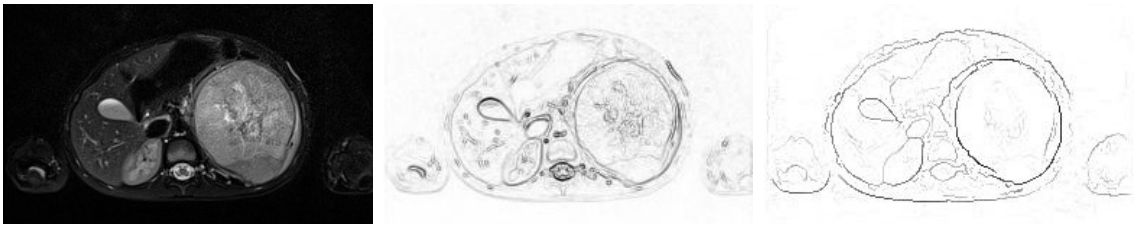
Werlberger et al. [192] measured the boundary length in a non-local metric, which achieves good results for small details, but suffers from expensive non-local computations. Of course, the image metric can be induced by more sophisticated edge indicators such as the traditional ones [32, 139], which are based on color and intensities, or more recent ones that include also texture information.

An intuitive idea is to induce the image metric by one of the state-of-the-art trained edge detectors by Dollár and Zitnick [51, 52]. Besides having better information about edges in the image, much fewer edges are detected compared to the simple gradient magnitude measure. This has also a positive effect on the computation time, as the propagation of labels is hampered less by unimportant structures.

A flexible framework for multi-label segmentation is the formulation as a minimal partitioning problem [63, 123], which can be solved via convex relaxation methods [34]. This framework, which is sometimes referred to as Potts multi-label segmentation model, is very flexible, since all the above data likelihoods can be incorporated. The boundary length is represented using the total variation of the label indicator functions, and is easily adjusted to a modification of the image metric. This convex relaxation framework is used for example in [127, 128, 130, 175, 192] for the task of interactive segmentation. In [165] it has been extended to a non-metric prior, in [164] to generalised ordering constraints, which constraints labels appearing adjacent to each other in a certain direction, in [50] to RGB-D data, and in [20] to the context of semantic segmentation.

6.1.1 A Multi-label Segmentation Model

First, let us discuss a generic multi-label segmentation model. Following Chambolle et al. [34], we consider a minimal partitioning problem of the rectangular image domain



(a) (b) (c)

Figure 6.1: Exemplary results for edge detection. (a) Input image, (b) Gradient magnitude image, (c) Result of structured edge detector. Edge maps are inverted and gamma corrected for visualisation. The structured edge detector shows more object edges and less clutter for unimportant structures.

$\Omega \subset \mathbb{R}^2$ into $\Omega_1, \ldots, \Omega_n \subset \mathbb{R}^2$ non-overlapping regions. The generic variational problem is

$$\begin{aligned} \min_{\Omega_1,\ldots,\Omega_n \subset \Omega} \; & \frac{1}{2} \sum_{i=1}^{n} \mathrm{Per}(\Omega_i; \Omega) + \sum_{i=1}^{n} \int_{\Omega_i} h_i(\boldsymbol{x}) \, d\boldsymbol{x} \,, \\ \text{s.t.} \quad & \Omega = \bigcup_{i=1}^{n} \Omega_i, \quad \Omega_i \cap \Omega_j = \emptyset, \quad \forall \; i \neq j \end{aligned} \tag{6.1}$$

where $h_i \colon \mathbb{R} \to \mathbb{R}_+$ are potential functions reflecting the cost for each pixel being assigned to a certain label $i = 1, \ldots, n$, and $\mathrm{Per}(\Omega_i; \Omega)$ denotes the perimeter of region Ω_i inside Ω. A weighting parameter $\lambda > 0$ is not required, as it can be absorbed in the functions h_i. In order to improve the alignment of region boundaries with image edges the perimeter is usually measured in a metric that is induced by the underlying image $\boldsymbol{f} \colon \Omega \to \mathbb{R}^d$. A common choice is a weighting with the image gradient with

$$\exp(-\gamma |\nabla \boldsymbol{f}(\boldsymbol{x})|), \tag{6.2}$$

where $\nabla \boldsymbol{f}$ denotes the Jacobian of $\boldsymbol{f}$ and $|\nabla \boldsymbol{f}|$ is its Frobenius norm. This reduces the measure of the boundary length where the image gradient magnitude is high. As Fig. 6.1(b) demonstrates, this choice is suboptimal when we seek for segmentations of objects. The image gradient magnitude shows clutter, i.e., it is high for unimportant edges. We favor the usage of a sophisticated edge detector instead and built on the fast structured edge detector [51,52]. In contrast to traditional edge detectors, it incorporates texture, color, and brightness. Figure 6.1(c) shows that this state-of-the-art edge detector is well-suited to identify also texture edges and illusory contours.

Let us now turn our attention to the potential functions in the second term of (6.1). Any method that proposes a new way to estimate these potential functions may be combined with the perimeter regularization discussed above. In the following, we recap the model proposed by Nieuwenhuis et al. [127]. However, in contrast to their formulation, ours is fully continuous, i.e.: the user input is assumed to be given on a (measurable) set instead of pixel positions.

Assume the user provides a (measurable) set of scribbles $\mathcal{S}_i \subset \Omega$ for each label i. Nieuwenhuis and Cremers [127] suggest to define the potential function $h_i(\boldsymbol{x})$ in (6.1) as the negative logarithm of the linearly to $[0, 1]$-scaled function $\tilde{h}(\boldsymbol{x})$ of

$$\frac{1}{|\mathcal{S}_i|} \int_{\mathcal{S}_i} k_{\rho_i(\boldsymbol{x})}(\boldsymbol{x} - \boldsymbol{y}) \; k_\sigma(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})) \, d\boldsymbol{y} \,, \tag{6.3}$$

where $|\mathcal{S}_i|$ denotes the area that is occupied by ith label, k_σ and k_{ρ_i} are Gaussians with standard deviation σ in color space and adaptive standard deviation $\rho_i(\boldsymbol{x}) = \alpha \inf_{\boldsymbol{y}\in\mathcal{S}_i} |\boldsymbol{x}-\boldsymbol{y}|$ in the spatial domain, respectively. Please note that instead of integrating over the set of scribbles they sum over all scribbled pixels. The idea of this spatially adaptive standard deviation is to reduce the influence of the color distribution from scribbles that are far away. This influence is reduced proportionally to distance from $\boldsymbol{x}$ to the closest scribble location.
A major drawback of this model is the assumption that all labels are correct. Formally, $h_i(\boldsymbol{x})$ must be set to $+\infty$ for $\boldsymbol{x} \in \mathcal{S}_i$, since (6.3) does not make sense for $\rho_i(\boldsymbol{x}) = 0$. Therefore, we adopt the potential functions to allow the segmentation method to correct possibly wrong scribbles/labels - this issue arises for instance when scribbles are provided by an uncertain human expert or are automatically set by an algorithm.
This is achieved by setting for scribble positions $\boldsymbol{x} \in \mathcal{S}_j$ the function values $\tilde{h}_i(\boldsymbol{x}) = 1-\zeta$ if $i = j$ and $\tilde{h}_i(\boldsymbol{x}) = \zeta/(n-1)$ otherwise, where $1-\zeta$ is the assumed probability for the scribble being correct.
To align image and region boundaries, the perimeter is commonly measured in a metric induced by the underlying image $\boldsymbol{f} : \Omega \to \mathbb{R}^3$. In this application, we weight the perimeter $\mathrm{Per}_g(\Omega_i;\Omega)$ of region boundaries in the metric

$$g(\boldsymbol{x}) = \exp(-\mathcal{E}(\boldsymbol{x})^\beta / \bar{\mathcal{E}}), \quad \bar{\mathcal{E}} := \frac{2}{|\Omega|}\int_\Omega |\mathcal{E}(\boldsymbol{x})|\, d\boldsymbol{x}.$$

Here $\mathcal{E}\colon \Omega \to \mathbb{R}$ is the output of the fast structured edge detector of [51, 52] and β is a positive parameter. Assume a (measurable) set of user-scribbles $\mathcal{S}_i \subset \Omega$ for each label i is given. We define the potential functions $h_i(\boldsymbol{x})$ in (6.1) as the negative logarithm of

$$\tilde{h}_i(\boldsymbol{x}) = \begin{cases} \left\{\frac{1}{|\mathcal{S}_i|}\int_{\mathcal{S}_i} G_\rho\, G_\sigma\, d\boldsymbol{y}\right\}_{\text{scale}}, & \boldsymbol{x} \notin \mathcal{S}_j, \\ 1-\zeta, & \boldsymbol{x} \in \mathcal{S}_j, i = j, \\ \zeta/(n-1), & \boldsymbol{x} \in \mathcal{S}_j, i \neq j, \end{cases} \tag{6.4}$$

and

$$G_\rho = k_{\rho_i(\boldsymbol{x})}(\boldsymbol{x}-\boldsymbol{y}),$$
$$G_\sigma = k_\sigma(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})).$$

Here $\{.\}_{\text{scale}}$ denotes linear rescaling to $[0,1]$, $|\mathcal{S}_i|$ is the area occupied by ith label, ζ is the assumed probability for a scribble being correct, and k_σ and k_{ρ_i} are Gaussians with standard deviation σ in intensity space and adaptive standard deviation

$$\rho_i(\boldsymbol{x}) = \alpha \inf_{\boldsymbol{y}\in\mathcal{S}_i} |\boldsymbol{x}-\boldsymbol{y}|$$

in the spatial domain, respectively. The spatially adaptive standard deviation attenuates the influence of the intensity distribution from scribbles that are far away proportionally to the distance of $\boldsymbol{x}$ to the closest scribble location.

6.1.2 Convex Optimization

The second term of the variational minimization problem (6.1) is obviously non-convex. The regions $\Omega_1, \ldots, \Omega_n$, their non-overlapping and covering criterion can be easily represented by label indicator functions $\theta \in BV(\Omega, \{0,1\}^n)$ where

$$\theta_i = \begin{cases} 1, & \boldsymbol{x} \in \Omega_i \\ 0, & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n \tag{6.5}$$

and BV is the bounded variation, i.e. functions with finite total variation (TV) [127]. In this representation, the perimeter of the regions is measured by the weighted total variation.

We follow Nieuwenhuis and Cremers [127] and rewrite the set indicated by θ_i by means of its distributional derivative $D\theta_i = \nabla\theta_i \; d\boldsymbol{x}$. Let $\boldsymbol{y}_i \in C_c^1(\Omega, \mathbb{R}^2)$ be the dual variables where C_c^1 is the space of smooth functions with compact support. According to [127], we can now apply the coarea formula [57] such that the weighted total variation is identical to the weighted perimeter of Ω_i:

$$\frac{1}{2}\mathrm{Per}_g = \frac{1}{2}\mathrm{Per}_g(\{\boldsymbol{x} |\ \theta_i(\boldsymbol{x}) = 1\}) \tag{6.6}$$

$$= \frac{1}{2}\mathrm{TV}_g(\theta_i) \tag{6.7}$$

$$= \frac{1}{2}\int_\Omega g \|D\theta_i\| \tag{6.8}$$

$$= \sup_{\boldsymbol{y} \in \mathcal{K}_g} \int_\Omega \boldsymbol{y}_i D\theta_i. \tag{6.9}$$

With integration by parts and due to its compact support, we can rewrite (6.9) as

$$\sup_{\boldsymbol{y} \in \mathcal{K}_g} \int_\Omega \boldsymbol{y}_i D\theta_i = \sup_{\boldsymbol{y} \in \mathcal{K}_g} \left(- \int_\Omega \theta_i \ \mathrm{div}\ \boldsymbol{y}_i \ d\boldsymbol{x} \right). \tag{6.10}$$

where

$$\mathcal{K}_g = \left\{ \boldsymbol{y}_i \in C_c^1(\Omega, \mathbb{R}^2) |\ \|\boldsymbol{y}_i(\boldsymbol{x})\| \leq \frac{g(\boldsymbol{x})}{2}, \boldsymbol{x} \in \Omega \right\}. \tag{6.11}$$

Hence, the general minimal partition problem of (6.1) with our potential functions is equivalent to

$$\min_{\theta \in B} \sup_{\boldsymbol{y} \in \mathcal{K}_g} \{ - \int_\Omega \theta_i \ \mathrm{div}\ \boldsymbol{y}_i \ d\boldsymbol{x} + \sum_{i=1}^n \int_{\Omega_i} \theta_i \tilde{h}_i(\boldsymbol{x}) \ d\boldsymbol{x} \}, \tag{6.12}$$

$$B = \{ \theta \in BV(\Omega, \{0,1\}^n) | \sum_{i=1}^n \theta_i = 1 \}. \tag{6.13}$$

Obviously, (6.12) is not convex and cannot be globally optimized. Following [127], we relax the ranges of B to

$$\tilde{B} = \{ \theta \in BV(\Omega, [0,1]^n) | \sum_{i=1}^n \theta_i = 1 \}. \tag{6.14}$$

The application of the primal–dual algorithm in [35] is now straightforward. The resulting update equations are given in Alg. 4; see Sec. 2.3. Here, the primal variables $\boldsymbol{x}_i$ are projected onto the simplex and the dual variables $\boldsymbol{y}_i$ onto $\mathcal{K}_g$.

Algorithm 4 Algorithm to solve for spatially varying color distributions (3.3) and (3.4)

Initialization: $\tau, \sigma > 0$, $\boldsymbol{x}^0 \in \mathcal{X}$, $\boldsymbol{y}^0 \in \mathcal{Y}$, and $\bar{\boldsymbol{x}}^0 = \boldsymbol{x}$
for $(t = 0;\ \ t < T;\ \ t{+}{+})$ **do**
$\boldsymbol{y}_i^{t+1} = \Pi_{\mathcal{K}_g}(\boldsymbol{y}_i^t + \sigma \nabla \bar{\boldsymbol{x}}_i^t)$
$\boldsymbol{x}_i^{t+1} = \Pi_{\tilde{B}}(\boldsymbol{x}_i^t - \tau(\text{div } \boldsymbol{y}^{t+1} - f_i))$
$\bar{\boldsymbol{x}}_i^{t+1} = 2\boldsymbol{x}_i^{t+1} - \boldsymbol{x}_i^t$

6.1.3 Additional Applications

In this section, we demonstrate the flexibility of our interactive segmentation method and show that it is not only applicable to medical images but also performs well on color and video data. We evaluated our approach on the GRAZ benchmark [153] as well as the FBMS-59 [131] data sets. We show results in terms of the metrics suggested in Sec. 2.4.

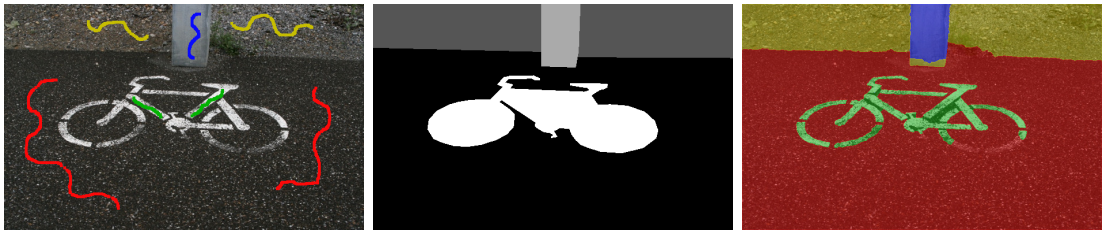
Segmentation of Color Images

The GRAZ benchmark dataset [153] consists of 262 seed-ground-truth pairs from 158 natural images for interactive multilabel segmentation. We use the following manually tuned parameters for experiments with space-variant color distributions: $\alpha = 15$, $\beta = 2$, $\sigma = 3$, and $\zeta = 0.05$.
We compare our results in Tab. 6.1 to the original approach by Nieuwenhuis et al. [127], as well as their advanced method that incorporates texture information [128]. The results indicate that using an advanced edge detector textural information in the data term can be neglected. Fig. 6.3 shows an exemplary result from our evaluation. The information from an advanced edge detection is sufficient for segmentations of high quality. Decreasing the diameter of user scribbles leads to slightly worse approximations of the

Table 6.1: Comparison of our approach to spatially variant approaches by Nieuwenhuis et al. [127, 128].

Method	Dim	Dice Score
Nieuwenhuis/Cremers, spatially constant [127]	3	0.89
Nieuwenhuis/Cremers, space-variant [127]	5	0.92
Nieuwenhuis/Cremers, space-variant [127]	13	0.93
Nieuwenhuis et al., space-variant + texture [128]	13	0.94
Our approach, spatially constant, no color	5	0.77
Our approach, spatially constant, color	5	0.90
Our approach, space-variant, no color	5	0.80
Our approach, space-variant, color	5	0.93
Our approach, space-variant, color	13	0.94

color distributions. However, our model can compensate this lack of information since textural and color information are also included in the edge detector.



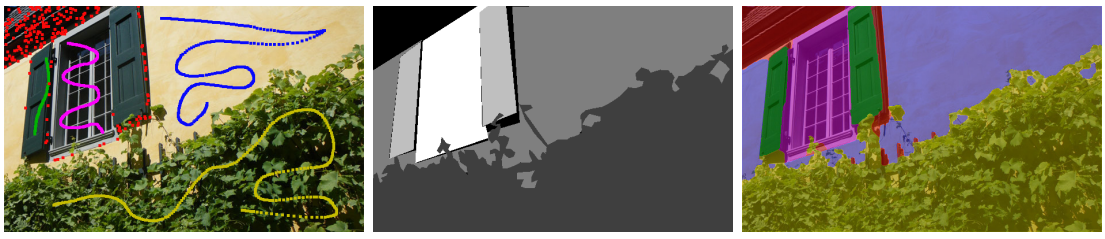
(a) (b) (c)

Figure 6.2: Limitations of GRAZ benchmark. (a) Scribble image, (b) ground truth label image, (c) our result. It is clearly visible, that the quality of image labels is limited and the segmentation outcome can not reflect the user intention. Dice score: 0.86

In fact, the texture based model [128] seems to rely on a large diameter.
The creators of the GRAZ benchmark made an important contribution towards evaluation of interactive multi-label segmentation, but current state-of-the-art segmentation reached the limit of these data sets. Small details are very important, and not all ground truth labellings are optimal, see Fig. 6.2. Overall, the results on this data set are very close to the optimum. Therefore it might not be desirable to further improve the Dice score on this data set.

Video Segmentation

The FBMS-59 [131] data set, an extended version of [30], contains 29 video sequences for training and 30 video sequences for testing. We used minimum cost multicuts [90] as well as [131] to generate and automatically label point trajectories. Ideally, they provide sparse, and temporally consistent labels for each frame in a video. In contrast to interactive or supervised segmentation, trajectory labels are single pixels spread over the whole image domain and tend to be erroneous. Video segmentation is a challenging task. Even the considered state-of-the-art methods [90, 131] provide erroneous trajectories. Therefore, methods that turn the sparse labels into dense segmentations, such as our method, must be able to correct some of the wrong labels. We incorporate this prior knowledge by increasing the uncertainty parameter ζ. The manually tuned parameters we used for all video frames are $\alpha = 30$, $\beta = 2$, $\sigma = 3$, and $\zeta = 0.2$.
We compare our approach to the recent approach [131] that densifies the sparse labels from the trajectories in each frame based on image gradients. We stick to point trajectories generated by [131] for the sake of comparability. Tab. 6.2 states our benchmark results



(a) (b) (c)

Figure 6.3: Exemplary segmentation result exclusively based on color variation and the structured edge detector. (a) Scribble image, dim $= 5$, (b) Ground truth labelling, (c) Our result.

Table 6.2: Results on the Video Segmentation Benchmark FBMS-59.

Method	Density	Dice	Precision	Recall	F$\geq$ 75%
Training set					
Ochs et al., MoSegDense [131]	100%	0.69	0.84	0.59	15/65
Ochs et al., MoSegSparse [131]	0.87%	0.72	0.85	0.62	17/65
Our approach, SPT-C, NC (SC [131])	100%	0.79	0.83	0.75	18/65
Our approach, SPT-V, C (SC [131])	100%	0.81	0.84	0.78	20/65
Keuper, MCe sparse, prior 0.5 (MT [90])	0.86%	0.79	**0.87**	0.73	31/65
Our approach, SPT-V, C (MT [90])	100%	**0.82**	0.85	**0.80**	24/65
Test set					
Ochs et al., MoSegDense [131]	100%	0.66	0.78	0.57	17/69
Ochs et al., MoSegSparse [131]	0.92%	0.69	0.80	0.61	24/65
Our approach, SPT-C, NC (SC [131])	100%	0.71	0.75	0.68	18/65
Our approach, SPT-V, C (SC [131])	100%	0.74	0.76	**0.72**	21/69
Keuper, MCe sparse, prior 0.5 (MT [90])	0.87%	**0.76**	**0.88**	0.68	25/69
Our approach, SPT-V, C (MT [90])	100%	0.75	0.81	0.71	23/69

for two variants of our model: In the first version, we include exclusively information about label position and rely on information already contained in our edge detector (SPT-C, NC). This version coincides with [131] except for the edge detection. In the second variant, we include all available information for segmentation and use spatially variant color distributions (SPT-V, C). Though the edge detector should already contain all relevant color and texture information in the image, favoring slightly color similarity improves the results. This is due to a suboptimal performance of the edge detector.We clearly outperform [131] on all error metrics.

Moreover, we complement the sparse labels of the state-of-the-art in motion segmentation from Keuper et al. [90] with our approach. The ability to correct also erroneous labels (see Fig. 6.4) allows us to even improve the results in the dense segmentation. Note that the difference in performance on test and training set is due to different challenges and the (still limited) number of video sequences rather than over-fitting.

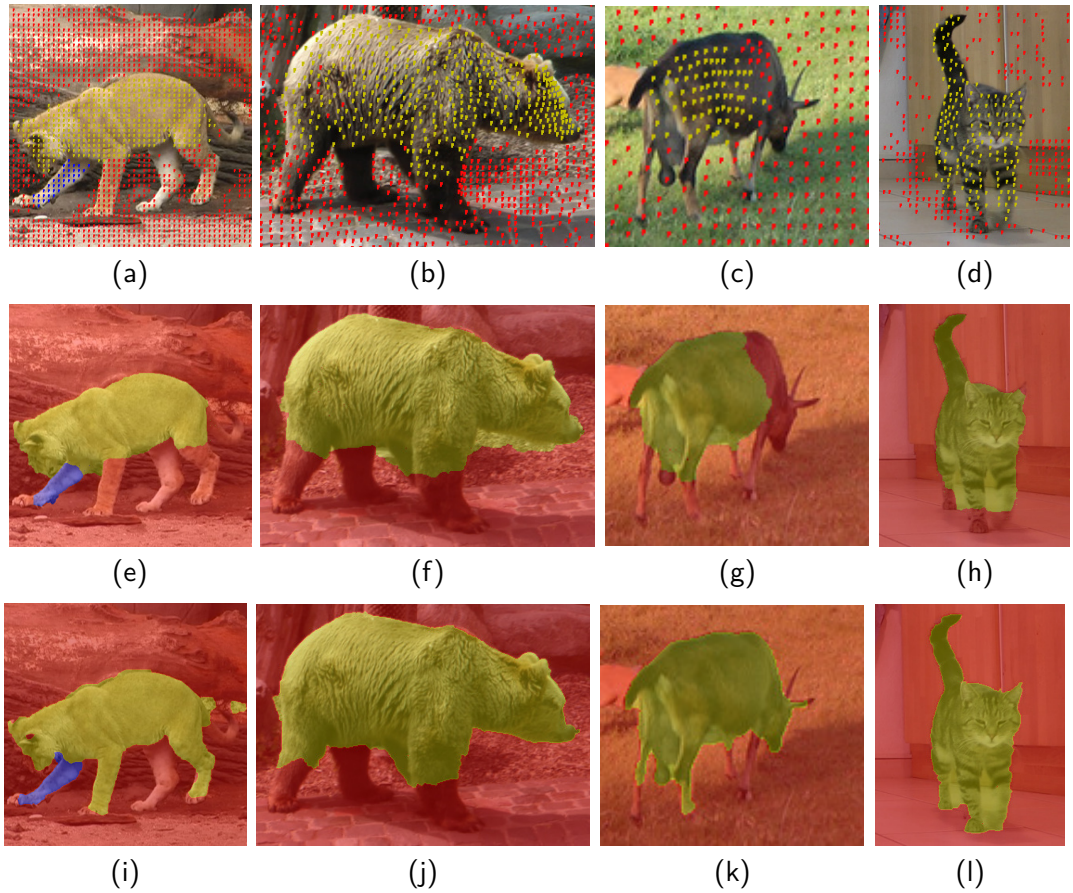


Figure 6.4: Exemplary results on FMBS-59 [131] data sets. ((a) - (d)) Labels from SC-point trajectories [131]. ((e) - (h)) Segmentation results of Ochs et al., MoSegDense [131]. ((i) - (l)) Our segmentation result, no color, spatially constant. Trajectory labels are enhanced for visualization. Our method is able to correct a significant amount of labels.

6.1.4 Summary

We illustrated a robust and flexible algorithm for interactive multi-label segmentation based on a sophisticated edge detector that includes texture, colour, and brightness. A remarkable feature of our method is the ability to correct some erroneous labels. In addition to our main scenario of labels set by an user, we also applied our approach to labels generated by an algorithm and even more likely to be erroneous. In all these settings, we could show its ability to produce segmentation results of high quality. Next, we evaluate our semi-automatic as well as fully automatic methods on the task of Wilms' tumors segmentation.

6.2 Evaluation of Segmentation Algorithms

One of the major drawbacks of human expert annotations is their lack of reproducibility while a strong human bias is present. In order to address this problem, a comprehensive evaluation of computer-based segmentation methods is essential.

6.2.1 Experiments

In the following, we conduct example evaluations on our benchmark data set for Wilms' tumor segmentation (Sec. 4.3) with six fully-automatic and our semi-automatic method:

- Chan-Vese active contours [36] with two level sets.
- K-means clustering [100] with intensities.
- Entropy Rate Superpixel Segmentation [99].
- Classification with a support vector machine [24] with intensities and HOG-features [45]; see Sec. 3.2.3.
- Random-forest classification [27], either with intensities or HOG-features [45].
- Segmentation with a U-Net [149].
- Our semi-automatic segmentation method; see Sec. 6.1.

To guarantee a fair evaluation, we equally split the data sets in training and test data, each containing seven data sets before and after chemotherapy. For each segmentation approach we include information from all modalities. Since the sampling rate in depth direction is substantially lower than in the other directions, we prefer to restrict ourselves to 2D segmentations we refrain from 3D segmentations as the interpolation error would be too high. Let us now sketch each of the evaluated segmentation approaches.

Chan-Vese Active Contours

We consider a cubic data domain $\Omega \subset \mathbb{R}^3$ and a volumetric data set $\boldsymbol{f} : \Omega \to \mathbb{R}^3$. In our setting, the co-domain describes the different MRI modalities T2, T1, and T1c. Then a segmentation of $\boldsymbol{f}$ by means of the Chan-Vese active contour model [36] minimizes the cost function

$$\begin{aligned} E(\boldsymbol{u}, C) = \lambda_{\mathtt{in}} \int_{C_{\mathtt{in}}} \|\boldsymbol{u}_{\mathtt{in}} - \boldsymbol{f}\|^2 d\boldsymbol{x} \\ + \lambda_{\mathtt{out}} \int_{C_{\mathtt{out}}} \|\boldsymbol{u}_{\mathtt{out}} - \boldsymbol{f}\|^2 d\boldsymbol{x} + \nu\, \ell(C) \end{aligned} \tag{6.15}$$

where the data domain Ω is split in two regions C_{in} and C_{out}. The function $\boldsymbol{f}$ is approximated by a piecewise constant function where $\boldsymbol{u}_{\mathtt{in}}$ and $\boldsymbol{u}_{\mathtt{out}}$ are the arithmetic means of $\boldsymbol{f}$ inside and outside the segment boundaries C, respectively. The positive weights $\lambda_{\mathtt{in}}$ and $\lambda_{\mathtt{out}}$ control the influence of each region to the final partitioning, $\|\,.\,\|$ denotes the Euclidean norm in $\mathbb{R}^3$, and C are the segment boundaries with a (Hausdorff) length of $\ell(C)$. This length is weighted with a parameter $\nu > 0$.

K-means Clustering

K-means clustering [100] is a vector quantization method that partitions n observations into k clusters. Data points are assigned to cluster centers, prototypes of corresponding classes, with minimal Euclidean distance. In our application, we want to split the observations into two classes, tumor and non-tumor points.
Given a set of data points $\boldsymbol{f} : \Omega \to D$ with $D \subset \mathbb{R}^3$ and $\Omega \subset \mathbb{R}^3$, k-means minimizes

$$\begin{aligned} E(D_1, D_2) &= \int_{D_1} \|\xi - \boldsymbol{u_1}\|^2 d\xi + \int_{D_2} \|\xi - \boldsymbol{u_2}\|^2 d\xi \\ D &= D_1 \cup D_2, \quad D_1 \cap D_2 = \emptyset, \end{aligned} \tag{6.16}$$

where $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$ are the arithmetic means of both classes. In this case, k-means clustering is equivalent to Otsu's method [98].

Support Vector Machine

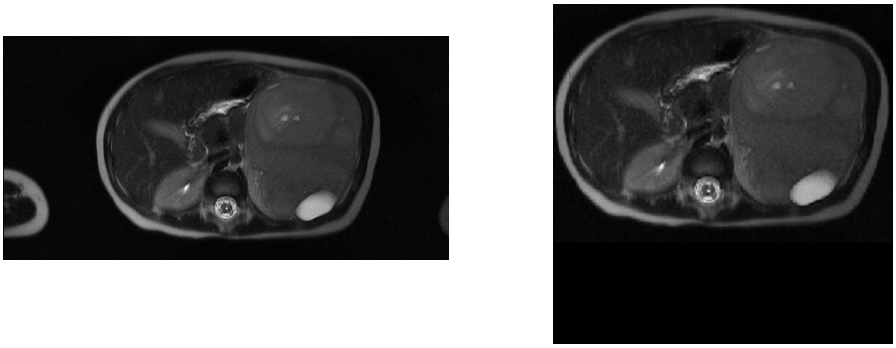
Support Vector Machines [24] are based on the concept of hyperplanes in a multidimensional space, separating between sets of objects having different classes, e.g. tumor and non-tumor points. In our application, we use a five-fold cross validation to find optimized hyperparameters. Training was performed using MATLAB (`www.mathworks.com/products/matlab`) and the problem was solved via Sequential Minimal Optimization [56]. Furthermore, we used Gaussian-like kernels and the classification error, i.e. the weighted fraction of misclassified observations, as loss function.

Random-forest Classification

Ensemble methods employ a finite set of different learning algorithms to get better predictive performance than using a single learning algorithm. Random forests [27] are ensemble approaches for classification combining a group of decision trees. A single tree is highly sensitive to noise, while the average of many decorrelated trees is not. Training all decision trees of a random forest on the same training data would result in strongly correlated trees. Bagging (bootstrap aggregation) generates new training sets $\mathbf{K}$ by sampling from the original training set $\mathbf{Y}$ uniformly and with replacement. In this way, decision trees are decorrelated by using different training data. Additionally, random forests use feature bagging, i.e. features are randomly sampled for each decision tree [77]. To estimate how well the results can be generalized, we use 2-fold-cross validation, i.e. we train two sets of models.

Entropy Rate Superpixel Segmentation

The method of Liu et al. [99] formulates the superpixel segmentation problem as maximization of the entropy rate of cuts in the graph. Optimizing this entropy rate encourages the clustering of compact and homogeneous regions, which also favors the superpixels to overlap with only one single object on the perceptual boundaries.
This technique starts with each pixel being considered as a separate cluster. Clusters are then gradually merged into larger superpixels. In this way, during segmentation, a



(a) (b)

Figure 6.5: Exemplary pre-processing step for the U-Net. (a) Original image containing abdomen and extremities, (b) Image after pre-processing.

hierarchy of superpixels is created until finally only one superpixel, the image itself, is left. In our case we want to segment a tumor, i.e. we use the hierarchy of superpixels to divide the image into three groups: tumor, body and background. Unfortunately we do not know in advance which superpixel contains which class. This objective function is optimized with a greedy algorithm.

UNet

In many areas of medical image processing, deep learning and especially convolutional neural networks (CNN) have proven to be very powerful tools. Within these, the UNet architecture [149] is one of the standard CNNs in the field of medical image segmentation. It learns segmentation in an end-to-end setting and only needs a few training examples in its original formulation; we refer to Sec. 7.1.2 for a detailed explanation.
Since our benchmark consists of real clinical data, they are available in different resolutions. Some of them also contain other parts of the body, e.g. the arms. Therefore, the amount of non-tumor areas outweighs the tumor areas substantially, such that it becomes necessary to balance the classes. This is done in three steps: First we determine the connected components, i.e. connected parts of the body, and remove everything except the largest one. Then we determine the maximum extent of the existing object and extract this part to a new, smaller image; see Fig. 6.5. This is then rescaled to a size of 512×512 pixels. We use the implementation presented in [2] to solve our segmentation problem and set up the network with batch size 5 and 50 epochs.

6.2.2 Results

In Tab. 6.3 we present the mean precision, recall and Dice score over the 14 test data sets of the different segmentation algorithms. Since the Chan-Vese method is region-based, it suffers from the fact that the visual appearance of Wilms' tumors can be highly heterogeneous. Our experiments show that intensities are an important feature to identify tumor areas, resulting in high precision values for the pixel-based classifiers k-means clustering and random forests. However, spatial information is essential as intensities of a tumor can overlap with those of the surrounding tissue. Accordingly, the pixel-based methods

Table 6.3: Results on the proposed benchmark data set (test data). k-means: k-means clustering, CV: Chan-Vese active contours, RF: Random Forest Classification, SVM: Support Vector Machine, INT: Intensity values, HOG: HOG-features, PP: Post-processing. Best results are depicted in bold face.

Method	Dice Score	Precision	Recall
Pre-Chemotherapy			
CV [36]	0.57	0.48	0.69
k-means [100] (INT)	0.53	0.76	0.41
Superpixel [99]	0.41	0.33	0.56
SVM [24] (INT + HOG [45])	0.71	0.71	0.72
RF [27] (INT + HOG [45])	**0.92**	**0.92**	0.91
U-net [149]	0.64	0.49	**0.94**
our approach	0.88	0.88	0.87
Post-Chemotherapy			
CV [36]	0.41	0.32	0.58
k-means [100] (INT)	0.35	0.50	0.27
Superpixel [99]	0.41	0.29	0.68
SVM [24] (INT + HOG [45])	0.68	0.69	0.67
RF [27] (INT + HOG [45])	0.81	0.73	**0.92**
U-net [149]	0.30	0.25	0.61
our approach	**0.84**	**0.80**	0.89

suffer from low recall. Using HOG-features in addition to intensities improves k-means clustering after chemotherapy, SVM classification as well as random forests both before and after chemotherapy.

The results of the superpixel-based method are unexpectedly poor both before and after chemotherapy. The optimum number of superpixels depends strongly on the image and it is also difficult to identify the respective segments. We could not find a parameter set that worked on all data sets.

Deep learning methods usually require a large amount of training data. The U-net used here deviates from this paradigm and can also be trained with smaller amounts of data. Tab. 6.3 shows that it gives a high mean recall, but a low mean precision. This indicates that although the network can recognize the basic structure of the nephroblastoma, it is not able to distinguish it from similar tissue.

Therefore, we suggest random forests trained on HOG-features as well as intensities as the baseline method for this benchmark data set. In order to ensure spatial consistency, we also apply Chan-Vese active contours on the predicted probabilities of the random forest. It turns out that predictions of this method lack too much global information and the resulting segmentation loses quality.

Our approach differs from the before mentioned fully-automatic segmentation approaches

by the fact that it is semi-automatic: A clinician who was not involved in the manual segmentations (see Chap. 5) and who is familiar with tumors, drew user scribbles in a single depth slice for each T_2 data set as initialization. We observe that this approach shows high-quality results before and after chemotherapy. Segmentation quality of random forests as well as semi-automatic segmentation lies within the variability of human experts; see Chap. 5. These observations highlight the challenges in the data set.

6.2.3 Summary

We evaluated seven computer-based algorithms. At this time, fully-automatic segmentations undersegment the tumor volume compared to human expert raters and the quality is insufficient, especially after chemotherapy.
Our experiments indicate that semi-automatic segmentation with our approach is an appropriate tool for segmentation of Wilms' tumors. Its results lie within the variability of the contouring performed by human expert raters on the same data. Moreover, it offers the advantage that specifying scribbles is much faster than a full segmentation by human experts.

6.3 Conclusions

We illustrated our robust and flexible interactive segmentation method for Wilms' tumor segmentation. We evaluated it together with a wide range of fully-automatic segmentation methods on our benchmark data set; see Sec. 4.3. It turns out, that fully automatic approaches oversegment the tumor and are therefore not as suited as our method for this kind of segmentation task. In the remainder of this thesis, we use our approach as a first step for further processing of MR images of Wilms' tumors.

7 Generalization of Deep Neural Networks

"Fast is fine, but accuracy is everything."

– Xenophon

Contents

Since AlexNet [93] won the "ImageNet Large Scale Visual Recognition Competition" challenge [48], the influence of deep neural networks has increased dramatically in all domains of image processing and pattern recognition: From classification, to object tracking and image and video segmentation, new approaches are typically based on deep learning strategies [11, 42, 80]. These approaches are also gaining more and more influence in the field of medical image processing. Since Ronneberger et al. proposed the UNet structure [149], this model is de facto the standard method in the field of medical image segmentation. While the original approach could be trained with relatively few examples in a short time, current models require large amounts of data with a very time consuming and computationally intense training cycle [82, 124]. Since several years it is common practice to compare the performance of segmentation approaches on benchmark data sets. One of the best known of these data sets is provided within the "Multimodal Brain Tumor Segmentation Challenge" (BraTS) [115].
Brain tumors account only for a very small fraction of all types of cancer, but are also among the most fatal forms of this deadly disease. Gliomas, developing from the glial cells, are the most frequent primary brain tumors. The fast growing and more aggressive types of gliomas called high-grade gliomas, come with a median overall survival rate

up to 15 months [111]. The standard diagnosis technique for brain tumor is magnetic resonance imaging (MRI) [191] providing detailed information about the tumor and the surrounding brain. Tumor segmentation is a crucial task in surgical and treatment planning. The clinicians' standard technique is still manual tumor segmentation, which tends to inter- and intra-rater variability [113]. Moreover, the time required to manually annotate and segment the data is high. Therefore, much research is performed to develop methods for automatic brain tumor segmentation; see [11] and the references therein. Fully automated segmentation is a challenging task, especially for high-grade gliomas as they usually show diffuse and irregular boundaries and have intensities overlapping with normal brain tissue. Moreover, acquisition parameters are not standardized, and different parameter settings can have a substantial impact on the visual appearance of the tumor. This makes it difficult to compare the quality of different methods for brain tumor segmentation. As a step towards an unbiased performance evaluation, BraTS database has been created [10,11,115], and many recent approaches report benchmark results on either the full data set or parts of it; see e.g. [82,124,176].
Since this data set has been used for seven years now to compare different approaches with each other, a major drawback has manifested itself over this long time: The main focus of the researchers is not to present the most robust network with best generalization behavior, but to maximize the performance metrics of the BraTS benchmark dataset. We are deeply convinced that the increasingly complicated models are not useful in a real clinical scenario as they heavily overfit the test set: this benchmark data set is saturated. We strongly believe that models do not get better in a general sense but current best approaches overfit the test set more than others. Typically, a test set is meant to be a biased version of a specific problem representation, i.e. all humans with high grade brain tumors in MRI sequences. In order to show a statistical significance of one benchmark result being superior to another one, an appropriate sample size is necessary [41]. Unfortunately, the sample size of the BraTS test set is too small to provide a statistical significant difference of the best performing methods [11].
In addition, the main strength of deep learning approaches of fitting the underlying data distribution is also their greatest weakness: in a clinical setting, the assumption that training and test data belong to the exact same distribution is typically not correct.

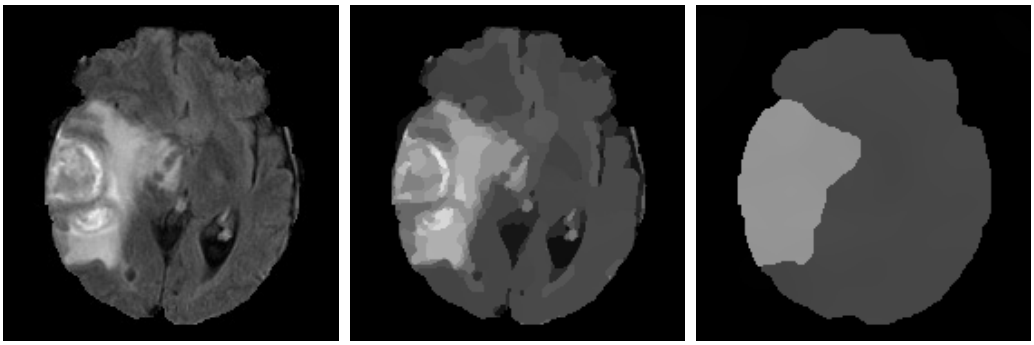
In this chapter, we evaluate the robustness of different segmentation methods with respect to disturbances in the underlying distribution. We investigate three state-of-the-art methods as well as a simple and intuitive scheme based on the powerful Mumford-Shah functional.
In addition, we suggest two simple and straight forward modifications that allow to increase the generalization performance of the evaluated deep neural networks. Finally, we demonstrate that our semi-supervised segmentation approach is a powerful post-processing step, that allows to robustify the predictions of deep neural networks with respect to disturbances in the test data set. Hence, we combine the best of two worlds: We still can learn the class distributions of the targeted objects while exploiting the robustness of energy formulations to modifications in the data.

7.1 Segmentation Approaches

The baseline of our evaluation is a segmentation approach that does not require training: a cascadic Mumford-Shah cartoon model. It was published before deep learning models start to dominate the field of medical image segmentation. Since we can almost eliminate an overfit to the underlying data set, none of the compared deep learning models should score below the performance of this method.
We begin our evaluation with the de-facto standard model for image segmentation with deep neural networks: the UNet architecture [149]. We continue with an extended version, namely the No NewNet approach showing high performance on BraTS 2018 [82]. Afterwards, we take the winner of last years' challenge into account: NVDLMED, using autoencoder regularization to improve the segmentation accuracy [124]. Last but not least, we investigate the third place of the BraTS 2018 challenge [200].
Please note that a full introduction to deep learning is beyond this thesis. We therefore refer the interested reader to the excellent introduction to deep learning by Goodfellow et al. [65].

7.1.1 Cascadic Mumford-Shah Cartoon Model

Brain tumor segmentation methods can be roughly categorized into two groups: semi-automatic and fully-automatic segmentation. Most of the classical fully-automatic segmentation approaches include classification [171] or clustering algorithms or rely on atlas-based methods. The majority of semi-automatic segmentation methods are based on active contour models. Active contours are often modeled implicitly as level-sets and can be further divided in edge- and region-based methods. Edge-based active contours heavily depend on the image gradient. Hence, they are less suited for segmenting edematous regions of high-grade gliomas. However, region-based active contours are more robust when parts of the object boundaries are diffuse, since they reward homogeneity within the segment.
The most popular approach for region-based active contours is the Chan-Vese model [38]. In its basic formulation, it segments an image into fore- and background. The granularity of this segmentation is steered by a single parameter that weights the length of the segment boundaries. The global behavior of this segmentation model is both a blessing and a curse - a blessing because it is very robust to noise and initialization; a curse because it is prone to false segments and intensity inhomogeneities. The Chan-Vese model can be seen as a simplification of the cartoon limit of the Mumford-Shah functional [122, 123]. Since Strekalovskiy et al. [166] proposed an efficient primal–dual algorithm to minimize this functional, optimizing this approach is computational feasible, too.
However, the crux in tumor segmentation approaches with this functional is to find an appropriate parameter setting and to identify segments with tumor tissue. If the steering parameter is too small, the result contains many small segments such that it is hard to determine the appropriate ones. In contrast, if the parameter is too large, the segmentation is too coarse, and the segmentation does not approximate the tumor boundaries very well.
The strategy of our cascadic approach is as follows: At the beginning, we segment the complete tumor. We tackle the difficulty of finding an appropriate parameter setting by solving the Mumford-Shah cartoon model iteratively with varying boundary weights



(a) (b) (c)

Figure 7.1: Exemplary results for different penalizations of the boundary length. **(a)** T_2-Flair input image. **(b)** Result for $\nu = 1000$. **(c)** Result for $\nu = 340000$.

until we reach a segmentation with only few segments. We refine these tumor boundaries afterwards by introducing a new confidence measure that increases the precision significantly.

Preprocessing Since MRI scans may suffer from non-uniformities within each data set, we first apply the N4ITK filter [170] to all scans to correct for these artifacts. Then we reduce the influence of the background on the segmentation outcome by replacing all gray values smaller than the average gray value μ with μ. Finally, we compute and equalize the histograms of each 3D scan to exploit the contrast in the data in the best possible way. All data channels are rescaled to intensity values in the interval $[0, 255]$.

The Mumford-Shah Cartoon Model Let us consider a cubic data domain $\Omega \subset \mathbb{R}^3$ and some volumetric data set $\boldsymbol{f} : \Omega \to \mathbb{R}^m$. For our application, its m channels describe different MRI modalities such as T_1, T_{1c}, T_2 and T_2-Flair. Then a segmentation of $\boldsymbol{f}$ by means of the Mumford–Shah cartoon model [122, 123] minimizes the energy functional

$$E(\boldsymbol{u}, C) = \sum_i \int_{\Omega_i} \|\boldsymbol{u} - \boldsymbol{f}\|^2 d\boldsymbol{x} + \nu\, \ell(C). \tag{7.1}$$

Here the a priori unknown number of segments Ω_i partition the data domain Ω, the function $\boldsymbol{u}$ denotes a piecewise constant approximation of $\boldsymbol{f}$, $\|\,.\,\|$ is the Euclidean norm in $\mathbb{R}^m$, and the segment boundaries C have a (Hausdorff) length of $\ell(C)$. The first term of the energy is a data term that penalizes fluctuations within each segment, while the second term favors short segment boundaries. The parameter $\nu > 0$ allows to weight the boundary length in relation to the inhomogeneities within each segment. Obviously the choice of ν is of crucial importance: The higher the value of this parameter, the less segments are contained in the final result. In Fig. 7.1, the number of segments decreases with increasing penalization of the boundary length. At the same time, the inhomogeneities within individual segments increases.

While early algorithms for the Mumford–Shah cartoon model are based on region merging

concepts [119], the approach by Strekalovskiy et al. [166] describes a very fast approximation of this model by means of primal–dual optimization ideas. They propose to minimize the energy

$$\begin{aligned} \min_u E_{MS}(u) &= \sum_{x\in\Omega} \|\boldsymbol{u}-\boldsymbol{f}\|^2 + R_{MS}(\nabla \boldsymbol{u}) \\ \text{with } R_{MS}(g) &= \begin{cases} 1 & g \neq 0, \\ 0 & \text{else} \end{cases} \end{aligned} \tag{7.2}$$

where the image is discretized into a finite rectangular grid Ω. The authors begin with the assumption of a convex and lower-semicontinuous regulariser R.
Let $R : \mathbb{R}^{d\times k} \to \mathbb{R}$ and R^* is its convex conjugate (2.15). Then $R = (R^*)^*$ holds, and states the convex envelope (2.14) of R, i.e. is the largest convex function pointwise below or equal to R:

$$R = (R^*)^* = \sup_{g\in\mathbb{R}^{d\times k}} \langle g^*, g\rangle - R^*(g). \tag{7.3}$$

Let us consider (7.1) with such a convex, lower-semicontinuous regulariser R instead of R_{MS} [166]. Then we can rewrite the energy as

$$E(\boldsymbol{u}) = \sup_{\mathbf{p}:\ \Omega\to\mathbb{R}^{d\times k}} \sum_{\boldsymbol{x}\in\Omega} |u(\boldsymbol{x}) - f(\boldsymbol{x})|^2 + \langle p(\boldsymbol{x}), \nabla u(\boldsymbol{x})\rangle - R^*(p(\boldsymbol{x})). \tag{7.4}$$

The minimization of (7.1) leads to a saddle-point problem. The authors suggest to apply the accelerated primal-dual method of [35] (Sec. 2.3.2), since the data term $D(\boldsymbol{u}) = \sum_{x\in\Omega} \|\boldsymbol{u}-\boldsymbol{f}\|^2$ is uniformly convex with constant $\gamma = 2$, i.e. for any $\boldsymbol{u}, \boldsymbol{v}$ it holds

$$D(\boldsymbol{u}) \geq D(\boldsymbol{v}) + \langle 2\boldsymbol{f}, \boldsymbol{u} - \boldsymbol{v}\rangle + \frac{\gamma}{2}\|\boldsymbol{u}-\boldsymbol{v}\|^2. \tag{7.5}$$

This results in the following update equations [166]:

$$\boldsymbol{p}^{n+1} = \text{prox}_{\sigma_n, R^*}(\boldsymbol{p}^n + \sigma_n \nabla \bar{\boldsymbol{u}}^n) \qquad \text{(Dual problem)} \tag{7.6}$$

$$\boldsymbol{u}^{n+1} = \text{prox}_{\tau_n, D}(\boldsymbol{u}^n + \tau_n\ div\ \boldsymbol{p}^{n+1}) \qquad \text{(Primal problem)} \tag{7.7}$$

$$\tau_{n+1} = \theta_n \tau_n, \quad \theta_n = \frac{1}{\sqrt{1+4\tau_n}} \quad \sigma_{n+1} = \frac{\sigma_n}{\theta_n} \tag{7.8}$$

$$\bar{\boldsymbol{u}}^{n+1} = \boldsymbol{u}^{n+1} + \theta_n(\boldsymbol{u}^{n+1} - \boldsymbol{u}^n) \tag{7.9}$$

where $\bar{\boldsymbol{u}}^0 = \boldsymbol{u}^0$ and $\tau_0\sigma_0\|\nabla\|^2 < 1$. Since for dimension $d \geq 2$ it holds that $\|\nabla\| < \sqrt{4d}$ it is possible to set $\tau_0 = \frac{1}{2d}$ and $\sigma_0 = \frac{1}{2}$ [35, 166].
The proximal operator of the primal update equation, see (7.7), poses trivial pointwise quadratic problems and can be computed directly as:

$$\text{prox}_{\tau_n, D}(\tilde{\boldsymbol{u}}) = \frac{\tilde{\boldsymbol{u}} + 2\tau\boldsymbol{f}}{1+2\tau}\ \text{, where } \tilde{\boldsymbol{u}} = \boldsymbol{u}^n + \tau_n\ div\ \boldsymbol{p}^{n+1}. \tag{7.10}$$

The central observation is that the non-convex regularizer affects the algorithm only within the dual problem (7.6) in the form of the convex conjugate R^* [166]. The basic

idea is then to make use of an inherent property (see Sec. 2.1) to reduce the proximal operator from R^* to R, i.e.

$$\operatorname{prox}_{\sigma,R^*}(\tilde{\boldsymbol{p}}) = \tilde{\boldsymbol{p}} - \operatorname{prox}_{\frac{1}{\sigma},R}\left(\frac{\tilde{\boldsymbol{p}}}{\sigma}\right), \tag{7.11}$$

$$\operatorname{prox}_{\sigma,R}(\tilde{\boldsymbol{h}}) = \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^{d\times k}} \frac{\|\boldsymbol{h}-\tilde{\boldsymbol{h}}\|^2}{2\tau} + \lambda \tag{7.12}$$

$$\text{where } \tilde{\boldsymbol{p}} = \boldsymbol{p}^n + \sigma_n \nabla \bar{\boldsymbol{u}}^n. \tag{7.13}$$

Although the energy is still not convex, there is an explicit formula for the minimizer:

$$\operatorname{prox}_{\sigma,R^*_{MS}}(\tilde{\boldsymbol{p}}) = \begin{cases} \tilde{\boldsymbol{p}} & \text{if } |\tilde{\boldsymbol{p}}| \leq \sqrt{2\lambda\sigma}, \\ 0 & \text{else.} \end{cases} \tag{7.14}$$

Due to its intrinsic parallelism, it is well-suited for parallel processing hardware such as GPUs. In its original formulation, this algorithm has been specified for 2D data sets. For our framework, we have extended this approach in a straightforward way to the segmentation of volumetric data.

Tumor Segmentation On MRI T_2-Flair scans, high-grade gliomas contain areas that are brighter than the brain tissue. We use this prior knowledge and segment for a bright outlier in intensity in the following way: We also include T_1 data in our segmentation, since our experiments show that this makes the segmentation process more robust against small distortions. We start with the parameter $\nu = 400,000$ and check if this gives a segmentation into two areas: the tumor and the background. Since the algorithm of Strekalovskiy et al. might give more than one segment, we postprocess the result with an Otsu thresholding [133]. If the area of the thresholded tumor is larger than 50% of the brain volume, this is an indication that ν was too large such that the tumor has been merged with its background. In this case, we reduce ν by 15% and start the procedure again. This approach is repeated recursively until we have a segmentation where the tumor volume is below 50% of the brain volume.

Confidence Refinement The total tumor segmentation from the previous subsection tends to give an area that is somewhat larger than the real tumor area: It favors sensitivity over precision. To refine this segmentation with a confidence refinement postprocessing, we investigate how well the data term of the Mumford-Shah cartoon model is fulfilled locally. It is sufficient to do this only in the T_2 channel, since this channel always contains the tumor and it has not been used before. Thus, in any location $\boldsymbol{x}$ of the tumor area, we measure the difference $d(\boldsymbol{x}) := |u(\boldsymbol{x}) - f(\boldsymbol{x})|$, where f and u denote the original resp. segmented intensity values of the T_2 channel. Since u and f have values in $[0, 255]$, the difference d lies in the same interval. Large values of $d(\boldsymbol{x})$ indicate that the local confidence in our segmentation result should be low. We quantize the maximal d-range $[0, 255]$ into 256 bins, and we discard those voxels from the segment where the distance lies in the highest bin with nonvanishing contributions.
This confidence refinement is a trade-off between gaining precision and loosing sensitivity:

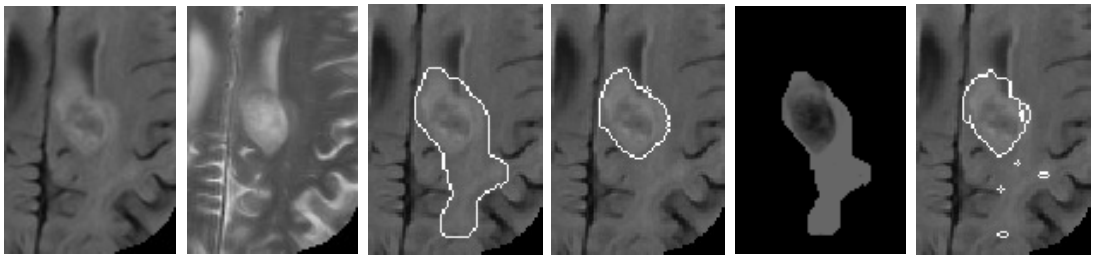


Figure 7.2: Illustration of the confidence refinement procedure. **(a)** T_2-Flair input image for (7.1). **(b)** T_2 input image for consistency refinement. **(c)** Segment boundaries before refinement. Dice score: 0.51, sensitivity: 0.82, precision: 0.37 (3D tumor volume). **(d)** Boundaries of the ground truth segmentation. **(e)** Difference map $d(\boldsymbol{x})$. **(f)** Segment boundaries after refinement. Dice score: 0.77, sensitivity: 0.7, precision: 0.86 (3D tumor volume).

In general, we loose a small amount of sensitivity but gain remarkably in precision: In Fig. 7.2, the Dice score [49] of the segmentation improved from 0.51 to 0.77, and the precision increased even from 0.37 to 0.86, while the sensitivity deteriorated only mildly from 0.82 to 0.7. This illustrates the usefulness of our confidence refinement.

Segmentation of the Tumor Fine-Structure

We use this first segmentation to determine further tumor subcomponents: We minimize the Mumford-Shah cartoon model again with a very small boundary penalization ($\nu = 1$), but this time exclusively on the T_{1c} scans and in the previously defined segment. Afterwards we use Otsu's thresholding to identify the active tumor, i.e. enhancing- and non-enhancing tumor core. In this way, we get a splitting of the complete tumor region into active tumor and necrosis/edema. To get the final subcomponents, we apply Otsu's method on both subcomponents and split the first component, i.e. active tumor, into its enhancing and non-enhancing part and the second subcomponent into necrosis and edema.

We show some exemplary results of our method in Fig. 7.3. Intuitively one would expect that the segmentation performance would deteriorate when the tumor is very similar to the background. Remarkably, this is not the case: We can determine exact tumor boundaries for low-contrast regions, even when the tumor is very small. Apart from that, our method is also able to identify subcomponents properly.
We developed this cascadic Mumford-Shah approach originally for BraTS2014. However, the prior knowledge that a tumor is on average brighter than the remaining brain tissue still holds. Our methods serves therefore as a simple and intuitive baseline for all further experiments.

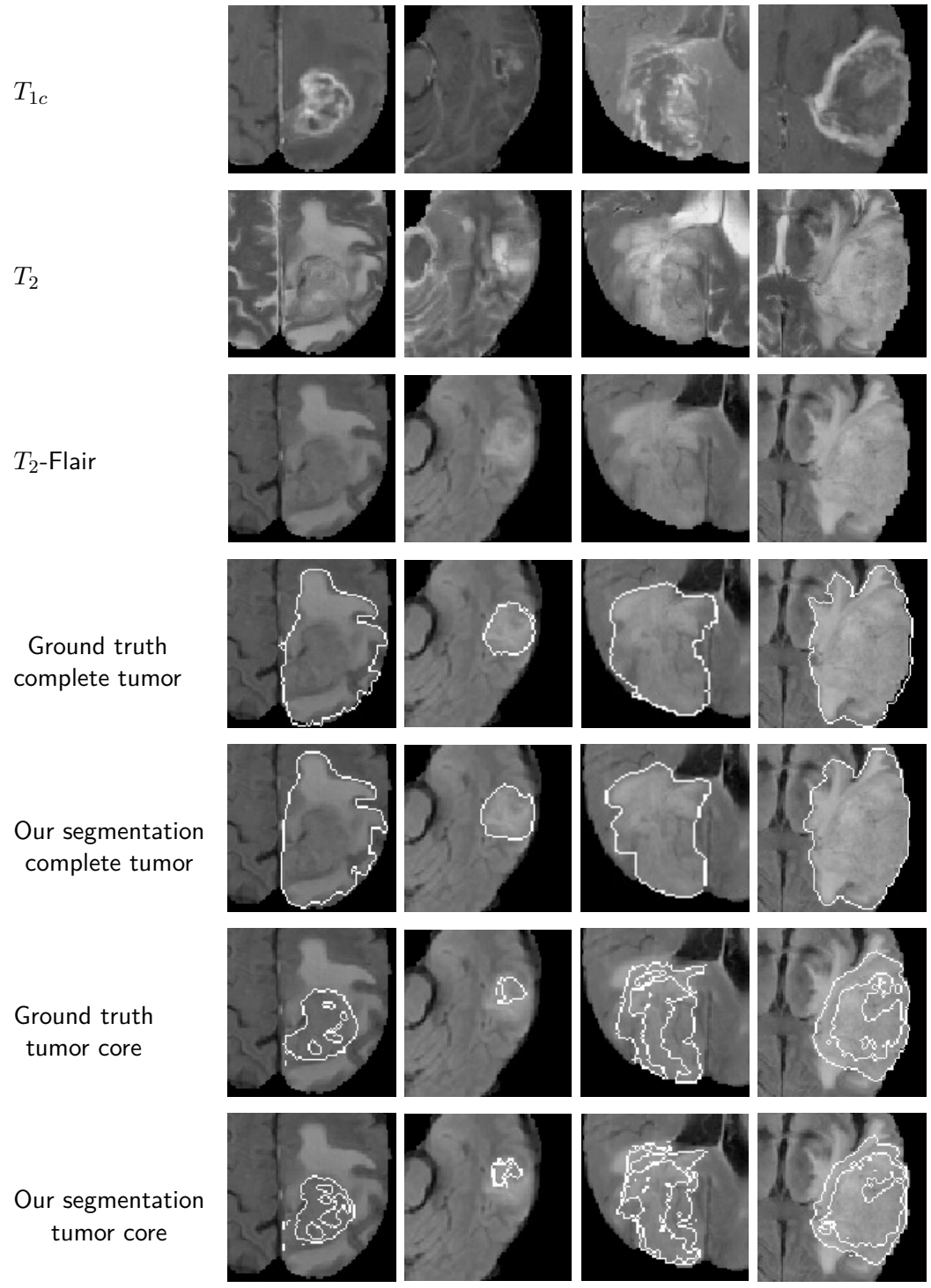


Figure 7.3: Exemplary results of our cascadic Mumford-Shah method for high grade brain tumors.

7.1.2 UNet

The field of computer vision is dominated by "*Convolutional Neural Networks*" (CNN). The main idea of CNNs is to learn a hidden feature mapping from the image domain to a latent space. In a classification setting, this feature mapping transfers the input image into a vector representation used as input to a classifier [40, 80, 89]. Although most of current architectures are designed for these kind of problems, their application to image segmentation is straight forward: Segmentation can be seen as a classification task on pixel level.

An simple idea is to classify patches extracted around each pixel to generate a multi-channel likelihood map with the same dimensions as the original image. Unfortunately, the amount of memory to handle feature maps of the full image resolution is typically infeasible.

An intuitive approach is to downsample the feature maps after a sequence of operations to avoid this curse of dimensionality while refining the abstraction level. However, this results in low-resolution outputs not applicable to image segmentation as we need to reconstruct a full-resolution image.

Obviously, it is much more difficult to reconstruct a full-resolution image from a vector representation, than vice versa. Ronneberger et al. [149] addressed this problem with the UNet architecture, a milestone in medical image segmentation.

The intuition behind its structure is to re-use already learned feature mappings: This architecture can be split into two components, an encoder and a decoder branch connected by a bottleneck; see Fig. 7.4. While the first one learns feature mappings and

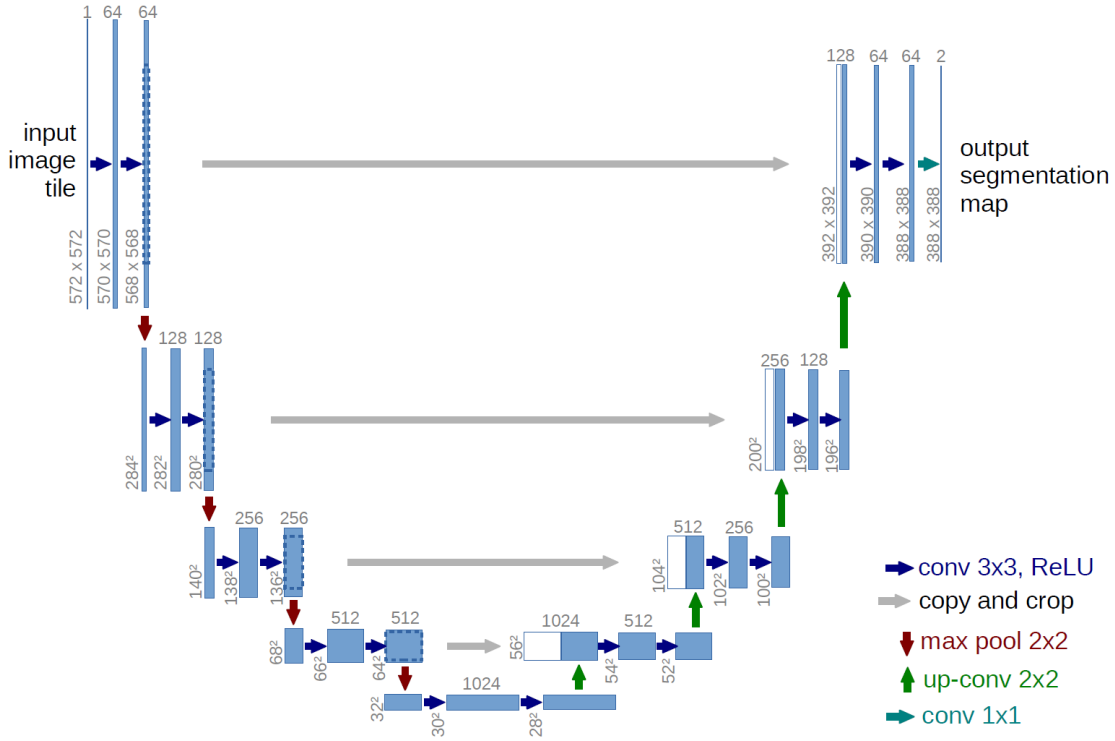


Figure 7.4: Basic structure of UNet approaches. Each blue box indicates a multidimensional feature map, arrows correspond to operations. Image courtesy of Ronneberger et al. [149].

$$\underbrace{\begin{bmatrix} 0 & 3 & 0 & 3 & 0 \\ 3 & 3 & 3 & 4 & 0 \\ 1 & 1 & 1 & 3 & 4 \\ 3 & 1 & 1 & 3 & 3 \\ 2 & 3 & 0 & 3 & 0 \end{bmatrix}}_{\text{Image matrix}} \quad \underbrace{\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}}_{\text{Convolution Kernel}} \quad \underbrace{\begin{bmatrix} 5 & 13 & 9 \\ 11 & 12 & 10 \\ 5 & 11 & 8 \end{bmatrix}}_{\text{Convolved Image}}$$

Figure 7.5: Example of an convolution operation.

contracts the image to its vector representation in the latent space (i.e. the bottleneck), the decoder part reconstructs an image of the original size using the previously learned feature maps [149]. In this way, the structural integrity is maintained while distortions due to lost locality are reduced.

However, before we can explain the architecture in detail, it is necessary to understand its building blocks. Probably the most important ingredient of each deep learning model for computer vision is the *convolution layer*.
The basic idea of convolutional layers follows observations made in classical image processing:

1. Local structures are important as they are semantically meaningful.
2. Essential information can appear everywhere in the image.

The application of filter (or kernel) benches addresses this aspects. Typically, each 2D filter is convolved with the whole 3D input volume, generating a 2D output feature map highlighting the presence of specific features in the input data. Concatenating the filter responses from all kernels in the filter bench results in a 3D output of each convolution layer. Please note, that the output dimension of the feature map can change between convolutional layers; see Fig. 7.4. After the first convolution block, the data volume changes the input dimensions of $572 \times 572 \times 1$ (gray valued image) to $568 \times 568 \times 64$. In order to downsample the tensor, a *max pooling* operation is applied. This simple operator selects the maximal value within a given patch, see Fig. 7.6 for an exemplary 2×2 max pooling. Then, the resulting feature map of $284 \times 284 \times 64$ is fed to the second convolution

$$\begin{bmatrix} \begin{array}{cc|cc} 10 & 20 & 30 & 40 \\ 50 & 60 & 70 & 80 \\ \hline 90 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{array} \end{bmatrix} \rightarrow \begin{bmatrix} \begin{array}{cc} 60 & 80 \\ 90 & 16 \end{array} \end{bmatrix}$$

Figure 7.6: Example of a 2×2 max pooling operation. The maximal value of each patch is selected as output for the given area.

block. The output of the first filter bench is now $282 \times 282 \times 128$.
Hence, not only the number of feature maps increased from 64 to 128, but also the image dimension reduced from 284 to 282. Fig. 7.5 shows a simple example of this operation. The left input matrix is convolved with the kernel depicted in the second matrix, resulting in a smaller output image or feature map. However, the most important difference to

classical image processing is that filter kernels are not predefined, but learned during network training.
Typically, each convolutional layer is followed by an activation function to introduce non-linearity to the network; see Sec. 3.2.1. Generally, activation functions decide whether a feature should be activated, i.e. its input is relevant for the model prediction. Probably the most frequently used activation function is the *ReLU function* (Rectified Linear Unit), defined as

$$f(x) = \max(0, x). \tag{7.15}$$

ReLU is a piecewise linear function, representing the identity for all positive values and zero for all negative inputs. Among other aspects, the sparse activations result in a faster computation time. Unfortunately, a detailed explanation is beyond the topic of this thesis and we refer the interested reader to [136].
The UNet architecture resembles an "U", justifying its name. The encoder part (i.e. the left side of the "U"), contracts the image to the latent space via a series of 3×3 convolutions with ReLU activation followed by a 2×2 max pooling operation. Here, the number of filters and their responses doubles after each convolution block while the image is downsampled. This has two major properties: The complex image features are learned and transferred to a more abstract representation.
The lowermost layer connects the encoder and decoder sections. It is composed of two 3×3 convolutions followed by an 2×2 *up-convolution* (or transposed convolution). In

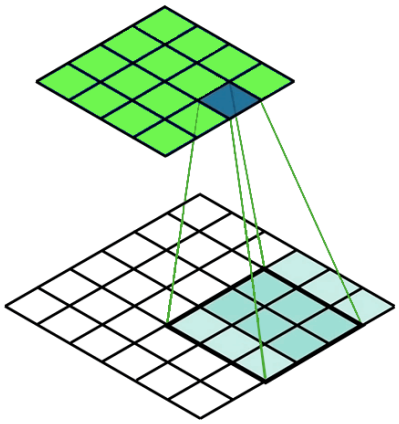


Figure 7.7: Sketched principle of up-convolutions. Each input pixel contributes to several output positions.

contrast to classical interpolation methods, transposed convolutions cannot applied directly: Their weights have to be learned during training.
Fig. 7.7 illustrates the basic principle. In our example, the pixel marked in dark blue distributes information to a neighborhood of positions in the output defined by the kernel size. Each learned weight in the filter is multiplied with the input value and added to the output position. This is then accumulated for every pixel in the output matrix, indicated by the blue shaded areas. Hence, the kernel defines a weighted neighborhood and decides to which amount an input value is distributed to the output positions.
The decoder part of the UNet architecture performs the opposite procedure than the

encoder. Each series of 3×3 convolutions with ReLU activation is now followed by an 2×2 up-convolution. Here, the number of filters and their responses is halved after each convolution block while the image is upsampled.
In order to introduce locality to the massively abstracted feature representations, Ronneberger et al. [149] apply *skip-connections*. Those connections allow to re-use already learned filters - they are simply concatenated at the beginning of each decoder block.
In the end, the final feature maps are fed into a 1×1 convolution operation to adjust the channel dimension to the number of classes in the segmentation.
Finally, the loss is calculated by a pixel-wise softmax

$$\sigma(\mathbf{c})_i = \frac{e^{c_i}}{\sum_{j=1}^{n} e^{c_j}} \text{ for } i = 1, \ldots, n \text{ and } \mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n \tag{7.16}$$

over the final feature map in conjunction with cross entropy [149]. Here, $\mathbf{c}$ is the output of the network for each pixel and n is the number of possible classes. Intuitively, we classify each pixel into one of the possible classes. All in all, this loss function results in larger weights at the border of segmented objects such that individual areas can be identified easily within the segmentation maps.
In its original formulation, the Unet architecture was developed for 2D cell images. However, its extension to 3D images (necessary for volumetric MRI data) is straightforward - the architecture is identical and only replaces all 2D operators with their corresponding 3d variants [42].
In the following we will discuss the No NewNet topology [82]. This recent work shows a very high performance on several datasets. It was developed on the basic assumption that already the original UNet architecture is very powerful and most extensions of its design are not necessary and too complicated. Since we will follow this assumption in a quite similar way, we explain this work in detail.

7.1.3 No NewNet

Since the publication of the UNet architecture, the encoder-decoder strategy has become the dominant approach in image segmentation. Nowadays, almost all new developments in this field are based on architectural modifications of this topology [82, 84, 124].
In the meantime it is almost impossible to predict which architecture might be suitable for a problem due to the multitude of possible extensions: Each of these possibilities has been tested on a specific data set. Unfortunately, it is an inherent part of deep learning that there is an architectural overfit to the data set used - making it almost impossible to decide whether an adjustment is appropriate in a different context.
Isensee et al. [82] implemented a number of these variants and evaluated their usefulness. It is not surprising that they found most of these extensions to be pointless in a general context - compared to a well trained UNet model. Overall, they claim that a generic UNet architecture with a few minor modifications can be sufficient to provide competitive performance.
Similar to the original UNet topology, the encoder part contracts the image to the latent space via a series of four convolution blocks. However, the authors changed these building blocks to a repetition of 3×3 convolutions with instance normalization [173] and leaky ReLU followed by a max pooling operation; see Fig. 7.8. Here, instance normalization is a specific form of regularization, while leaky ReLU is a variant of the original ReLU

activation with a small negative slope for values below zero.
The decoder branch of the No NewNet architecture follows the same strategy as in the UNet topology - the only adjustment are the basic convolution blocks already used in the encoder; see Fig. 7.4.
All in all, their modifications of the network topology can be reduced to the injection of normalization to the network. Obviously, this is fully consistent with current findings that regularization as well as normalization lead to wider optima (with higher generalization performance) in the loss surface [83].
In order to optimize the performance of the model on BraTS benchmark data, the authors suggest a set of additional extensions. In a first step, Isensee et al. exchange the final softmax with a sigmoid function and optimize the segmentation similar to our approach to identify the tumor fine structure; see Sec. 7.1.1. Here, the basic idea is, that the whole tumor includes also the tumor core. Consequently, the area to be segmented reduces to the complete tumor when tumor fine structures have to be identified. In a further adjustment, the authors include more data, namely in-house data (not publicly available) as well as benchmark data from a different challenge. In addition, the authors suggest to apply an unweighted sum of Dice loss and negative log-likelihood as loss function. Finally, their toolchain applies a postprocessing step: The replacement of voxels identified as enhancing tumor with necrosis as long as their amount in an image is below some threshold.
All in all, each of those steps contributed some improvement to the overall performance; see Tab. 7.1. Obviously, most of their changes had a minor impact on the final Dice scores. Here, the cascadic segmentation of tumor fine structures as well as the unweighted sum of two loss functions increased the overall performance in the Dice score by neglectable 0.002 and 0.005, respectively. However, the postprocessing step as well as the training on additional data noticeably improved the error metric by 0.032 (enhancing core) and 0.013 (complete tumor), respectively. Therefore we deeply believe that their main improvement in performance was caused by the inclusion of more training data, i.e. by reducing the overfit of the model to the training distribution.

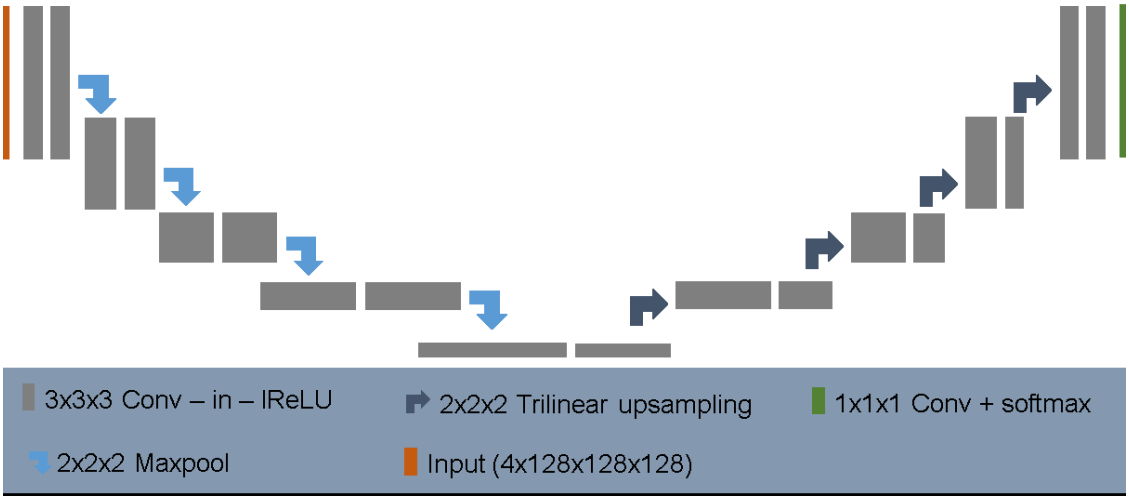


Figure 7.8: Architecture of the No NewNet model. Each gray box corresponds to a series of convolution, instance normalization, and leaky relu. Arrows indicate up and down sampling operations, respectively. Image courtesy of Isensee et al. [82].

Table 7.1: BraTS18 evaluation of the No NewNet architecture (training data).

Model	Dice Score		
	Enhancing	Complete	Core
Baseline	0.734	0.898	0.822
Baseline + reg	0.738	0.900	0.829
Baseline + reg + post + loss	0.768	0.903	0.836
Baseline + reg + cotr(ds1)	0.759	0.913	0.853
Baseline + reg + cotr(ds1) + post	0.787	0.913	0.853
Baseline + reg + cotr(ds1) + post + loss	0.786	0.918	0.857
Baseline + reg + cotr(ds2) + post + loss	0.763	0.904	0.844

7.1.4 NVDLMED: Autoencoder Regularization

The winner of last years' BraTS challenge, also followed a basic UNet architecture [124]. While the backbone can still be reduced to an encoder-decoder structure, the author dramatically increased the model size and extended most of the basic topology by additional operations; see Fig. 7.9. Although the encoder branch is still similar, its building blocks are massively changed. An additional skip connection inside each convolutional block is combined with a group normalization [195], while the max pooling operation for downsampling is exchanged with another convolution and stride 2. The decoder branch follows the same strategy but with only one convolution block per spatial level.
Probably the most important change is an additional variational autoencoder branch reconstructing the input image to itself. This sub-network is then used during the training phase as regularization.
In order to improve the model performance, NVDLMED is built on an ensemble of 10 different networks. Unfortunately, this setting results in a very large network, that can only be trained on NVidia V100 GPUs, or on a CPU cluster.

7.1.5 Cascadic Neural Networks

Zhou et al. [200] approach the task of brain tumor segmentation from a slightly different perspective. While most of the state-of-the-art methods consider the identification of the complete tumor and its subcomponents as a single problem, the authors decompose the segmentation challenge into three different sub-tasks. In a first step their method performs a coarse segmentation to detect the complete tumor. Afterwards, the segmentation is refined and intra-tumoral classes are segmented. Finally, this segmentation is again optimized to classify the enhancing tumor core. This cascade of segmentation tasks is realised with two different network topologies. On the one hand, Zhou et al. make use of 3D FusionNets [180]; see Fig. 7.10. to extract the multi-scale context information. On the other, they apply one-pass multi-task networks [201]. In addition, Zhou et al. [200] perform several modifications, such that the final ensemble contains seven different neural network architectures whose results are averaged for the final model prediction.

7.1.6 Preprocessing for Deep Neural Networks

Typically MR images are recorded from different hospitals with varying scanners and no standardized parameter settings. This results in strong variations in the MR intensities: Even the same sequence of the same patient (e.g. T_2) acquired at the same machine, can differ dramatically due to inconsistent parameter choices.
Deep neural networks learn the data distribution provided by the training set. Hence, it is essential that the value range in the training data corresponds to the range present in the test set. In order to compensate for these variations, we follow [82] and adjust each modality independently. In a first step, we substract the mean of the brain region and normalize by its standard deviation. Afterwards, we remove outliers by clipping and rescale the images to the range $[0, 1]$.

7.1.7 Postprocessing

Although deep neural networks proved to produce segmentation results of high quality, post-processing is a necessary step in a medical context. The brain tumor segmentation challenge contains high- and low-grade gliomas. While the high-grade tumors typically consist of an enhancing tumor core, it is rarely present in low-grade abnormalities.
In order to compensate for this prior knowledge, we follow [82] and apply a postprocessing step to remove potentially false labels of the enhancing tumor core in low-grade gliomas. Our segmentation approach (see Sec. 6.1) already proved its ability to correct for false labels. Hence, we postprocess the segmentation masks of the deep learning models as follows: In a first step, we sample every eight voxel in the output mask to sparsify the data. Afterwards we incorporate this mask in the cost term of our semi-automatic approach and densify the segmentation.

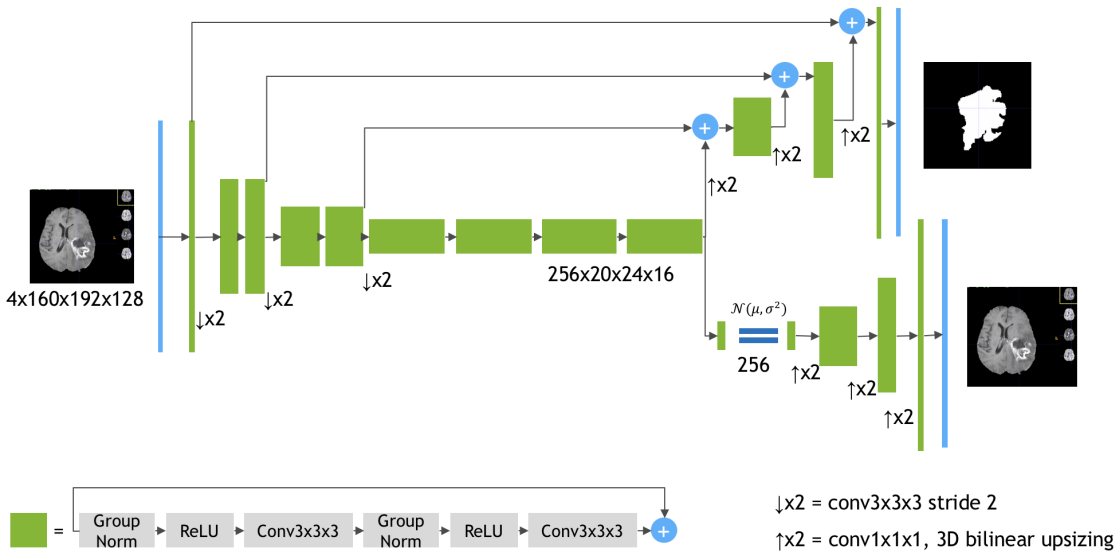


Figure 7.9: Architecture of NVDLMED. In contrast to the basic UNet structure, a second decoder branch is implemented.

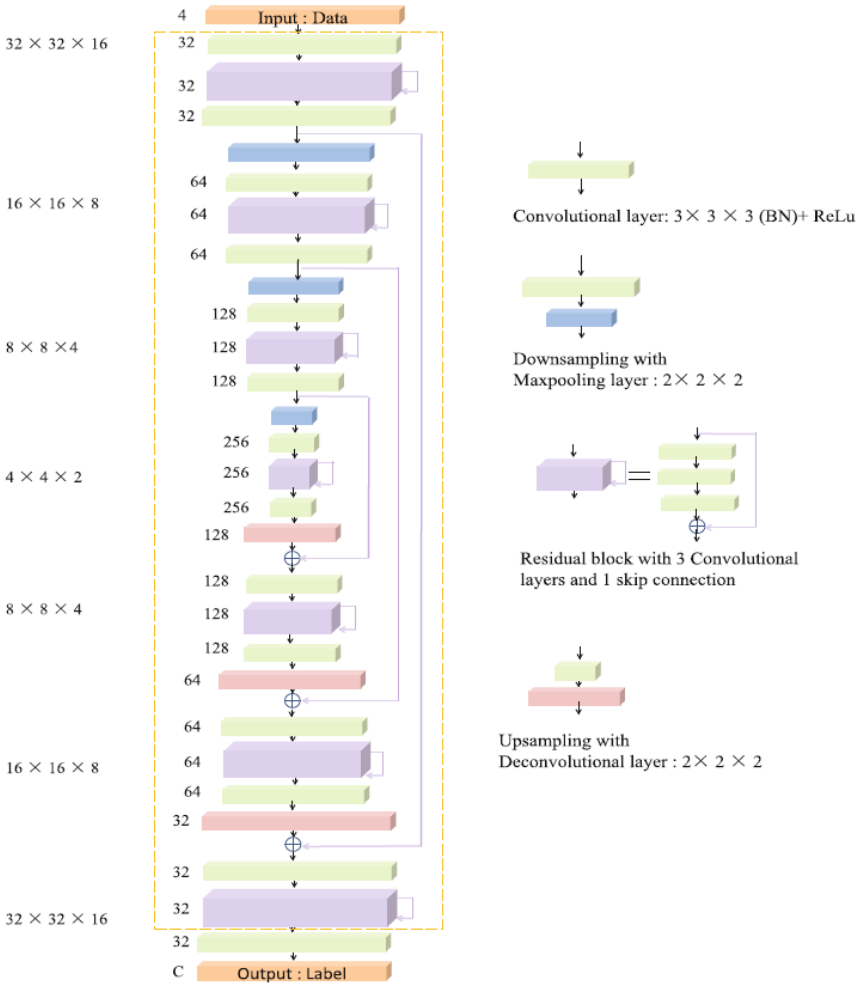


Figure 7.10: Topology of FusionNets. Although the basic UNet structure is still present, the architecture changed dramatically.

7.2 Improving the Generalization Performance

Overfitting is one of the major problems in training of deep neural networks. Typically, this issue is caused by a lack of training data in combination with complex models. Especially in the situation of medical image segmentation, the amount of data is rather limited. There are several approaches to relax this problem: Obviously, the most straight forward idea is to add more training data. However, this is typically a severe problem. Another possibility is to reduce the capacity of a model by reducing its size. Of course, it is also an option to regularize either the weights or the loss functions of a model. One more strategy is to include normalization layers: Recent work [83] indicates that normalization layer lead to wider optima and therefore better generalization.

In the following, we discuss two approaches: The first one, octave convolutions [40], addresses the reduction of weights in a neural network while not reducing its capacity. This advanced operator allows to exploit the mixture of frequencies inherent to each image. Second, we illustrate the stochastic weight averaging [83] that enables the optimization algorithm to converge to wider and therefore better generalizing optima in the loss surface.

7.2.1 Octave Convolutions

The fundamental aspect of convolution layers is their ability to identify local structures in their input data. These characteristics are then assigned to a new filter response - typically the image resolution does not change during this process.
However, each image can be divided into its low-frequency signal, which describes the coarse structure and the global layout, and its high-frequency signal, containing fine details; see Fig. 7.11.
Although this is well known in the classic image processing community, this inherent information cannot be exploited by standard convolutional layers. Recently there are several attempts to express this structure within layers of deep neural networks [40, 89]. The multigrid approach of Ke et al. [89] maps every convolutional layer into a pyramid of operations. In this way, features at different scales can be extracted. However, this type of strategy obviously has a massive disadvantage: The amount of required parameters

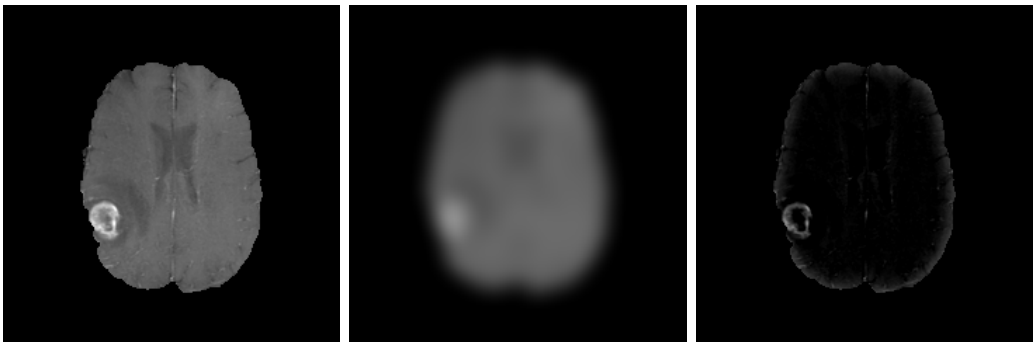


Figure 7.11: Illustration of low and high frequency parts in an image. The input image of a MRI of the brain (left) is split into its low frequencies (middle) and high frequencies (right).

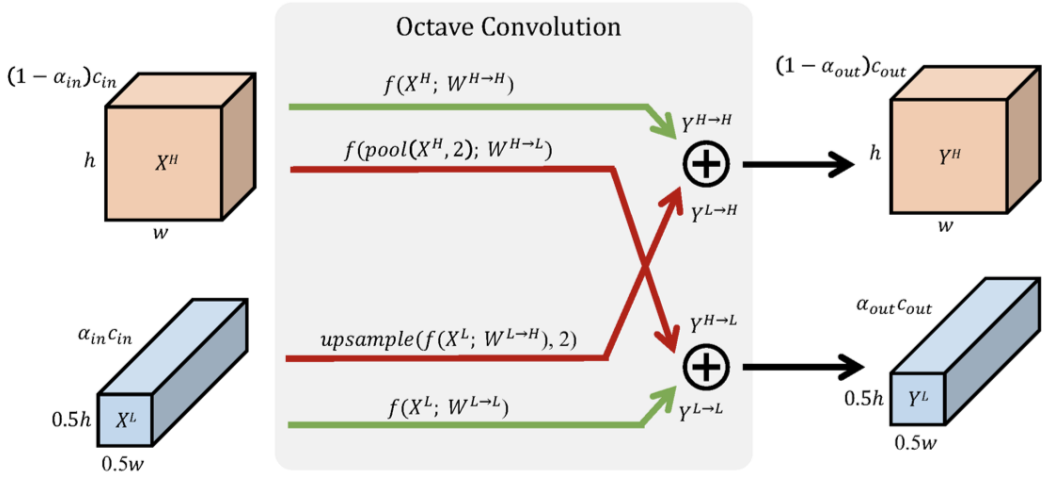


Figure 7.12: Detailed design of octave convolutions. Red arrows indicate communication between low- and high-frequency components, green arrows depict regular information updates. Image courtesy of Chen et al. [40].

increases with the number of scales in the pyramids.

Octave convolutions use a similar concept but interpret output feature maps as mixtures of information at different frequency scales [40]. Hence, these advanced convolutions factorize the output maps only into two groups: low and high frequencies. The corresponding smoothly changing low-frequency maps are then stored in a low resolution tensor (half of the original input resolution) to reduce spatial redundancy [40]; see Fig. 7.12.

Following this idea, octave convolutions process low frequency information with corresponding (low frequency) convolutions. This not only increases the receptive field in the original pixel space, but also collects more contextual information. Since the resolution for the low-frequency filter responses can be reduced, this saves both computational load and memory consumption.

The effort for such an octave convolution architecture consists in an additional hyper parameter $\alpha \in [0; 1]$ indicating the ratio of low frequency components. Then, the input feature map $X \in \mathbb{R}^{c\times w\times h}$ can be written as

$$\begin{aligned} X_H &\in \mathbb{R}^{(1-\alpha)\times c\times w\times h} \\ X_L &\in \mathbb{R}^{\alpha\times c\times \frac{w}{2}\times \frac{h}{2}} \end{aligned} \tag{7.17}$$

where w is the input width, h the input height, and c the number of channels. In addition to their ability to exploit spatial redundancy, octave convolutions also enable an efficient communication between low- and high-frequency components of the filter responses.

Let $X = \{X^L, X^H\}, Y = \{Y^L, Y^H\}$ be the factorized input and output tensors, respectively. Then the output feature map is defined as

$$\begin{aligned} Y^H &= Y^{H\to H} + Y^{L\to H} \\ Y^L &= Y^{L\to L} + Y^{H\to L}. \end{aligned} \tag{7.18}$$

Here $Y^{A\to B}$ denotes the update from group A to B, i.e. $Y^{H\to H}, Y^{L\to L}$ state intra-frequency updates and $Y^{H\to L}, Y^{L\to H}$ inter-frequency communication [40]; see Fig. 7.12.

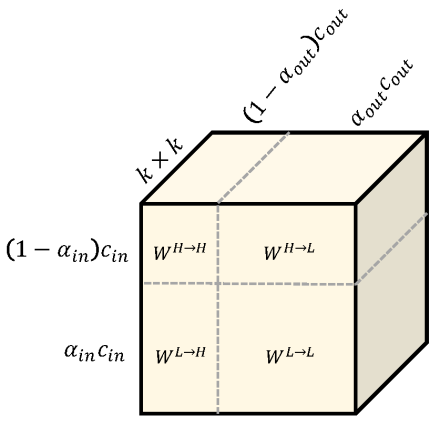


Figure 7.13: Illustration of the octave convolution kernel. The kernel is split in intra- and inter-frequency parts. Image courtesy of Chen et al. [40].

In order to compute the output feature maps, the convolution kernel is split accordingly; see Fig. 7.13.
Obviously, filter responses of intra-frequency maps can be computed with regular convolutions. However, up- and down-sampling (or pooling) operations for inter-frequency computations can also be fold up into the convolutions; see [40] for more details.
In total, the application of octave convolutions is straight-forward. Due to its inherent design, it is a plug-and-play component, not leading to any architectural changes. In some of our experiments, we replaced all standard convolutions with their octave variants. Although this change had no consequences with respect to network architecture, it dramatically reduced the size of the models while improving their generalization behavior; see Sec. 7.3.

7.2.2 Stochastic Weight Averaging

The training of deep neural networks is a tedious and time consuming task. While in most cases, the capacity of the model architecture is large enough to solve the depicted problem, finding reasonable hyperparameters (e.g. learning rate, batch size, etc) can be challenging: Especially the learning rate has massive influence to the training procedure and an optimal value is of crucial importance. In medical image segmentation, neural network architectures tend to be complicated and can easily overfit due to a limited amount of training data. In this scenario, an appropriate learning rate is even more important.

Typically, deep neural networks do not converge to a global minimum. Therefore, the quality of the model is evaluated with respect to its generalization performance. In general, local optima with flat basins tend to generalize better than those in sharp areas [80, 83, 159]. Since even small changes in the weights can lead to dramatic changes in the model prediction, these solutions are not stable.

If the learning rate is too low, the model converges to the nearest local optimum and may hang in a sharp basin. Once the learning rate is high enough, the inherent random motion of the gradient steps not only prevents the solution from being trapped in one of the sharp regions, but can also help the optimizer to escape. Obviously, finding a reasonable learning rate boils down to the trade-off between convergence and generalization.

Probably the most common strategy to solve this problem is the usage of a cyclic scheme [103, 159]. In *cosine annealing*, the learning rate cyclically decreases from a given maximal value following the cosine function [103]. It turned out, that each of the local optima at the end of the cycles had similar performance, but lead to different but not overlapping errors in the model prediction; see Fig. 7.14. Hence, Huang et al. [80] suggested to combine the local optima of each cycle into an ensemble prediction.

Unfortunately, computation time at inference increases dramatically with the number of snapshot models used in the ensemble.

Stochastic weight averaging follows the same idea but at a fraction of computational

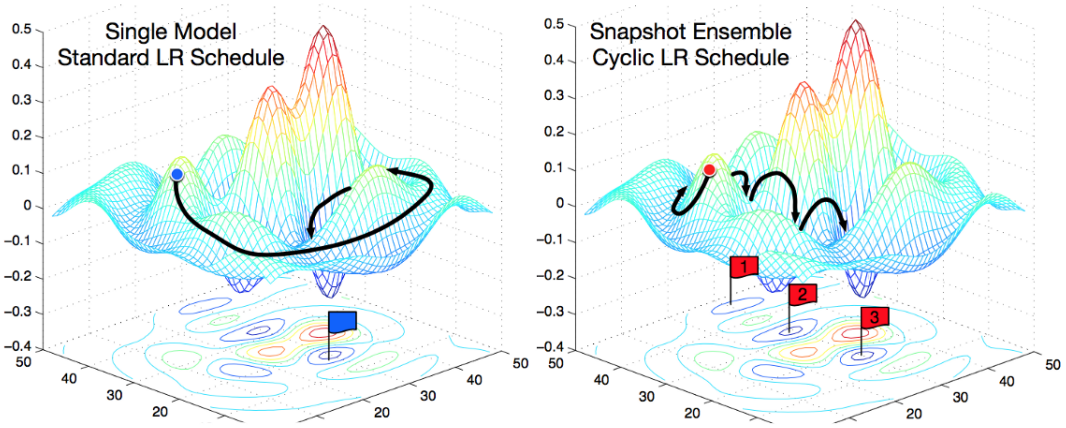


Figure 7.14: Illustration of different model snapshots. While the standard learning rate schedule slowly converges to the minimum, snapshot ensembles are a combination of different local optima. Image courtesy of Huang et al. [80]

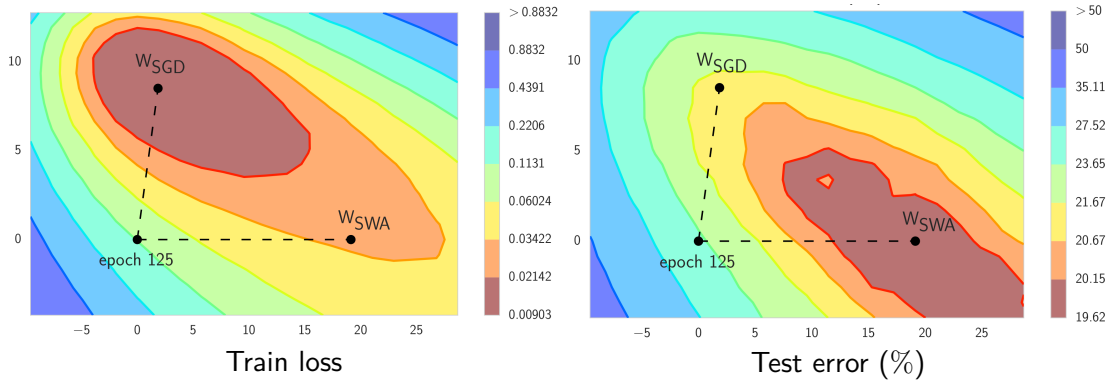


Figure 7.15: Illustration of SWA and SGD showing the weights suggested by SGD and SWA at convergence. SWA started the weights of SGD after 125 training epochs. Image courtesy of Izmailov et al. [83]

load. The basic idea is to conduct an equal average of the weights traversed by the optimizer with a learning rate schedule [83]. Intuitively, by taking the average of several local optima in the loss surface, a wider basin can be reached with better generalization performance [9, 83].
In contrast to ensemble approaches, we only need two models: The first one keeps track of the running average of the model weights, while the second one is traversing the weight space. At the end of each learning rate cycle, the state of the second model is used to update the weights of the running average model as

$$w_{\text{swa}} = \frac{w_{\text{swa}} * n_{\text{models}} + w}{n_{\text{models}} + 1}. \tag{7.19}$$

Here, w_{swa} are the weights of the running average model, while w are the weights of the model traversing the weight space, respectively. The total number of models to be averaged is given by n_{models}. All in all, stochastic weight averaging significally improves generalization performance [9], being less prone to the shifts between train and test error loss; see Fig. 7.15. In general the strategy can be divided in two phases: In the first phase of 75% of training time, the learning rate schedule follows a standard scheme - e.g. it is fixed to a specific value and decays after several epochs. In the second phase, the learning rate can be set to a constant value or follow a cyclic scheme to encourage the exploration of the loss surface; see Fig. 7.16.

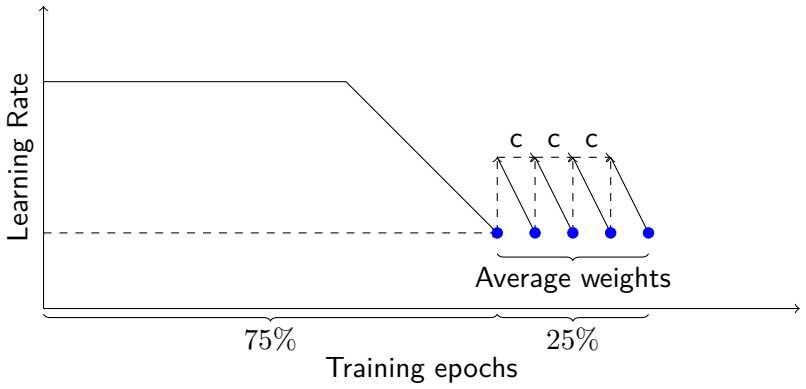


Figure 7.16: Sketch of stochastic weight averaging.

7.3 Experiments

The brain tumor segmentation challenge is a widely accepted benchmark data set [10, 11, 115]. The challenge contains skull-stripped and spatially registered multimodal MR images (T_1, T_{1c}, T_2, and T_2-Flair) with a voxel size of $1mm$ in every direction. Tumors are of different shape, size and location in each data set.
In 2018, the BraTS challenge contained 285 training instances accompanied with 66 validation and 191 test cases. Unfortunately, the testing data set allows only for a single submission, disqualifying this compound for our analysis. However, we found the validation data set to be rather small and therefore not expressive. We decided to rely in our evaluation on five-fold cross validation on the training data set. Consistent with Isensee's view [82], we are convinced that the conclusions drawn from the training set with cross validation are more general in nature and more robust to changes in the underlying distributions.
We performed nearly all network training on four NVidia Titan V with 12GB memory and 5120 cuda cores. In case of NVDLMED, we do not have a graphics card with sufficiently large memory: We trained this network for several weeks on Intel Xeon Gold 6132 ("Skylake") with 28 CPU cores and 192GB of main memory.
We set all hyperparameters of the considered networks as described in their publications and used code provided by the authors whenever possible.
In order to generate a baseline for our experiments, we evaluated all analyzed approaches on the original BraTS2018 training data; see Tab. 7.2. Here, "CMS" denotes our cascadic Mumford-Shah method (Sec. 7.1.1) while "CascNN" means the cascadic segmentation approach with multiple neural networks (Sec. 7.1.5), and "No NewNet" refers to the No NewNet approach with region optimization and postprocessing (Sec. 7.1.3). Please note that "NVDLMED" represents a single NVDLMED network (Sec. 7.1.4) not an ensemble of several models.
The neural networks surpass the cascaded Mumford Shah approach as expected. Nevertheless, the assumption that a brain tumor has higher average intensities in T_2-Flair images is a reliable prior knowledge: This intuitive method shows a remarkable performance when the entire tumor is considered. However, the most significant difference between the results of the various networks is shown in their accuracy to identify the enhancing tumor core.

Typically, a medical benchmark data set is intended as a biased version of a particular

Table 7.2: BraTS18 evaluation for different segmentation approaches in terms of Dice score. No additional disturbances.

Method	Enhancing	Complete	Core
CMS	0.70	0.84	0.76
UNet	0.73	0.89	0.82
CascNN	0.78	0.89	0.84
No NewNet	0.77	0.90	0.84
NVDLMED	0.82	0.91	0.86

Table 7.3: BraTS18 evaluation for different segmentation approaches. Gaussian Noise ($\sigma = 0.02$) is added to the validation data.

Method	Enhancing	Complete	Core
CMS	0.69	0.82	0.74
UNet	0.71	0.82	0.75
CascNN	0.65	0.76	0.76
No NewNet	0.72	0.83	0.79
NVDLMED	0.68	0.80	0.74

problem, i.e. in the case under consideration, all patients with high-grade brain tumors in MRI sequences. BraTS addresses this issue by providing comprehensive multi-institutional routine examinations of glioblastoma multiforme (GBM/HGG) and low grade gliomas (LGG) with pathologically confirmed diagnosis [10, 11]. However, care was mostly taken to create a representative visual representation of the brain tumors themselves. In a real clinical scenario, time and cost pressures usually prevail. For this reason, the assumption that voxels have a size of $1mm$ in all directions is not realistic. In fact, exactly the opposite is typically true: While in-slice images are taken at high resolution, across-slice images are mostly sampled at lower resolution.
In addition, noise also plays an important role in MRI images. These recordings are very costly and time-consuming: Often, MR sequences differ dramatically in sampling rates and suffer from heavy noise disturbances. All in all, real clinical MR images do not correspond to the scheme of the BraTS benchmark data. In order to ensure the applicability of segmentation approaches tested on BraTS data in everyday clinical practice, it is necessary for them to show high generalization performance.
For this reason, we analyze the outcomes of the different approaches when the distribution of the validation data set does not exactly match that of the training data. In a first step, we add Gaussian noise with zero mean and standard deviation $\sigma = 0.02$ to the validation data. The results are depicted in Tab. 7.3. The Dice scores indicate that the prior information about tumor appearance used in the cascadic Mumford-Shah approach is highly robust to disturbances. Although this approach performed worse than the considered neural networks in the original setting, it copes relatively well with the noisy data and the Dice score is only marginally reduced (≈ 0.02 for all categories).
On the contrary, all of the tested neural networks have a major problem with the different distribution in the validation data. All of them show a significant decline in their segmentation performance. This problem obviously also becomes more serious the more complicated the respective architecture is. While the basic UNet as well as the No NewNet model drop by a Dice score of ≈ 0.06 on average, the much more complex CascNN and NVDLMED show a significant decline by a Dice score of ≈ 0.11 and ≈ 0.12, respectively. This is consistent with our assumption that the best performing models are not the ones that generalize best on the test data, but only have the strongest overfit. This conclusion unfortunately disqualifies models trained and evaluated on BraTS data to be directly applied in a real clinical scenario.
The two approaches No NewNet and NVDLMED were almost equal in the evaluation of the BraTS18 challenge and our analysis in Tab. 7.2. Since the NVDLMED in particular

Table 7.4: BraTS18 evaluation for different segmentation approaches. Gaussian noise ($\sigma = 0.02$) is added to training and validation data.

Method	Enhancing	Complete	Core
CMS	0.69	0.82	0.74
UNet	0.72	0.87	0.81
CascNN	0.76	0.89	0.81
No NewNet	0.75	0.90	0.84

shows a strong overfit on the data set while its training is extremely computationally intensive, we exclude this network in the following from our evaluation.
An obvious remedy to cope with noisy validation data is to add the same noise distribution to the training data as well; see Tab. 7.4. In fact, this additional information helps the three deep learning approaches to handle the altered distribution and their performance returns close to the original value. Of course it would be possible to add different noise distributions to the training data. However, at training time it is usually not known how much noise is present in the test set. Another approach would be to include a preprocessing step to denoise the input images. Unfortunately, this idea also has a massive disadvantage: Small details might be lost. In our opinion, both approaches only lead to disguising the problem, but not to solving it. For this reason we address the overfitting in the network topology itself.
In the following we consider the No NewNet (without the adjustments suggested by the authors) as our baseline. Similar to our first experiments, we add Gaussian noise with zero mean and standard deviation $\alpha = 0.02$ and $\alpha = 0.04$ to our validation data; see Tab. 7.5. It turns out that the model in its simplest form performs similar to our cascadic Mumford-Shah method when not much noise is present in the data. However, as soon as the noise is seriously altering the data distribution, the model prediction collapses and is outperformed by the classical approach. Obviously, the generalization performance is limited and the network overfits the training data.
Octave convolutions (see Sec. 7.2.1) have already shown in various applications that, in addition to a massive reduction in model size, they also contribute to improving generalization performance [40]. Consequently, we exchange all ordinary convolutions in the model by 3D octave convolutions ($\alpha = 0.75$). Although this minor change does not alter the network topology, in both settings the performance increases by ≈ 0.02 and 0.03 in Dice score. This indicates a better generalization at inference to the validation data.
In a next step, we apply stochastic weight averaging (see Sec. 7.2.2) with a cycle length of 10 after 75% of the training epochs. This adaptation of the training cycle obviously has a massive influence on the generalization behavior. The averaging of multiple minima in the loss surface allows the model to cope well with the disturbed data while neither the model capacity nor the training time is increased: While the model improves in the first scenario by ≈ 0.04 on average, the performance gain of ≈ 0.05 in the second setting with heavier noise disturbances is massive. The results of both of these modifications let us conclude that overfitting is indeed a serious problem - otherwise our changes would not lead to such drastic improvements.
Afterwards, we use the sparsified results as input for our semi-automatic segmentation

approach; see Sec. 7.1.7. In the first setting, this postprocessing step mainly corrects for false positive labels of the enhancing tumor core; see Tab. 7.5. However, in the second scenario the robust energy formulation stabilizes the segmentation and increases the overall performance for all classes.
In the end, we evaluated our final model (No NewNet+OctConv+SWA+post) on the original BraTS data without additional noise. We did not observe any drop in its performance: With Dice scores of 0.79 for the enhancing tumor core, 0.90 for the whole tumor and 0.85 for the tumor core our approach is on par with current state-of-the-art approaches.

Please note, that we do not want to propose the next neural network trained on BraTS data. We rather want to highlight that generalization is a serious problem when improving on the benchmark metrics is the main goal. Of course one might argue, that those networks are never meant to be directly applied in a clinical setting. We only partly agree with this opinion. First, BraTS was originally designed to allow for a fair comparison and especially to push research in the direction of brain tumor segmentation. In this context, neural networks that can only be applied to benchmark data sets counteract the goal of a medical image segmentation challenge. Second, networks with a high performance on these data sets should at least perform similar on real data - but in our experiments, all approaches except the No NewNet architecture showed a much lower performance than in the benchmark setting - and even dropped below our method that exclusively rely on reliable prior information. Third, we are deeply convinced that increasing complex models do not lead to a satisfying real-world performance.
Similar to Isensee et al. [82], we implemented several suggested network extensions and

Table 7.5: BraTS18 evaluation for different adaptations of No NewNet. Gaussian Noise is added to the validation data.

Slight disturbance ($\sigma = 0.02$)			
Method	Enhancing	Complete	Core
CMS	0.69	0.82	0.74
Baseline	0.69	0.82	0.76
Baseline + OctConv	0.71	0.84	0.77
Baseline + OctConv + SWA	0.74	0.88	0.83
Baseline + OctConv + SWA + post	0.78	0.89	0.84
Moderate disturbance ($\sigma = 0.04$)			
Method	Enhancing	Complete	Core
CMS	0.67	0.79	0.71
Baseline	0.66	0.72	0.70
Baseline + OctConv	0.69	0.77	0.72
Baseline + OctConv + SWA	0.71	0.82	0.79
Baseline + OctConv + SWA + post	0.73	0.85	0.81

found them mostly pointless. Our experiments even indicate, that they might be harmful as soon as training and validation data are not generated by the exactly same distribution. Hence, we fully agree that a well trained UNet architecture is sufficient to solve this segmentation task.
All in all, we improved the generalization performance of the No NewNet architecture by straight forward adjustments in the model and the training procedure itself.
Although we neither changed its topology nor did we need to include the noise distribution in our training data, we could robustify the network while improving its generalization performance. Since this model is actually only a slightly modified version of the original UNet, we are deeply convinced that our suggested modifications also apply to similar structures.

7.4 Summary and Conclusions

Within this chapter we have addressed the general problem of model overfitting of deep neural networks in brain tumor segmentation. Although the basic assumption to learn a class distribution from the training data is very powerful, it is also an Achilles heel when training and validation data slightly differ.
In a first step, we added noise to the validation data. Unfortunately, our evaluations showed that such small variations lead to a massive drop in network performance for two of the three best performing methods of BraTS 2018. Afterwards, we analyzed the behavior of networks when training and validation data both are disturbed in the same way. It turned out, that this additional information allows the network to cope with noisy data. However, since adding noise to the training data can have massive side-effects, we suggested several straightforward modifications to be included in network designs. Last but not least, we showed that these adjustments dramatically improve the generalization performance. Although we did not include the disturbance in the training data, we could reach with the same network topology nearly the same performance than without adding noise.

8 Wilms' Tumor Classification

"The truth is rarely pure and never simple."

– Oscar Wilde

Contents

In about 40% of all children with nephroblastoma, so-called nephrogenic rests can be detected; see Sec. 2.5. Since these only occur in 0.6% of all childhood autopsies, they are considered a premalignant lesion of Wilms' tumors [19].
The diffuse or multifocal appearance of nephrogenic rests is called nephroblastomatosis [101, 135]. Despite the histological similarity, nephroblastomatosis does not seem to have any invasive or metastatic tendencies - it is not malignant.
In order to adapt the therapy accordingly and not to expose children to an unnecessary medical burden on the one hand and to maximize their chances of survival on the other, it is necessary to distinguish Wilms' tumor and its precursor nephroblastomatosis at the beginning of treatment.
Its visual appearance has been described before as small homogenous abdominal mass [43, 148]. However, all existing publications describe the visual appearance on usually very small data sets [69, 148]. So far, it has never been validated statistically to what extent the described features are sufficient for classification.
Nevertheless, the classification is not reliable and it is currently not possible to distinguish Wilms' tumors and their precursor lesion: Patients with the benign nephroblastomatosis are treated the same way as those with a nephroblastoma. In order to distinguish these two diseases, a profound classification fulfills the following criteria: First, the classification accuracy has to be high. Second, the method should be robust to noise. Last but not least, false positives should be as rare as possible.
Especially the third criterion is important: It is less worse to expose a child to an unnecessary chemotherapy than to miss the treatment and reduce the chances of survival.

Hence, we ask the following questions:

- Are the clinical assumptions about the visual appearance of nephroblastomatosis correct? Can we use them for classification?
- Are there other properties that we can use for this problem?

Thus, in a first step we review the current clinical practice in Sec. 8.1. For this purpose, we use our proposed data set (Sec. 4.4) and evaluate whether the assumed properties can solve the classification problem between these two entities. Afterwards, we investigate more texture properties of nephroblastomatosis that allow us to distinguish these two diseases and dramatically simplify the problem; see Sec. 8.2.

8.1 Current Clinical Practice

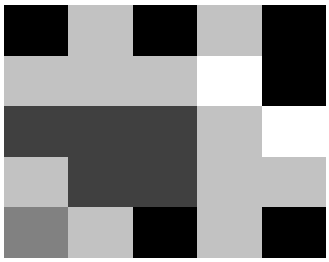
Nowadays, therapy and treatment planning of Wilms' tumors and nephroblastomatosis are equivalent: The distinction of these two diseases is not reliable. However, clinicians assume nephroblastomatosis to be smaller and more homogeneous than Wilms' tumors [135]. Unfortunately, these observations are not based on a reproducible measurement - they solely rely on the opinion of the respective human observer. Indeed, nephroblastomatosis can be visually very similar to a Wilms' tumor, see Fig. 8.1. Although the basic assumptions about its appearance (homogeneity and size) are correct (Fig. 8.1(a)), it is difficult to make a proper differentiation to Wilms' tumors. Obviously, Wilms' tumors can also be very homogeneous; see Fig. 8.1(b-c).
Hence, the question is not if homogeneity is present in case of nephroblastomatosis. The

Figure 8.1: Exemplary MR images from our classification data set. (a) Nephroblastomatosis, (b) Wilms' tumor (blastemal dominant), (c) Wilms' tumor (regressive).

more important aspect is if these criteria - homogeneity and size - are decisive enough to be the fundamental information for treatment decisions. In the following, we evaluate the expressiveness of these two features for the classification of Wilms' tumors and nephroblastomatosis.

8.1.1 Feature Extraction

Haralick et al. [73] suggested a number of texture features. In the last decades, his classical but time proven characteristics influenced the progress in many applications ranging from segmentation [199] to object tracking [109] and classification [152]. Many other methods have been proposed since then, but they are still widespread and widely used.
The main building block is the computation of so-called co-occurrence matrices. These spatial dependence matrices represent the distribution of co-occurring pixel values and serve in most cases as texture detector.
Generally, for two pixels with values a and b the occurrence of value a being horizontally adjacent to b is counted. Figure 8.2 shows an image with five different gray values, its corresponding matrix and the co-occurrence matrix. Consider for example the tuple $(3, 0)$. It occurs four times in the image matrix. Thus, the co-occurrence matrix has a value of 4 at position $(3, 0)$.
In this basic formulation, these features are not rotational invariant. In order to approximate this property, a straightforward idea is to not only consider the right neighbor but all of them with respect to offsets Δ_x in x and Δ_y in y direction, respectively.



Image

$$\begin{bmatrix} 0 & 3 & 0 & 3 & 0 \\ 3 & 3 & 3 & 4 & 0 \\ 1 & 1 & 1 & 3 & 4 \\ 3 & 1 & 1 & 3 & 3 \\ 2 & 3 & 0 & 3 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 0 & 0 & 4 & 1 \\ 0 & 3 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 3 & 2 & 1 & 3 & 0 \\ 0 & 0 & 0 & 2 & 0 \end{bmatrix}$$

Image matrix Co-occurrence matrix

Figure 8.2: Exemplary matrix and its co-occurrence matrix. The co-occurrence matrix reflects the distribution of co-occurring pixels.

More formally, these spatial dependence matrices are then defined as

$$\mathrm{GLCM}_{\Delta_x,\Delta_y}(a,b) = \sum_{x=1}^{n_x}\sum_{y=1}^{n_y} \begin{cases} 1, & \text{if } \boldsymbol{f}(x,y) = a \text{ and } \boldsymbol{f}(x+\Delta_x, y+\Delta_y) = b \\ 0, & \text{otherwise} \end{cases} \tag{8.1}$$

where x and y are pixel positions, and a, b their values in the image $\boldsymbol{f}$. Typically, the co-occurrence matrix is normalized to probabilities p:

$$p(a,b) = \frac{\mathrm{GLCM}(a,b)}{\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1}\mathrm{GLCM}(i,j)} \tag{8.2}$$

where N_g denotes the number of different intensity levels. In this way, Haralick proposed several measures for texture characteristics - among those also a measure for homogeneity, defined as:

$$h = \sum_{a=0}^{N_g-1}\sum_{b=0}^{N_g-1} \frac{1}{1+(a-b)^2}\, p(a,b). \tag{8.3}$$

Here, $p(a,b)$ is the entry in the normalized co-occurrence matrix p at position (a,b). In our example in Fig. 8.2 there are 5 different intensity values in the range of $[0,..,4]$, i.e. $N_g = 5$. It is obviously straightforward to extract the size of the detected objects as well as their homogeneity.

8.1.2 Experiments and Evaluation

In order to evaluate the validity of the clinical assumptions, we made several experiments. First, we analyzed the expressiveness of size and homogeneity in a standard scenario, i.e.

Table 8.1: Evaluation of clinical assumptions for classification: Homogeneity and size.

		Predicted
	Wilms' tumor	Nephroblastomatosis
Wilms' tumor	0.83 ± 0.08	0.17 ± 0.08
Nephroblastomatosis	0.19 ± 0.07	0.815 ± 0.07

the combination of both features. Afterwards, we evaluated the classification performance when either homogeneity or size are neglected.
Let us now analyze if the amount of information contained in the two properties size and homogeneity is sufficient to separate both diseases. We extract these two features from all images and use it for classification. For this purpose we randomly select 54 out of our 148 Wilms' tumors; see Sec. 4.4. We then subdivide these into 27 test and training data sets again by chance. We proceed analogously with nephroblastomatosis data sets. Since the diffuse anaplastic and necrotic subtypes are under-represented, we made sure that they occur exclusively in the test sets. From each of our data sets we draw the middle slice of the annotated tumor region and train a random forest classifier (Sec. 3.2) to distinguish these two classes with 3-fold cross validation. We repeat this procedure 5 times and calculate the average accuracy at the end.
The results of this procedure are shown in Tab. 8.1. It turns out that the average accu-

Table 8.2: Classification of nephroblastomatosis based on objects' size.

		Predicted
	Wilms' tumor	Nephroblastomatosis
Wilms' tumor	0.79 ± 0.12	0.21 ± 0.12
Nephroblastomatosis	0.22 ± 0.09	0.78 ± 0.09

racy of 0.82 indicates that homogeneity and size are valuable properties to distinguish a nephroblastoma from its precursor lesion. However, $\approx 18\%$ of the images are misclassified. Thus, this result is not sufficient to build clinical decisions on. Especially the high false positive rate of 0.17 is dangerous: In all these cases, Wilms' tumors are classified as nephroblastomatosis. This might result in a lack of chemotherapy, and no sufficient treatment of this malignant abdominal mass.
However, in order to investigate the decisiveness of each feature, we also made a classification based on solely one of these features - either size or homogeneity. Table 8.2 shows

Table 8.3: Classification of nephroblastomatosis based on objects' homogeneity.

		Predicted
	Wilms' tumor	Nephroblastomatosis
Wilms' tumor	0.71 ± 0.16	0.29 ± 0.16
Nephroblastomatosis	0.37 ± 0.10	0.63 ± 0.10

the results in terms of classification accuracy when homogeneity is not considered. The size of the objects is remarkably reliable.
Although the classification average accuracy drops to 0.79, its quality is comparable to the previous setting. Obviously, homogeneity alone is not decisive, see Tab. 8.3. The average classification accuracy reduces dramatically to 0.67.
We cannot make any judgement whether nephroblastomatosis are always homogeneous objects. However, we conclude from the results that in some cases this property also applies to Wilms' tumors. Thus this feature is not meaningful and homogeneity is no separating criterion between nephroblastomatosis and Wilms' tumors.

8.1.3 Summary

In this section, we showed that the current clinical assumptions of nephroblastomatosis are not sufficient for classification. Although these basic assumptions are mostly correct, they do not provide a reliable basis for a treatment decision.
Especially homogeneity is not decisive: Our experiments indicate, that nephroblastomatosis show this property, but not exclusively - due to their triphasic nature, Wilms' tumors also can appear homogeneous.
Nevertheless, it turns out that the size of an object is a first indication for a nephroblastomatosis. However, this characteristic alone is not a reliable separating criterion: Wilms' tumors can be small, too. Thus, we investigate in the next section the application of more visual texture properties to the classification procedure.

8.2 Robust Classification of Nephroblastomatosis

The experiments in the last section gave us important insights about current clinical practice: First, we verified that the size of an abdominal mass is helpful in the distinction of Wilms' tumors and their precursor nephroblastomatosis. Second, the widely accepted assumption about increased homogeneity of nephroblastomatosis in comparison to nephroblastoma is not decisive.
However, homogeneity is only one out of many textural characteristics. Haralick et al. [73] suggested a wide range of texture measures establishing the basic assumption that gray-level co-occurence matrices contain all available textural information of an image.
These second order texture features are extensively used in recent years in the area of medical image analysis to diagnose and differentiate cancer [161, 193, 199]. In the following, we include more of these texture information to improve the classification performance.

8.2.1 Feature Extraction

We do not make any assumptions about the visual appearance of nephroblastomatosis and follow the basic assumption that all textural information of an image is contained in co-occurrence matrices. A straight forward idea is now to include all available texture properties. Thus, we extract all of Haralicks' texture features (Tab. 8.4) as well as

Table 8.4: Definition of Haralicks' texture features [73]. IMC: Information Measure of Correlation.

Name	Definition
Angular Second Moment	$h_1 = \sum_a \sum_b p(a,b)^2$
Contrast	$h_2 = \sum_n n^2 \sum_a \sum_b p(a,b), \lvert a-b\rvert = n$
Correlation	$h_3 = \frac{\sum_a \sum_b (ab)p(a,b) - \mu_x \mu_y}{\sigma_x \sigma_y}$
Joint Variance	$h_4 = \sum_a \sum_b (a-\mu)^2 p(a,b)$
Inverse Difference Moment	$h_5 = \sum_a \sum_b \frac{1}{1+(a-b)^2} p(a,b)$
Sum Average	$h_6 = \sum_k k\ p_{x+y}(k)$
Sum Variance	$h_7 = \sum_k (k-h_8)^2 p_{x+y}(k)$
Sum Entropy	$h_8 = \sum_k p_{x+y}(k) \log\{p_{x+y}(k)\}$
Entropy	$h_9 = \begin{cases} -\sum_a \sum_b p(a,b)\log(p(a,b)) & p(a,b) \neq 0 \\ 0 & \text{otherwise} \end{cases}$
Dissimilarity	$h_{10} = \sum_k k\ p_{x-y}(k)$
Difference Variance	$h_{11} = \sum_k (k-s_1)^2 p_{x-y}(k)$
Difference Entropy	$h_{12} = -\sum_k p_{x-y}(k) \log(p_{x-y}(k))$
IMC 1	$h_{13} = \frac{\mathrm{HXY} - \mathrm{HXY1}}{\max\{\mathrm{HX},\mathrm{HY}\}}$
IMC 2	$h_{14} = \sqrt{1 - \exp(-2(\mathrm{HXY2}\text{-}\mathrm{HXY}))}$

Table 8.5: Definition of Soh and Tsoutsalis texture features [160].

Name	Definition
Autocorrelation	$s_1 = \sum_a \sum_b p(a,b)ab$
Cluster Prominence	$s_2 = \sum_a \sum_b (a+b-\mu_x(a)-\mu_y(b))^4 p(a,b)$
Cluster Shade	$s_3 = \sum_a \sum_b (a+b-\mu_x(a)-\mu_y(b))^3 p(a,b)$
Cluster Tendency	$s_4 = \sum_a \sum_b (a+b-\mu_x(a)-\mu_y(b))^2 p(a,b)$
Maximum Probability	$s_5 = \max p(a,b)$

all measures of Soh and Tsoutsalis [160]; see Tab. 8.5. Here, $p(a,b)$ is the normalized co-occurrence matrix and

- marginal row probability: $p_x(a) = \sum_b \mathrm{GLCM}(a,b)$
- marginal column probability: $p_y(b) = \sum_a \mathrm{GLCM}(a,b)$
- mean intensity value: $\mu = \sum_a p(a)a$
- standard deviation: $\sigma = \sqrt{\frac{1}{N_g-1}\sum_a (a-\mu)^2}$
- gray level sum distribution:
  $p_{x+y}(k) = \sum_a \sum_b p(a,b)$, where $k = a+b$, and $k = 2, \ldots, 2N_g$
- gray level difference distribution:
  $p_{x-y}(k) = \sum_a \sum_b p(a,b)$, where $k = a-b$, and $k = 0, \ldots, N_g - 1$
- $\mathrm{HX} = -\sum_a p_x(a) \log(p_x(a))$
- $\mathrm{HY} = -\sum_b p_y(b) \log(p_y(b))$
- $\mathrm{HXY} = -\sum_a \sum_b p(a,b) \log(p(a,b))$
- $\mathrm{HXY1} = -\sum_a \sum_b p(a,b) \log(p_x(a)p_y(b))$
- $\mathrm{HXY2} = -\sum_a \sum_b p_x(a)p_y(b) \log(p_x(a)p_y(b))$

Since the intuitions of these definitions are not obvious, we list interpretations [178] for most of these mathematical terms in Tab. 8.6.

We want to identify the set of textural measures that is decisive for classification of Wilms' tumors and nephroblastomatosis. Fortunately, we can make use of an inherent property of random forests: their ability to decide whether a feature is decisive. Important features will be used in a higher hierarchical level, than less expressive characteristics. Hence, we can identify separating criteria by following the decision structure of a random forest.

Table 8.6: Intuition of textural measures.

Name	Intuition
Angular Second Moment (Energy)	Measures monotonic transitions of intensity values. Higher textural uniformity results in a higher energy.
Autocorrelation	Amount of coarseness/fineness of the occurring textures.
Cluster Prominence	Measures skewness and asymmetry of the co-occurrence matrix.
Cluster Shade	Measures skewness and uniformity of the co-occurrence matrix.
Cluster Tendency	Amount of grouped voxels with a similar intensity value.
Contrast	Variations in an image.
Correlation	Amount of linear dependencies among neighboring gray values.
Joint Variance	Heterogeneity or intensity level variability of an image.
Inverse Difference Moment	Homogeneity in an image.
Maximum Probability	Denotes the most dominant pair occurring in the image.
Sum Average	Mean intensity level sum distribution.
Sum Variance	Dispersion of the sum average measure.
Sum Entropy	Disorder in relation to the sum average measure.
Entropy	Amount of randomness of textural patterns.
Dissimilarity	Measures the mean of the gray level difference distribution. The larger the value, the more dissimilar are neighboring pixels.
Difference Variance	Similar to joint variance but related to the mean intensity level sum distribution. Intensity level pairs deviating from the mean value are punished.
Difference Entropy	Measures randomness in relation to the gray level difference distribution.

8.2.2 Experiments

Let us begin with our default scenario: The original data from our classification benchmark; see Sec. 4.4. While collecting the data, we made sure that no extremely degraded data set is included. Hence, we can assume these imaging data as moderately disturbed while no massive noise is observable.
In a first step, we use all textural features as well as objects' size for classification. We now exploit the hierarchical structure of decision trees and extract the ordering by importance for all included features. Afterwards, we are able to drop features with minor influence to the final classification result.
In the next step, we add different amounts of noise, i.e. we disturb the data. One might think that this is no realistic scenario, but please note that we excluded all data from our benchmark where we were not satisfied with the image quality. Typically, the main reason for exclusion was noise.
Thermal noise in the patient is the main reason for noise in MRI data [75]. It is present as Gaussian noise in the original signal. During the reconstruction, i.e. an inverse discrete Fourier transform, the measurements are contained in both real and imaginary channels. In order to construct the final image, the two channels (with two independent Gaussian noise distributions; see (2.10)) are squared and summed up. In the end, the MRI image is the square root of the previous computation. Hence, MRI data typically suffer from Rician noise [68].
However, Nowak [129] showed that Rician noise is approximated by Gaussian noise when the signal-to-noise ratio is large enough. Hence, we can treat the MRI noise as white noise in our case.
We start with the calculation of all 19 features as well as all mean values and standard deviations. We use this information to train a bagged random forest classifier with 300 ensemble learners, i.e. decision trees. Proceeding analogously to our previous approach, we evaluate these features on five randomly selected data sets and 3-fold cross validation. Table 8.7 shows the result of the classification in terms of accuracy. It turns out that this additional information dramatically improves the classification performance to an accuracy of 0.93 - especially the false positive rate of 0.17 has been more than halved to a moderate rate of 0.07.
The hierarchical structure of random forests allows us to identify the most important textural information for nephroblastomatosis classification, namely size, information measure correlation 1 and 2, cluster prominence, sum entropy, dissimilarity, maximum probability, energy and autocorrelation. Surprisingly, the feature of homogeneity is negligible when the above information is given: This aspect is already covered by the dissimilarity measure. Table 8.8 lists the results based on the aforementioned nine features selected by their importance according to the structure of the random forest. Obviously, these features contain most information necessary for classification and the amount of required proper-

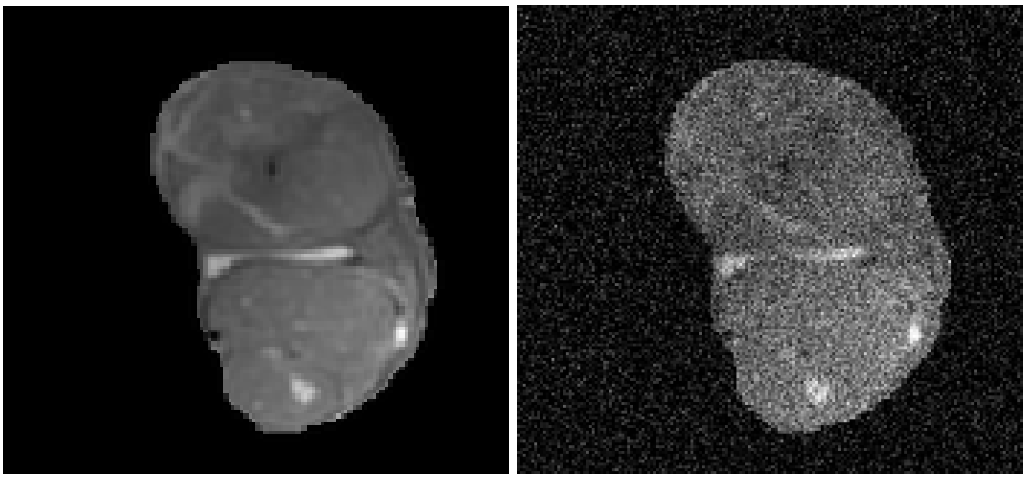
Table 8.7: Classification outcome with all Haralick and Soh textural information.

		Predicted
	Nephroblastoma	Nephroblastomatosis
Nephroblastoma	0.93 ± 0.05	0.07 ± 0.05
Nephroblastomatosis	0.06 ± 0.02	0.94 ± 0.02

Table 8.8: Classification result with optimized feature selection.

	Predicted	
	Nephroblastoma	Nephroblastomatosis
Nephroblastoma	0.93 ± 0.06	0.07 ± 0.06
Nephroblastomatosis	0.06 ± 0.03	0.94 ± 0.03

ties is substantially reduced. This beneficial trade-off results not only in less computation time for feature extraction but also in dimensionality reduction and less overfitting.
In a real clinical scenario, it is of high importance that classification methods are robust to noise. In order to simulate disturbed data, we add Gaussian noise with zero mean and standard deviation $\sigma = 0.02$; see Fig. 8.3.



(a) (b)

Figure 8.3: Exemplary image from our classification data set with additional noise. (a) Input image of a nephroblastoma, (b) Input image with additional Gaussian noise, $\sigma = 0.02$.

The results in Tab. 8.9 show, that the included textural information is robust to noise: Although a lot of textural information is lost due to the noise, the classification remains reliable.

Table 8.9: Classification result with optimized feature selection and moderate noise disturbances, $\sigma = 0.02$

	Predicted	
	Nephroblastoma	Nephroblastomatosis
Nephroblastoma	0.89 ± 0.12	0.11 ± 0.12
Nephroblastomatosis	0.10 ± 0.05	0.90 ± 0.05

8.2.3 Summary

We showed that the classical but time proven Haralick texture features can serve as a reliable source of information for classification of nephroblastomatosis and Wilms' tumors. In a first step, we included all textural information of Haralick as well as Soh and Tsatsoulis.
It turns out, that we can restrict the information necessary for classification and reduce the amount of computational load while preserving the classification performance.
Afterwards we showed that our proposed combination of textural measures is robust to noise and provides in all tested scenarios a reliable classification.

8.3 Conclusions

Since decades, clinicians assumed homogeneity and size to be identifying criteria of nephroblastomatosis. We used gray level co-occurrence matrices to extract Haralick's texture measure of homogeneity. Based on these information, we trained a random forest classifier and evaluated the influence of both criteria to the final classification decision. It turns out that homogeneity is not a decisive property but objects' size is a first indication of Wilms' tumor precursor lesion.
In a next step, we investigated the expressiveness of other textural measures based on co-occurrence matrices. In order to identify separating criteria, we trained again a random forest classifier allowing us to extract the most important features for classification, namely size, information measure correlation 1 and 2, cluster prominence, sum entropy, dissimilarity, maximum probability, energy and autocorrelation. This combination allows for a reliable classification with low false positive rate: We improved the classification performance of nephroblastomatosis and Wilms' tumors by $\approx 11\%$.
All in all, we demonstrated in this chapter that the distinction between nephroblastomatosis and nephroblastoma is not as trivial as previously assumed. However, we were able to solve this problem and proposed further intuitive features that make the distinction much more reliable. This significantly reduces the risk of misdiagnosis and thus minimizes the medical burden on affected children.

9 Subtype Prediction of Wilms' Tumors

"Prediction is very difficult, especially if it's about the future."

– Nils Bohr

Contents

Wilms' tumor is a solid tumor, consisting mainly of three types of tissue: blastema, epithelium and stroma [182]. In Europe, diagnosis and therapy follow the guidelines of the International Society of Pediatric Oncology (SIOP) [67, 88]; see Sec. 2.5.3.
One of the most important characteristics of this therapy protocol is a preoperative chemotherapy. During this therapy, the tumor tissue changes, and a total of nine different subtypes can develop [67]. Depending on this and the local stage, the patient is categorized into one of the three risk groups (low-, intermediate-, or high-risk patients) and further therapy is adapted accordingly; see Sec. 2.5.3. Of course, it would be of decisive importance for therapy and treatment planning to determine the corresponding subtype as early as possible. It is currently not known how this can be achieved.
However, there are no research results in this direction so far. Unfortunately, diffusion-weighted MR images are not yet recorded as standard. Due to a relatively low incidence of this disease, it is also difficult to sensitize the clinical staff in this direction. However, a T_2 sequence is part of the therapy protocol and always recorded - even if there are no parameter specifications.

In the US therapy and treatment protocol follow a different strategy without a preoperative chemotherapy and direct tumor extraction. Their histological data allows to approximate the occurrence of Wilms' tumors subtypes at the time of diagnosis before any treatment. Clinicians assume - based on subtype distributions before and after chemotherapy - that mainly blastemal tissue is destroyed during this phase of therapy, see Fig. 9.1. It is not possible to validate this assumption, as there is currently no possibility to determine the histological components without a biopsy, exclusively based on imaging data.
Obviously, the task is very challenging and there are two main questions that arise and need to be answered:

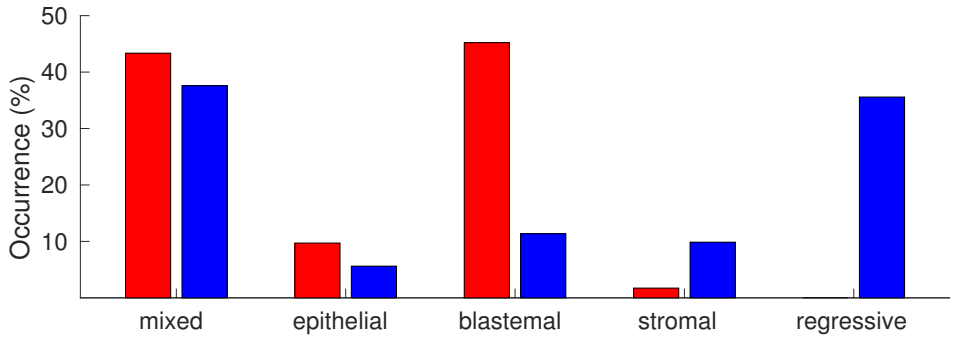


Figure 9.1: Subtype distribution of the most common Wilms' tumor subtypes without (red) and with (blue) pre-operative chemotherapy. These distributions indicate a change in tumor structure during chemotherapy.

- Do T_2 sequences contain enough information about tumor development for a classification before any treatment?
- Can we confirm the clinical assumption about the effect of chemotherapy to blastemal tissue?

We want to close this gap and address the considered unsolvable problem of subtype determination prior to chemotherapy based on simple but standard T_2. First, we propose a way to build a visual representation of different tumor entities in Sec. 9.1. Afterwards, we use this visual description for classification. We evaluate our approach and are able to draw first conclusions.

9.1 Visual Representation of Subtypes

The first step after diagnosis of nephroblastoma is a chemotherapy. At this stage, all efforts are made to avoid a spreading of the tumor to other organs at all costs: A biopsy is not possible. Since the course of therapy is monitored, the initial imaging and the final subtype are known, but the original histological condition of the tumor is unknown.
Unfortunately, this is a major limitation as chemotherapy can have completely different effects on various people. Many factors influencing the development of the tumor as well as the amount of their effect are unknown. Whether and with which probability a tumor of a certain subtype develops into a different type under the use of chemotherapeutic agents is currently unknown. Although there are suspicions that, for example, blastemal dominant tumors show a higher response to chemotherapy, the probability of which subtype developing from it is just as unknown as the conditions under which blastemal regions persist. So we do not have any information about the original condition and cannot incorporate any prior knowledge.
Most classification methods are either based on separating objects based on given points of interest and their features, or learn the distribution of objects and their properties directly from the data. Since the latter class of approaches requires large amounts of data to extract necessary information while Wilms' tumors are relatively rare, it is not applicable in this situation.
Since we obviously cannot use any prior knowledge and it is therefore difficult to identify points of interests, there are now tree possibilities:

- We make an educated guess based on information useful for other classification problems to extract points of interest.
- We do not use points of interest and utilize all pixels of the objects for classification.
- We extract image sections and use them to approximate a visual representation of the tumor as a whole.

It is well known that image edges contain very important information about objects. Typically, however, they are not sufficient, and more information about object-specific patterns and textures is essential. Unfortunately, it is not yet known whether and to what extent subtypes differ in this respect.
Assuming that each pixel would be used for classification, there are tree major problems: Firstly, the number of dimensions of the classification problem increases with the number of pixels and thus also the computational complexity. Second, it is not scale invariant - objects of the same class but of different size are difficult to match. And last but not least, the use of all pixels results in a massive overfitting: Since the number dimensions in which the classification is done is too large, the individual training instances are memorized; see Sec. 3.2.1.
We decided to choose a mixture of both previous approaches. Since we have no prior knowledge, we extract parts of the image, so-called patches, and use them to generate a visual representation of the tumors. On the one hand we can incorporate important information about e.g. edges, on the other hand we do not specify exact points of interest. We use a bag of visual words model to provide a robust visual representation of the available image information - always with the constraint that the amount of data is limited [197].

9.1.1 A Bag of Visual Words Model

In case of a limited amount of data, bag of visual words models are a popular method to create a robust representation of the contained image information. Intuitively, each object is a composition of its subcomponents, i.e. its *visual words*. The key idea of these models is now to classify objects based on the frequency histogram of their visual words. Generally, each classification with bags of visual words consists of several steps:

1. Extract the local features, i.e. the visual words.
2. Aggregate a codebook of visual words via clustering of the extracted features.
3. Generate the histogram of the visual words for all training images.
4. Train a classifier to separate the images based on their visual representations, i.e. their histograms.

Typically, features are extracted at points of interest identified via prior knowledge. As we mentioned before, we cannot incorporate prior information and decided therefore to extract patches. First, we subdivide the images into 7×7 patches.
We do not have any information about the average size, orientation or shape of each subtype before chemotherapy, such that the features used for classification have to be invariant under image scaling, rotation, and small changes. In addition, MRI sequences are acquired with different parameter settings resulting in an uniform change of intensity levels for each sequence. SURF features (Sec. 3.2.4) are fortunately invariant under all these conditions and therefore well suited for our purpose [15]. Hence, we calculate the SURF features of the central pixel of all patches, i.e. every 8th pixel position.

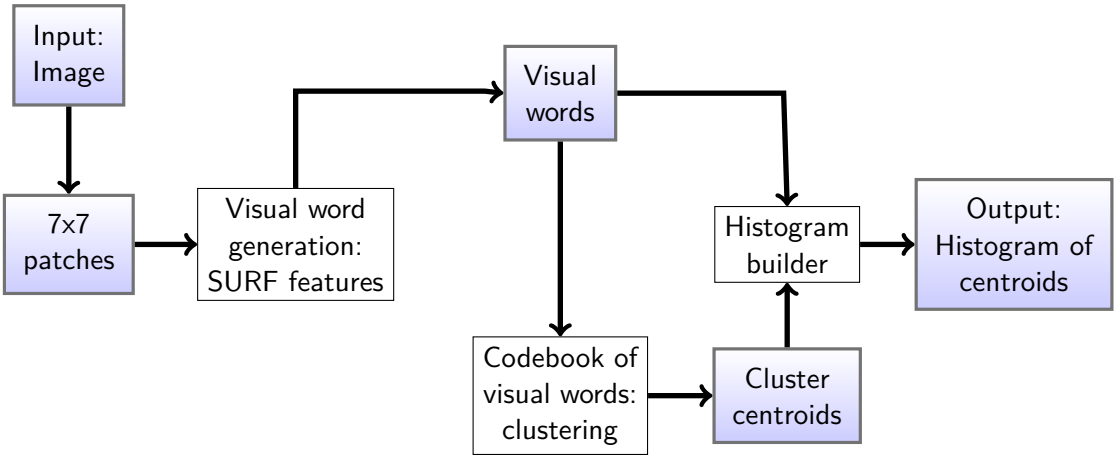


Figure 9.2: Schematic view of the bag of visual words model that we use for subtype prediction.

The next step is to cluster all extracted features to generate a codebook of visual words. Here, the cluster centroids represent the vocabularies of the visual dictionary. We follow [185] and use k-means clustering to group the visual words into vocabularies.
Using these visual vocabularies, we can now aggregate the frequency histograms for the SURF features of each training and test image. We use this information to train a bagged random forest classifier [27].

9.1.2 Experiments

A Wilms' tumor consists of the tissue types stroma, epithelium, and blastema [182]. Depending on the chosen therapy strategy, the subtypes are distributed differently, see Fig. 2.14. During the preoperative chemotherapy, various subtypes emerge, some of which differ dramatically in their prognosis. In the following we consider the standard group of intermediate risk patients. This consists of mainly regressive, epithelial dominant, stromal dominant, and mixed (none of the tissue types predominates) tumors. Since the blastemal dominant type has the worst prognosis, we also include it. We evaluate how far we can get in subtype determination with simple but standard T_2 sequences. Since this problem is much more complex than the distinction between nephroblastoma and nephroblastomatosis, we need more data. Therefore, we select one slice from each annotated tumor from the lower third of the annotation, one from the upper third and the middle slice. In this way we generate a total of 54 images of a blastemal dominant tumor, 150 of a regressive tumor, 87 of a mixed tumor, 84 of a stromal dominant tumor and 51 of an epithelial dominant tumor.
Depending on the classification problem, we always take as many images as there are in the smaller class and divide them randomly into training and test sets. In this way we ensure that the results are not aimed at the frequency of the images but only at the discrimination. Then we calculate the visual vocabulary for each data set to generate a bag of visual words of 100 vocabularies.
With this information we then train a random forest with 300 ensemble learners and 3-fold cross validation. We repeat this process 5 times, analogous to the differentiation of nephroblastomatosis, including the newly generated training and test set. Here, we optimize the size of the vocabulary on the training set and select a value from the interval $[10, 100]$.

We compare all selected subtypes with all others in Fig. 9.3. Our results are strictly above the chance level (dashed line) while average accuracy of regressive is 0.70, epithelial dominant 0.72, stromal dominant 0.66, mixed 0.67, and blastemal dominant 0.64. This indicates that we are on the right way and that it should be possible to distinguish these subtypes based on imaging data.
There are also several cases where our classification is surprisingly accurate. The accuracy of the distinction between regressive and epithelial dominant subtypes is 0.80. This leads to the following conclusions: 1. tumors that are epithelial dominant prior to chemotherapy are less likely to regress than those that are rich in stroma or blastemal tissue. This coincides with subtype distributions before and after chemotherapy. 2. epithelial areas can be distinguished from other types of tissue by visual features.
Furthermore, the differentiation between regressive and mixed subtypes is relatively accurate with 0.73. This allows conclusions similar to those of the epithelial type. In addition, the epithelial dominant subtype is also well distinguishable from the stromal dominant one, i.e. classification accuracy of 0.7.

We also tried to use neural networks to solve our classification problem. Unfortunately it turned out that we do not have a sufficient amount of data to re-train enough layers of a pretrained network. Therefore, all our attempts with neural networks showed low performance.

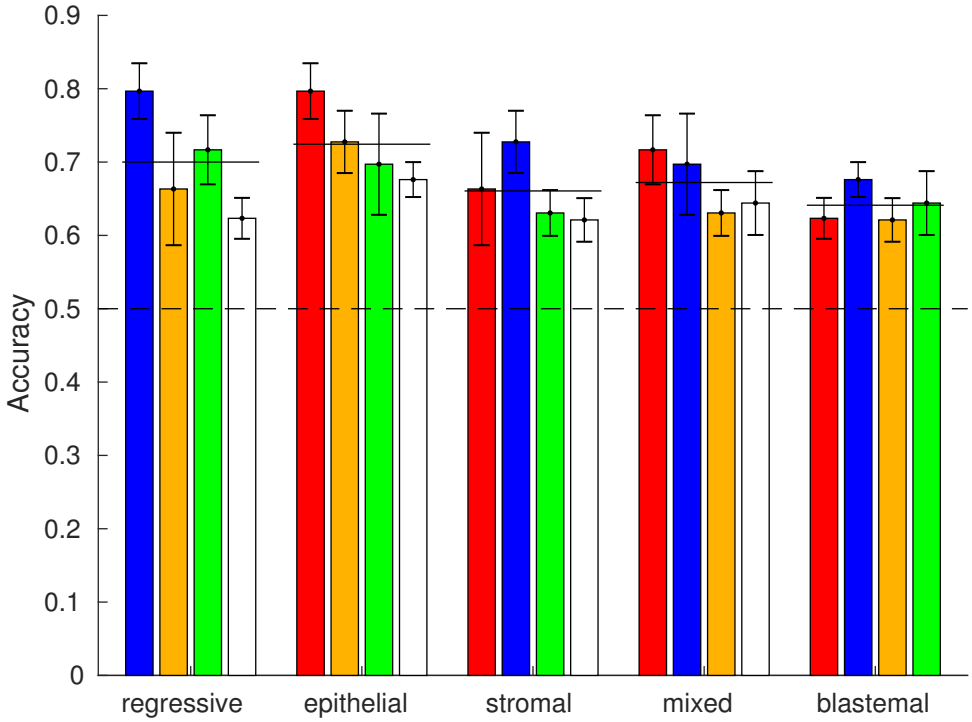


Figure 9.3: Evaluation results showing mean and standard deviation of between-class classification accuracy for regressive, epithelial, stromal, mixed and blastemal subtypes. Mean performance is indicated with black lines. The dashed line marks the chance level. red: regressive, blue: epithelial, yellow: stromal, green: mixed, white: blastemal.

We ensured that the main parameter settings of images included in our data set are as similar as possible. However, several parameters differ dramatically in many cases. Since these cannot be compensated, the data is unfortunately not completely comparable and a considerable parameter noise is present. We firmly believe that the classification would improve significantly if this kind of noise in the data were lower. We therefore hope that in the near future a standardization of MRI sequences will be established in the medical area.

9.2 Summary and Conclusions

The prediction of subtype evolution under chemotherapy is a major but considered unsolvable problem in Wilms' tumor treatment planning. Depending on the evolved subtype an early adaptation of the therapy strategy seemed to be a desirable but not realizable task. In this chapter, we addressed this problem and made first steps towards a subtype prediction.

Although we can only show a proof of concept that it is fundamentally possible to estimate the development, the rewards for research in this direction might be immense.

Even though the imaging is not standardized and therefore shows a high parameter noise, there are still visual features that allow a distinction. In all our experiments, the classification accuracy shows performance above the chance level. In some cases, the results clearly indicate that a prediction is within possible range. We hope that we can foster more researchers to investigate these challenge.

10 Summary and Outlook

"Nature will bear the closest inspection. She invites us to lay our eye level with her smallest leaf, and take an insect view of its plain."

– H. D. Thoreau

Contents

10.1 Summary

In this work we dealt with Wilms' tumors, the most frequent malignant kidney tumor in childhood. We approached this disease from the perspective of medical image processing and dedicated our interest to three overarching goals: improvement of therapy planning, reduction of false diagnoses, and prediction of the course of disease.
First, we addressed the possibilities to optimize the therapy planning. We considered the important information of tumor volume as it has a massive impact on the treatment schedule. We compiled a multi-sequence benchmark data set of 17 patients before and after chemotherapy, respectively. This contribution allowed us to investigate several research aspects. Firstly, it provided us with the necessary information to evaluate human expert annotations. We evaluated inter-rater variability of human experts, i.e. how reliable an annotation of a single human is and how much humans vary in their decisions about tumor and non-tumor areas.
Unfortunately, we found that annotations of single domain experts are not as reliable as previously assumed: While the differences between human raters are acceptable before chemotherapy, the situation changes after this milestone in therapy planning: inter-rater variability of human annotators increases dramatically after chemotherapy. We observed that human bias has more impact and is remarkably present in their annotations in the later stages of treatment. In addition, tumor delineation becomes more challenging during the curse of the therapy.
We further investigated the differences in the determined volume of single raters and their consensus truth. It turned out that the deviations cannot be neglected and might influence the final decisions of follow up treatment.
This lead us to the conclusion that no single domain expert should define tumor extensions after chemotherapy - the present variability even between radiologists is remarkable and can heavily influence treatment decisions.
The second investigation we could address with our segmentation benchmark data set

is the analysis of the current clinical practice for approximating tumor volume with an ellipsoid shape. We proved that the gold standard of volume determination is erroneous and volumes of nephroblastoma should not be approximated by an ellipsoid shape - the approximation is too coarse: Its surface is complex and oversimplifications lead to large approximation errors. We conclude therefore that a reliable and reproducible annotation of tumor outlines is essential for treatment planning.
We also used our data set to evaluate out-of-the-box methods for Wilms' tumor segmentation and develop a semi-automatic method for these kind of applications. We implemented a robust and flexible interactive segmentation method. In order to show its flexibility, we not only applied it to its basic target of nephroblastoma segmentation but also to standard image and video sequences. In the end, we evaluated our approach together with a wide range of fully-automatic segmentation methods on our benchmark data set. It turns out, that fully automatic approaches typically oversegment the tumor and are therefore not as suited as our method for this kind of segmentation task.
To demonstrate the topicality of our method, we employed it also as a postprocessing step for a deep learning approach to brain tumor segmentation. This scenario highlights the strength of the model as well: the accuracy of the final result is significantly improved.

In the second part of this thesis, we turned our attention towards the reduction of false diagnosis. Wilms' tumors are especially similar to their precursor lesion, the nephroblastomatosis. A reliable and trustworthy separation of these two diseases is essential: while Wilms' tumors are malignant and patients' need to be exposed to a chemotherapy and a possible irradiation, nephroblastomatosis is a benign abdominal mass. We compiled another data set of MRI sequences to address this classification problem. One main aspect of this data set is its ability to provide data to verify current assumptions about visual appearance and to identify sufficient features for their distinction.
We followed this idea and first analyzed the current clinical practice to identify nephroblastomatosis - namely size and homogeneity of the object to be examined. We showed that these assumptions are not decisive: although nephroblastomatosis tend to be smaller than Wilms' tumors, homogeneity is no valuable measure in this context. In principle, nephroblastoma as well as its precursor lesion can both be homogeneous and a stringent separation is on account of this not possible. Starting from this observations, we investigated the visual appearance of nephroblastomatosis. We extracted several texture features and revealed a collection of properties that reliably separate Wilms' tumors and its precursor lesion. We used these information to dramatically improve the classification accuracy allowing for a trustworthy identification.
Our classification data set also paves the way for first attempts of the prediction of the tumor evolution under chemotherapy. During the preoperative chemotherapy, Wilms' tumors react in various aspects to the applied chemotherapeutic agents. While some of them shrink and downgrade to necrotic masses, others evolve to highly dangerous subtypes. In order to adopt patients' treatment as early as possible, the prediction of subtype evolution during chemotherapy is highly desirable. However, this major challenge is considered an unsolvable problem in Wilms' tumor treatment planning.
We nevertheless addressed this problem in order to analyze to what extent this question is really impossible to answer. It is not yet known which data may be helpful to predict the development of the subtype. At the same time, it is a standard in the therapy protocol that T_2 sequences of the patient are recorded when a Wilms' tumor is diagnosed.
We employed the data set we provided previously and created a visual representation for

each of the subtypes that we used to classify these. Although we can only show a proof of concept that it is fundamentally possible to estimate the development, the rewards for research in this direction might be immense.
Even though the imaging is not standardized and therefore shows a high parameter noise, there are still visual features that allow a distinction. In all our experiments, the classification accuracy shows performance above the chance level. In some cases, the results clearly indicate that a prediction is within possible range. We hope that we can foster more researchers to investigate these challenge.

10.2 Outlook

The basis of any medical image analysis is the existence of a sufficient amount of data that reflects a realistic distribution of the disease under consideration. Therefore, one of the most important contributions of this work is the preparation of images for Wilms' tumors. Obviously, however, we have only paved the way - until the results of image processing in this area are of sufficient accuracy to automate diagnosis and therapy planning, much more data is needed. Nevertheless, the next step is not only to provide more data, but also to standardize the image acquisition.
Without massive parameter noise, it is possible to estimate the characteristics of Wilms' tumors more accurately: These information can be incorporated as texture features for segmentation or classification tasks of nephroblastoma.

We have designed our semi automatic segmentation approach to be as simple and intuitive as possible. Also, we never trained the edge detector for medical data - which is a double-edged sword. On the one hand, our approach has no bias to specific problems. On the other hand, he does not learn specific properties of medical data and Wilms' tumors in particular. Re-training would therefore be one way to increase segmentation performance. However, the problem with the available amount of data exists here as well.

It is also straight-forward to include further information in the data term of the energy functional of our segmentation approach. Finding the optimal features, however, is usually a challenge not to be underestimated. In order to avoid the time-consuming manual design, it is an option to use representation learning. Of course, the amount of data available for Wilms' tumors is not sufficient to extract those features directly. However, similar data often share comparable properties. It should be possible to identify suitable features based on MRI data of other cancer instances, e.g. liver or brain tumors.

Of course, an exact visual representation of the individual subtypes of nephroblastoma is an important ingredient in the prediction of Wilms' tumor development. We believe that especially the epithelial dominant class is a promising candidate for improvement. We are convinced that at least individual subtypes can be identified with certainty when the visual representation can rely on data without massive parameter-noise. Most importantly, we are convinced that this research will maximize the chances of survival of the affected children. If it is possible to detect especially blastemal dominant tumors (after chemotherapy) early, the therapy can be adapted much earlier and hopefully the recovery process of the child can be improved.

10.3 Closing Words

In recent years, the interest in the area of deep learning has increased dramatically. Many problems which can be specified exactly, were nearly solved. This has led to a general feeling that any question can be addressed.
Of course, the improvements through neural networks are remarkable. However, it should not be forgotten that the prerequisite for these approaches is a sufficiently large amount of data for the respective task. In many areas, especially in medical image processing, the amount of available data is often a limitation. Not only ethical concerns or privacy constraints, but also non-standard acquisition procedures make the situation difficult. This leads to situations where overfitting in particular is a problem - especially since this aspect is not considered an issue by some, but rather a feature.
We are deeply convinced that classical image processing based on mathematically well-founded knowledge can make a significant contribution in this area. Especially in cases where the statistical approach of deep learning fails, its strengths reside.

A List of Publications

2019

Robustness and Generalization of Brain Tumor Segmentation Models

Sabine Müller, Joachim Weickert, Norbert Graf
Technical Report.

Abstract. In this work we address the generalization behavior of deep neural networks in the context of brain tumor segmentation. While current topologies show an increasingly complex structure, the overall benchmark performance does improve negligibly. In our experiments, we demonstrate that a well trained UNet shows the best generalization behavior and is sufficient to solve this segmentation problem. We illustrate even more, why extensions of this model cannot only be pointless but even harmful in a realistic scenario. We suggest also two simple modifications (that do not alter the topology) to further improve its generalization performance.

Wilms' tumor in childhood: Can pattern recognition help for classification?

Sabine Müller, Joachim Weickert, Norbert Graf
In Proc. 23rd Conference on Medical Image Understanding and Analysis (MIUA 2019, Liverpool, UK, July 2019). Communications in Computer and Information Science, Springer, 2019.

Abstract. Wilms' tumor or nephroblastoma is a kidney tumor and the most common renal malignancy in childhood. Clinicians assume that these tumors develop from embryonic renal precursor cells - sometimes via nephrogenic rests or nephroblastomatosis. In Europe, chemotherapy is carried out prior to surgery, which downstages the tumor. This results in various pathological subtypes with differences in their prognosis and treatment.
First, we demonstrate that the classical distinction between nephroblastoma and its precursor lesion is error prone with an accuracy of 0.824. We tackle this issue with appropriate texture features and improve the classification accuracy to 0.932.
Second, we are the first to predict the development of nephroblastoma under chemotherapy. We use a bag of visual model and show that visual clues are present that help to approximate the developing subtype.
Last but not least, we provide our data set of 54 kidneys with nephroblastomatosis in conjunction with 148 Wilms' tumors.

Benchmarking Wilms' Tumor in Multi-Sequence MRI Data: Why Does Current Clinical Practice Fail? Which Popular Segmentation Algorithms Perform Well?

Sabine Müller, Iva Farag, Joachim Weickert, Yvonne Braun, Andreas Hötker, André Lollert, Jonas Dobberstein, Norbert Graf
Journal of Medical Imaging, Vol. 6, No. 3, Paper 034001, July 2019.

Abstract. Wilms' tumor is one of the most frequent solid and malignant tumors in childhood. Accurate segmentation of tumor tissue is a key step during therapy and treatment planning. Since it is difficult to obtain a comprehensive set of tumor data of children, there is no benchmark so far allowing evaluation of the quality of human or computer-based segmentations. The contributions in our paper are threefold: (i) We present the first heterogeneous Wilms' tumor benchmark data set. It contains multi-sequence MRI data sets before and after chemotherapy, along with ground truth annotation, approximated based on the consensus of five human experts. (ii) We analyze human expert annotations and interrater variability. It turns out that current clinical practice of determining tumor volume is inaccurate and that manual annotations after chemotherapy may differ substantially. (iii) We evaluate seven computer-based segmentation methods, ranging from classical approaches to recent deep learning techniques. We show that the best ones offer a comparable quality to human expert annotations.

2016

Robust Interactive Multi-label Segmentation with an Advanced Edge Detector.

Sabine Müller, Peter Ochs, Joachim Weickert, Norbert Graf
In B. Andres, B. Rosenhahn (Eds.): Pattern Recognition. Lecture Notes in Computer Science, Vol. 9796, 117-128, Springer, Cham, 2016.

Abstract. Recent advances on convex relaxation methods allow for a flexible formulation of many interactive multi-label segmentation methods. The building blocks are a likelihood specified for each pixel and each label, and a penalty for the boundary length of each segment. While many sophisticated likelihood estimations based on various statistical measures have been investigated, the boundary length is usually measured in a metric induced by simple image gradients. We show that complementing these methods with recent advances of edge detectors yields an immense quality improvement. A remarkable feature of the proposed method is the ability to correct some erroneous labels, when computer generated initial labels are considered. This allows us to improve state-of-the-art methods for motion segmentation in videos by 5-10% with respect to the F-measure (Dice score).

Automatic brain tumor segmentation with a fast Mumford-Shah algorithm.

Sabine Müller, Joachim Weickert, Norbert Graf
In M. A. Styner, E. D. Angelini (Eds.): Medical Imaging 2016: Image Processing (San Diego, CA, February 2016), SPIE Vol. 9784, 97842S, 2016

Abstract. We propose a fully-automatic method for brain tumor segmentation that does not require any training phase. Our approach is based on a sequence of segmentations using the Mumford-Shah cartoon model with varying parameters. In order to come up with a very fast implementation, we extend the recent primal-dual algorithm of Strekalovskiy et al. (2014) from the 2D to the medically relevant 3D setting. Moreover, we suggest a new confidence refinement and show that it can increase the precision of our segmentations substantially. Our method is evaluated on 188 data sets with high-grade gliomas and 25 with low-grade gliomas from the BraTS14 database. Within a computation time of only three minutes, we achieve Dice scores that are comparable to state-of-the-art methods.

B Medical Terms

- **abdomen:** the anterior portion of the body between the thorax and the pelvis
- **bilateral:** pertaining to both sides
- **embryogenesis:** phase of prenatal development involved in establishment of the characteristic configuration of the embryonic body.
- **epigenesis:** development of an organism from an undifferentiated cell, consisting in the successive formation and development of organs and parts (not existent in the zygote)
- **germline:** cell line from which egg or sperm cells are derived
- **haematogenous:** Originating in, or carried by, the blood.
- **histogenesis:** The formation and development of body tissues.
- **metachronous:** consecutive development of tumors
- **metanephric:** embryological structure that give rise to the kidney
- **parenchyma:** the essential or functional elements of an organ
- **pathogenic variant:** alteration in a gene associated with an abnormal phenotype or increased disease risk
- **renal:** pertaining to the kidney
- **thorax:** body part between the neck and abdomen. Its walls are formed by the ribs.
- **thrombus:** a stationary blood clot along the wall of a blood vessel
- **vena cava:** one of the two major veins of the blood circulatory system that carry blood from other veins to the right atrium of the heart.

Figures

Tables

Index

L

M

N

O

P

R

S

T

U

V

W

References

[1] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), 2274–2282 (May 2012) 38

[2] Akeret, J., Chang, C., Lucchi, A., Refregier, A.: Radio frequency interference mitigation using deep convolutional neural networks. Astronomy and Computing 18, 35–39 (Jan 2017) 89

[3] Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46(3), 175–185 (Feb 1992) 43

[4] Ambrosio, L., Tortorelli, V.M.: Approximation of functional depending on jumps by elliptic functional via t-convergence. Communications on Pure and Applied Mathematics 43(8), 999–1036 (Dec 1990) 40

[5] Angelina, S., Suresh, L.P., Veni, S.K.: Image segmentation based on genetic algorithm for region growth and region merging. In: Proc. 2012 International Conference on Computing, Electronics and Electrical Technologies. pp. 970–974. IEEE, Tamil Nadu, India (Mar 2012) 38

[6] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(5), 898–916 (May 2011) 38, 40, 79

[7] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From contours to regions: An empirical evaluation. In: Proc. 2009 Conference on Computer Vision and Pattern Recognition. pp. 2294–2301. IEEE, Miami, FL (Jun 2009) 38

[8] Arrow, K.J., Hurwicz, L., Uzawa, H.: Studies in linear and non-linear programming. Stanford Mathematical Studies in the Social Sciences 122(3), 381–382 (Jul 1959) 24

[9] Athiwaratkun, B., Finzi, M., Izmailov, P., Wilson, A.G.: There are many consistent explanations of unlabeled data: Why you should average. In: Proc 2019 International Conference on Learning Representations. New Orleans, LA, USA (May 2019) 113

[10] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific Data 4, 170117 (Sep 2017) 94, 114, 115

[11] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge. arXiv preprint arXiv:1811.02629 (Nov 2018) 93, 94, 114, 115

[12] Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision 92(1), 1–31 (Nov 2011) 51

[13] Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International Journal of Computer Vision 12(1), 43–77 (Feb 1994) 51

[14] Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces, vol. 408. Springer (May 2011) 11

[15] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) Proc. 2006 European Conference on Computer Vision, pp. 404–417. Lecture Notes in Computer Science, Springer, Graz, Austria (May 2006) 47, 49, 136

[16] Beckwith, J.B.: Wilms tumor and other renal tumors of childhood: an update. Journal of Urology 136(1), 320–324 (Jun 1986) 28

[17] Beckwith, J.B.: Precursor lesions of Wilms' tumor: clinical and biological implications. Pediatric Blood & Cancer 21(3), 158–168 (Oct 1993) 29, 30

[18] Beckwith, J.B.: New developments in the pathology of Wilms' tumor: Pediatric Oncology. Cancer Investigation 15(2), 153–162 (Jun 1997) 29

[19] Beckwith, J.B., Kiviat, N.B., Bonadio, J.F.: Nephrogenic rests, nephroblastomatosis, and the pathogenesis of Wilms' tumor. Pediatric Pathology 10(1-2), 1–36 (Jul 1990) 29, 119

[20] Bergbauer, J., Nieuwenhuis, C., Souiai, M., Cremers, D.: Proximity priors for variational semantic segmentation and recognition. In: Proc. 2013 IEEE International Conference on Computer Vision, Workshop on Graphical Models for Scene Understanding. pp. 15–21. Sydney, Australia (Dec 2013) 24, 79

[21] Beucher, S.: Segmentation d'images et morphologie mathématique. Ph.D. thesis, Ecole Nationale Supérieure des Mines de Paris (Jun 1990) . . . 38

[22] Bloch, F.: Nuclear induction. Physical Review 70(7-8), 460 (Oct 1946) . . . 15, 17

[23] Bloembergen, N., Purcell, E.M., Pound, R.V.: Relaxation effects in nuclear magnetic resonance absorption. Physical Review 73(7), 679 (Apr 1948) . . . 17

[24] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proc. 1992 Workshop on Computational Learning Theory. pp. 144–152. ACM, Pittsburgh, Pennsylvania, USA (Jul 1992) . . . 39, 43, 87, 88, 90

[25] Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: Proc. 2001 IEEE International Conference on Computer Vision. pp. 105–112. Vancouver, Canada (Jul 2001) . . . 40, 79

[26] Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (Aug 1996) . . . 44

[27] Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (Oct 2001) . . . 39, 43, 44, 87, 88, 90, 136

[28] Breslow, N.E., Norris, R., Norkool, P.A., Kang, T., Beckwith, J.B., Perlman, E.J., Ritchey, M.L., Green, D.M., Nichols, K.E.: Characteristics and outcomes of children with the Wilms' tumor-Aniridia syndrome: a report from the National Wilms Tumor Study Group. Journal of Clinical Oncology 21(24), 4579–4585 (Dec 2003) . . . 28

[29] Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: Proc. 2011 Conference on Computer Vision and Pattern Recognition. pp. 2225–2232. IEEE, Colorado Springs, CO, USA (Jun 2011) . . . 39

[30] Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Kostas Daniilidis, Petros Maragos, N.P. (ed.) Proc. 2010 European Conference on Computer Vision, Lecture Notes in Computer Science, vol. 6315, pp. 282–295. Springer, Heraklion, Greece (Sep 2010) . . . 84

[31] Call, K.M., Glaser, T., Ito, C.Y., Buckler, A.J., Pelletier, J., Haber, D.A., Rose, E.A., Kral, A., Yeger, H., Lewis, W.H., et al.: Isolation and characterization of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus. Cell 60(3), 509–520 (Feb 1990) . . . 28

[32] Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 679–698 (Nov 1986) . . . 45, 79

[33] Chambolle, A.: An algorithm for total variation minimization and applications. Journal of Mathematical Imaging and Vision 20(1-2), 89–97 (Jan 2004) . . . 9

[34] Chambolle, A., Cremers, D., Pock, T.: A convex approach to minimal partitions. SIAM Journal on Applied Mathematics 5(4), 1113–1158 (Oct 2012) . . . 41, 79

[35] Chambolle, A., Pock, T.: A first-order primal–dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision 40(1), 120–145 (May 2011) . . . 23, 24, 25, 82, 97

[36] Chan, T.F., Sandberg, B.Y., Vese, L.A.: Active contours without edges for vector-valued images. Journal of Visual Communication and Image Representation 11(2), 130–141 (Jun 2000) . . . 87, 90

[37] Chan, T.F., Shen, J.J.: Image processing and analysis: variational, PDE, wavelet, and stochastic methods, vol. 94. SIAM (Sep 2005) . . . 49

[38] Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on Image Processing 10(2), 266–277 (Feb 2001) . . . 39, 40, 79, 95

[39] Chen, W.T., Liu, W.C., Chen, M.S.: Adaptive color feature extraction based on image color distributions. Transactions on Image Processing 19(8), 2005–2016 (Jun 2010) . . . 45

[40] Chen, Y., Fang, H., Xu, B., Yan, Z., Kalantidis, Y., Rohrbach, M., Yan, S., Feng, J.: Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. arXiv preprint arXiv:1904.05049 (Apr 2019) . . . 101, 109, 110, 111, 116

[41] Chow, S.C., Shao, J., Wang, H., Lokhnygina, Y.: Sample size calculations in clinical research. Chapman and Hall/CRC (Aug 2017) . . . 94

[42] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Proc. 2016 International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–432. Springer, Athens, Greece (Oct 2016) . . . 93, 104

[43] Cox, S.G., Kilborn, T., Pillay, K., Davidson, A., Millar, A.J.: Magnetic resonance imaging versus histopathology in Wilms' tumor and nephroblastomatosis: 3 examples of noncorrelation. Journal of Pediatric Hematology/Oncology 36(2), 81–84 (Mar 2014) . . . 119

[44] Cremers, D.: Optimal solutions for semantic image decomposition. Image and Vision Computing 30(8), 476–477 (Aug 2012) . . . 39

[45] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. 2005 IEEE Conference on Computer Vision and Pattern Recognition. pp. 886–893. San Diego, CA (Jun 2005) 45, 87, 90

[46] David, R., Graf, N., Karatzanis, I., Stenzhorn, H., Manikis, G.C., Sakkalis, V., Stamatakos, G.S., Marias, K.: Clinical evaluation of DoctorEye platform in nephroblastoma. In: Proc. 2012 International Advanced Research Workshop on In Silico Oncology and Cancer Investigation. pp. 1–4. IEEE, Athens, Greece (Oct 2012) 77

[47] Davidoff, A.M.: Wilms' tumor. Current Opinion in Pediatrics 21(3), 357–364 (Jun 2009) . . . . . VII, IX, 2, 28

[48] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. 2009 Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE, Miami, FL (Jun 2009) 51, 93

[49] Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (Jul 1945) 99

[50] Diebold, J., Demmel, N., Hazırbaş, C., Möller, M., Cremers, D.: Interactive multilabel segmentation of RGB-D images. In: Aujol, J.F., Nikolova, M., Papadakis, N. (eds.) Scale Space and Variational Methods in Computer Vision, Lecture Notes in Computer Science, vol. 9087, pp. 294–306. Springer, Berlin (Jun 2015) 24, 79

[51] Dollár, P., Zitnick, C.L.: Structured forests for fast edge detection. In: Proc. 2013 IEEE International Conference on Computer Vision. pp. 1841–1848. Sydney, Australia (Apr 2013) 79, 80, 81

[52] Dollár, P., Zitnick, C.L.: Fast edge detection using structured forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(8), 1558–1570 (Dec 2014) 79, 80, 81

[53] Dumoucel, S., Gauthier-Villars, M., Stoppa-Lyonnet, D., Parisot, P., Brisse, H., Philippe-Chomette, P., Sarnacki, S., Boccon-Gibod, L., Rossignol, S., Baumann, C., et al.: Malformations, genetic abnormalities, and Wilms' tumor. Pediatric Blood & Cancer 61(1), 140–144 (Aug 2013) 28

[54] Elishakoff, I.: Eigenvalues of inhomogeneous structures: unusual closed-form solutions. CRC Press (Oct 2004) 22

[55] Erginel, B.: Wilms' tumor and its management in a surgical aspect. In: van den Heuvel-Eibrink, M.M. (ed.) Wilms' Tumor. Codon Publications, Brisbane (Mar 2016) 31

[56] Fan, R.E., Chen, P.H., Lin, C.J.: Working set selection using second order information for training support vector machines. Journal of Machine Learning Research 6, 1889–1918 (Dec 2005) . . . 88

[57] Federer, H.: Geometric measure theory. Classics in Mathematics, Springer (Jan 1996) . . . 82

[58] Felzenszwalb, P.F., Huttenlocher, D.P.: Image segmentation using local variation. In: Proc. 1998 Conference on Computer Vision and Pattern Recognition. pp. 98–104. IEEE, Santa Barbara, CA, USA (Jun 1998) . . . 38

[59] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59(2), 167–181 (Sep 2004) . . . 40, 79

[60] Freund, Y., Schapire, R., Abe, N.: A short introduction to boosting. Journal of the Japanese Society For Artificial Intelligence 14(771-780), 1612 (Sep 1999) . . . 43

[61] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. 2012 Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE, Providence, RI, USA (Jun 2012) . . . 51

[62] Geirhos, R., Temme, C.R., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A.: Generalisation in humans and deep neural networks. In: Advances in Neural Information Processing Systems, Proc. 2018 Conference on Neural Information Processing Systems. pp. 7538–7550. Montréal, Canada (Dec 2018) . . . 39

[63] Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741 (Nov 1984) . . . 40, 79

[64] Goldstein, T., Li, M., Yuan, X.: Adaptive primal-dual splitting methods for statistical learning and image processing. In: Advances in Neural Information Processing Systems, Proc. 2015 Conference on Neural Information Processing Systems. pp. 2089–2097. Montréal, Canada (Dec 2015) . . . 23

[65] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. Adaptive Computation and Machine Learning, MIT Press (Jan 2017) . . . 95

[66] Graf, N., Reinhard, H., Semler, J.O.: SIOP 2001/GPOH Therapieoptimierungsstudie zur Behandlung von Kindern und Jugendlichen mit einem Nephroblastom (2003), www.kinderkrebsinfo.de, accessed: 2019-03-07 . . . 65, 73

[67] Graf, N., Tournade, M.F., de Kraker, J.: The role of preoperative chemotherapy in the management of Wilms' tumor: The SIOP studies. Urologic Clinics of North America 27(3), 443–454 (Aug 2000) 2, 30, 133

[68] Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. Magnetic Resonance in Medicine 34(6), 910–914 (Feb 1995) 128

[69] Gylys-Morin, V., Hoffer, F., Kozakewich, H., Shamberger, R.: Wilms' tumor and nephroblastomatosis: imaging characteristics at gadolinium-enhanced MR imaging. Radiology 188(2), 517–521 (Aug 1993) 119

[70] Hadamard, J.: Sur les problèmes aux dérivées partielles et leur signification physique. Princeton University Bulletin 13, 49–52 (Jan 1902) 22

[71] Hadamard, J.: Le probleme de Cauchy et les équations aux dérivées partielles linéaires hyperboliques. Acta Mathematica (Jan 1908) 22

[72] Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. Cell 144(5), 646–674 (Mar 2011) 1

[73] Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3(6), 610–621 (Nov 1973) 45, 121, 125

[74] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proc. 2017 IEEE International Conference on Computer Vision. pp. 2961–2969. Venice, Italy (Dec 2017) 39

[75] He, L., Greenshields, I.R.: A nonlocal maximum likelihood estimation method for Rician noise reduction in MR images. IEEE Transactions on Medical Imaging 28(2), 165–172 (Jul 2008) 128

[76] Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33(1), 115–126 (Oct 2006) 56

[77] Ho, T.K.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), 832–844 (Aug 1998) 44, 88

[78] Hogeling, M.: Anesthesia for children. In: Nouri, K., Benjamin, L., Alshaiji, J., Izakovic, J. (eds.) Pediatric Dermatologic Surgery, pp. 49–61. John Wiley & Sons (Apr 2019) 2, 67

[79] Hu, M.K.: Visual pattern recognition by moment invariants. Transactions on Information Theory 8(2), 179–187 (Feb 1962) 45

[80] Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot ensembles: Train 1, get M for free. In: Proc. 2017 International Conference on Learning Representations. Toulon, France (Apr 2017) 93, 101, 112

[81] Humphrey, J.D., Delange, S.L.: An Introduction to Biomechanics: Solids and Fluids, Analysis and Design. Springer (Jan 2016) 1

[82] Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.: No New-Net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) International MICCAI Brainlesion Workshop: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 234–244. Lecture Notes in Computer Science, Springer, Cham (Sep 2018) 93, 94, 95, 104, 105, 107, 114, 117

[83] Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. In: Proc. 2018 Conference on Uncertainty in Artificial Intelligence. Monterey, CA, USA (Aug 2018) 105, 109, 112, 113

[84] Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proc. 2017 Conference on Computer Vision and Pattern Recognition Workshops. pp. 11–19. IEEE, Honolulu, HI, USA (Jul 2017) 104

[85] Jiang, Y., Li, Z., Zhang, L., Sun, P.: An improved SVM classifier for medical image classification. In: Proc. 2007 International Conference on Rough Sets and Intelligent Systems Paradigms. pp. 764–773. Springer, Warsaw, Poland (Jun 2007) 43

[86] Kappes, J., Andres, B., Hamprecht, F., Schnorr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Lellmann, J., Komodakis, N., et al.: A comparative study of modern inference techniques for discrete energy minimization problems. In: Proc. 2013 Conference on Computer Vision and Pattern Recognition. pp. 1328–1335. Portland, OR, USA (Jun 2013) 51

[87] Karami, E., Prasad, S., Shehata, M.: Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images. In: Proc. 2015 Newfoundland Conference on Electrical and Computer Engineering. pp. 212–217. St. John's, Canada (Nov 2015) 49

[88] Kaste, S.C., Dome, J.S., Babyn, P.S., Graf, N.M., Grundy, P., Godzinski, J., Levitt, G.A., Jenkinson, H.: Wilms' tumor: prognostic factors, staging, therapy and late effects. Pediatric Radiology 38(1), 2–17 (Jan 2008) 2, 30, 133

[89] Ke, T.W., Maire, M., Yu, S.X.: Multigrid neural architectures. In: Proc. 2017 Conference on Computer Vision and Pattern Recognition. pp. 4067–4075. Honolulu, HI, USA (Jul 2017) . . . 101, 109

[90] Keuper, M., Andres, B., Brox, T.: Motion trajectory segmentation via minimum cost multicuts. In: Proc. 2015 IEEE International Conference on Computer Vision. pp. 3271–3279. Santiago, Chile (Dec 2015) . . . 84, 85

[91] Kim, S., Chung, D.H.: Pediatric solid malignancies: neuroblastoma and Wilms' tumor. Surgical Clinics of North America 86(2), 469–487 (Apr 2006) . . . 2, 28

[92] Koepfler, G., Lopez, C., Morel, J.-M.: A multiscale algorithm for image segmentation by variational method. SIAM Journal on Numerical Analysis 31(1), 282–299 (Feb 1994) . . . 40

[93] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, Proc. 2012 Conference on Neural Information Processing Systems. pp. 1097–1105. Lake Tahoe, NV, USA (Dec 2012) . . . 51, 93

[94] Lange, J., Peterson, S.M., Takashima, J.R., Grigoriev, Y., Ritchey, M.L., Shamberger, R.C., Beckwith, J.B., Perlman, E., Green, D.M., Breslow, N.E.: Risk factors for end stage renal disease in non-WT1-syndromic Wilms' tumor. The Journal of Urology 186(2), 378–386 (Aug 2011) . . . 29

[95] Lauterbur, P.C.: Image formation by induced local interactions: examples employing nuclear magnetic resonance. Nature 242(5394), 190–191 (Mar 1973) . . . 15

[96] Lee, D.: Plato: The Republic. Penguin Books (Jan 1974) . . . 21

[97] Lellmann, J., Schnörr, C.: Continuous multiclass labeling approaches and algorithms. SIAM Journal on Imaging Sciences 4(4), 1049–1096 (Feb 2011) . . . 40

[98] Liu, D., Yu, J.: Otsu method and k-means. In: Proc. 2009 IEEE International Conference on Hybrid Intelligent Systems. vol. 1, pp. 344–349. IEEE, Shenyang, China (Aug 2009) . . . 88

[99] Liu, M.Y., Tuzel, O., Ramalingam, S., Chellappa, R.: Entropy rate superpixel segmentation. In: Proc. 2011 Conference on Computer Vision and Pattern Recognition. pp. 2097–2104. IEEE, Colorado Springs, CO, USA (Jun 2011) . . . 38, 87, 88, 90

[100] Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28(2), 129–137 (Mar 1982) . . . 39, 49, 87, 88, 90

[101] Lonergan, G.J., Martinez-Leon, M.I., Agrons, G.A., Montemarano, H., Suarez, E.S.: Nephrogenic rests, nephroblastomatosis, and associated lesions of the kidney. RadioGraphics 18(4), 947–968 (Aug 1998) 29, 119

[102] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. 2015 Conference on Computer Vision and Pattern Recognition. pp. 3431–3440. Boston, MA, USA (Jun 2015) 39

[103] Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: Proc. 2017 International Conference on Learning Representations. Toulon, France (Apr 2017) 112

[104] Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. 1999 International Conference on Computer Vision. vol. 99, pp. 1150–1157. IEEE, Corfu, Greece (Sep 1999) 45, 47

[105] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (Nov 2004) 47, 49

[106] Lukacs, E.: A characterization of the normal distribution. The Annals of Mathematical Statistics 13(1), 91–93 (Jan 1942) 9

[107] MacLennan, G.T., Cheng, L.: Neoplasms of the kidney. In: Cheng, L., Bostwick, D.G. (eds.) Essentials of Anatomic Pathology, pp. 1645–1679. Springer (Feb 2016) 30

[108] Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al.: ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269 (Jan 2017) 51, 53

[109] Mangale, S., Khambete, M.: Gray level co-occurrence matrix feature based object tracking in thermal infrared imagery. Journal of Electronic Imaging 27(3), 033021 (May 2018) 121

[110] Marescaux, J., Diana, M.: Next step in minimally invasive surgery: hybrid image-guided surgery. Journal of Pediatric Surgery 50(1), 30–36 (Jan 2015) 1

[111] Marsh, J.C., Goldfarb, J., Shafman, T.D., Diaz, A.Z.: Current status of immunotherapy and gene therapy for high-grade gliomas. Cancer Control 20(1), 43–48 (Jan 2013) 94

[112] Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 2001 IEEE International Conference on Computer Vision. vol. 2, pp. 416–423. IEEE, Vancouver, Canada (Jul 2001) 51

[113] Mazzara, G.P., Velthuizen, R.P., Pearlman, J.L., Greenberg, H.M., Wagner, H.: Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. International Journal of Radiation Oncology - Biology - Physics 59(1), 300–312 (May 2004) 65, 94

[114] Mejia-Foster's AP Language and Composition: Allegory of the cave, http://mejialangcomp.weebly.com/platos-allegory-of-the-cave.html, accessed: 2019-03-12 21

[115] Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BraTS). IEEE Transactions on Medical Imaging 34(10), 1993–2024 (Dec 2014) 51, 53, 93, 94, 114

[116] Millenson, M.L.: Demanding medical excellence: Doctors and accountability in the information age. University of Chicago Press (Mar 2000) 1

[117] Mktyscn: Wikipedia – semi continuity., https://en.wikipedia.org/wiki/Semi-continuity, accessed: 2019-05-14 10

[118] Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société Mathématique de France 93, 273–299 (Jan 1965) 11

[119] Morel, J.M., Solimini, S.: Variational Methods in Image Segmentation, Progress in Nonlinear Differential Equations and Their Applications, vol. 12. Birkhäuser, Basel (dec 1994) 97

[120] Mrabet, Y.: Wikimedia, Human anatomy planes., https://commons.wikimedia.org/wiki/File:Human_anatomy_planes.svg, accessed: 2019-03-13 14

[121] Mullen, L., Gaillard, F.: Radiopaedia - MRI sequences, https://radiopaedia.org/articles/mri-sequences-overview, accessed: 2019-03-19 18

[122] Mumford, D., Shah, J.: Boundary detection by minimizing functionals, I. In: Proc. 1985 Conference on Computer Vision and Pattern Recognition. pp. 22–26. IEEE Computer Society Press, San Francisco, CA (Jun 1985) 95, 96

[123] Mumford, D., Shah, J.: Optimal approximation of piecewise smooth functions and associated variational problems. Communications on Pure and Applied Mathematics 42(5), 577–685 (1989)
. . . . . 39, 40, 79, 95, 96

[124] Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) International MICCAI Brainlesion Workshop: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 311–320. Springer, Cham (Jan 2018)
. . . . . 93, 94, 95, 104, 106

[125] National Cancer Institute: What is cancer?, https://www.cancer.gov/about-cancer/understanding/what-is-cancer, accessed: 2019-03-06
. 1

[126] Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical Programming 103(1), 127–152 (May 2005)
. 24

[127] Nieuwenhuis, C., Cremers, D.: Spatially varying color distributions for interactive multilabel segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(5), 1234–1247 (Aug 2012)
. . . . . 24, 26, 39, 40, 41, 79, 80, 82, 83

[128] Nieuwenhuis, C., Hawe, S., Kleinsteuber, M., Cremers, D.: Co-sparse textural similarity for interactive segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Proc. 2014 European Conference on Computer Vision, Lecture Notes in Computer Science, vol. 8694, pp. 285–301. Springer, Cham, Switzerland (Sep 2014)
. . . . . 24, 26, 39, 40, 79, 83, 84

[129] Nowak, R.D.: Wavelet-based Rician noise removal for magnetic resonance imaging. IEEE Transactions on Image Processing 8(10), 1408–1419 (Oct 1999)
. 128

[130] Ochs, P., Brox, T.: Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In: Proc. 2011 IEEE International Conference on Computer Vision. pp. 1583–1590. Barcelona, Spain (Nov 2011)
. . . . . 24, 79

[131] Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(6), 1187–1200 (Jun 2014)
. . . . . 26, 83, 84, 85, 86

[132] Ono, S.: Primal-dual plug-and-play image restoration. IEEE Signal Processing Letters 24(8), 1108–1112 (May 2017)
. 23

[133] Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics 9(1), 62–66 (Jan 1979)
. . . . . 38, 98

[134] Oun, R., Moussa, Y.E., Wheate, N.J.: The side effects of platinum-based chemotherapy drugs: a review for chemists. Dalton Transactions 47(19), 6645–6653 (May 2018) . . . 32

[135] Owens, C.M., Brisse, H.J., Olsen, Ø.E., Begent, J., Smets, A.M.: Bilateral disease and new trends in Wilms' tumor. Pediatric Radiology 38(1), 30–39 (Jan 2008) . . . 119, 121

[136] Pascanu, R., Montufar, G., Bengio, Y.: On the number of response regions of deep feed forward networks with piece-wise linear activations. arXiv preprint arXiv:1312.6098 (Dec 2013) . . . 103

[137] Pastore, G., Znaor, A., Spreafico, F., Graf, N., Pritchard-Jones, K., Steliarova-Foucher, E.: Malignant renal tumours incidence and survival in European children (1978–1997): Report from the Automated Childhood Cancer Information System Project. European Journal of Cancer 42(13), 2103–2114 (Sep 2006) . . . VII, IX, 2, 28

[138] Pelletier, J., Bruening, W., Kashtan, C.E., Mauer, S.M., Manivel, J.C., Striegel, J.E., Houghton, D.C., Junien, C., Habib, R., Fouser, L., et al.: Germline mutations in the Wilms' tumor suppressor gene are associated with abnormal urogenital development in Denys-Drash syndrome. Cell 67(2), 437–447 (Oct 1991) . . . 28

[139] Perona, P., Malik, J.: Scale space and edge detection using anisotropic diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 629–639 (Jul 1990) . . . 79

[140] Popov, S., Sebire, N., Vujanic, G.: Wilms' tumor–histology and differential diagnosis. In: van den Heuvel-Eibrink, M.M. (ed.) Wilms' Tumor. Codon Publications, Brisbane (Mar 2016) . . . 29

[141] Rabi, I.I., Zacharias, J., Millman, S., Kusch, P.: A new method of measuring nuclear magnetic moment. Physical Review 53(4), 318–327 (Feb 1938) . . . 15

[142] Rance, T.: Case of fungus haematodes of the kidneys. Journal of Medical Physics 32(185), 19–25 (Jul 1814) . . . 28

[143] Raykar, V.C., Yu, S., Zhao, L.H., Jerebko, A., Florin, C., Valadez, G.H., Bogoni, L., Moy, L.: Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proc. 2009 International Conference on Machine Learning. pp. 889–896. ACM, Montreal, Canada (Jun 2009) . . . 56

[144] Real, E., Aggarwal, A., Huang, Y., Le, Q.V.: Regularized evolution for image classifier architecture search. In: Proc. 2019 Conference on Artificial Intelligence. pp. 4780–4789. AAAI Press, Honolulu, HI, USA (Jan 2019) . . . 43

[145] Rockafellar, R.T.: Convex Analysis. Princeton Landmarks in Mathematics and Physics, Princeton University Press (Jan 1970) 9, 10, 23, 24

[146] Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM Journal on Control and Optimization 14(5), 877–898 (Jul 1976) 11

[147] Rogowska, J.: Overview and fundamentals of medical image segmentation. Handbook of Medical Imaging, Processing and Analysis pp. 69–85 (Dec 2000) 38

[148] Rohrschneider, W.K., Weirich, A., Rieden, K., Darge, K., Tröger, J., Graf, N.: US, CT and MR imaging characteristics of nephroblastomatosis. Pediatric Radiology 28(6), 435–443 (Jun 1998) 119

[149] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, Cham (Oct 2015) 39, 87, 89, 90, 93, 95, 101, 102, 104

[150] Rother, C., Kolmogorov, V., Blake, A.: GrabCut: interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics 23(3), 309–314 (Aug 2004) 40, 79

[151] Rump, P., Zeegers, M., Van Essen, A.: Tumor risk in Beckwith–Wiedemann syndrome: A review and meta-analysis. American Journal of Medical Genetics 136(1), 95–104 (Jul 2005) 28

[152] Samanta, S., Ahmed, S.S., Salem, M.A.M.M., Nath, S.S., Dey, N., Chowdhury, S.S.: Haralick features based automated glaucoma classification using back propagation neural network. In: Satapathy, S.C., Biswal, B.N., Udgata, S.K., Mandal, J.K. (eds.) Proc. 2014 International Conference on Frontiers of Intelligent Computing: Theory and Applications. pp. 351–358. Springer (Nov 2014) 121

[153] Santner, J., Pock, T., Bischof, H.: Interactive multi-label segmentation. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) Proc. 2010 Asian Conference on Computer Vision, Lecture Notes in Computer Science, vol. 6492, pp. 397–410. Springer, Queenstown, New Zealand (Nov 2010) 26, 40, 79, 83

[154] Santner, J., Unger, M., Pock, T., Leistner, C., Saffari, A., Bischof, H.: Interactive texture segmentation using random forests and total variation. In: Proc. 2009 British Machine Vision Conference. pp. 66.1–66.12. British Machine Vision Association, London, UK (Sep 2009) 40, 79

[155] Saxe, R.: It's Interesting., https://its-interesting.com/2015/12/15/mother-child-mri/, accessed: 2019-03-14 13

[156] Schädler, M.R., Meyer, B.T., Kollmeier, B.: Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. Journal of the Acoustical Society of America 131(5), 4134–4151 (May 2012) . . . 45

[157] Scott, R.H., Stiller, C.A., Walker, L., Rahman, N.: Syndromes and constitutional chromosomal abnormalities associated with Wilms' tumor. Journal of Medical Genetics 43(9), 705–715 (Sep 2006) . . . 28

[158] Siemens Healthcare: MAGNETOM Spectra, http://www.healthcare.siemens.de, accessed: 2019-02-04 . . . 16

[159] Smith, L.N.: Cyclical learning rates for training neural networks. In: Proc. 2017 Conference on Applications of Computer Vision. pp. 464–472. IEEE, Santa Rosa, CA, USA (Mar 2017) . . . 112

[160] Soh, L.K., Tsatsoulis, C.: Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Transactions on Geoscience and Remote Sensing 37(2), 780–795 (Mar 1999) . . . 126

[161] Soomro, M.H., Giunta, G., Laghi, A., Caruso, D., Ciolina, M., De Marchis, C., Conforto, S., Schmid, M.: Haralick's texture analysis applied to colorectal T2-weighted MRI: A preliminary study of significance for cancer evolution. In: Proc.2017 IEEE International Conference on Biomedical Engineering. pp. 16–19. IEEE, Innsbruck, Austria (Feb 2017) . . . 125

[162] Sprawls, P.: Magnetic Resonance Imaging: Principles, Methods, and Techniques. Medical Physics Publishing, Atlanta, GA (2000) . . . 19

[163] Sridharan, S., Macias, V., Tangella, K., Kajdacsy-Balla, A., Popescu, G.: Prediction of prostate cancer recurrence using quantitative phase imaging. Scientific Reports 5, 9976 (Sep 2015) . . . 1

[164] Strekalovskiy, E., Cremers, D.: Generalized ordering constraints for multilabel optimization. In: Proc. 2011 International Conference on Computer Vision. pp. 2619–2626. Barcelona, Spain (Jan 2011) . . . 24, 79

[165] Strekalovskiy, E., Nieuwenhuis, C., Cremers, D.: Nonmetric priors for continuous multilabel optimization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Proc. 2012 European Conference on Computer Vision, pp. 208–221. Lecture Notes in Computer Science, Springer, Berlin (Oct 2012) . . . 24, 79

[166] Strekalovskiy, E., Cremers, D.: Real-time minimization of the piecewise smooth mumford-shah functional. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Proc. 2014 European Conference on Computer Vision, Lecture Notes in Computer Science, vol. 8692, pp. 127–141. Springer (Sep 2014) . . . 40, 95, 97

[167] Stutz, D., Hermans, A., Leibe, B.: Superpixels: An evaluation of the state-of-the-art. Computer Vision and Image Understanding 166, 1–27 (Jan 2018) . . . 38

[168] Teem: NRRD: Nearly Raw Raster Data, http://teem.sourceforge.net/nrrd/index.html, accessed: 2019-04-25 . . . 52

[169] Tikhonov, A.N.: Regularization of incorrectly posed problems. Soviet Mathematics Doklady 4(6), 1624–1627 (1963) . . . 22

[170] Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. IEEE Transactions on Medical Imaging 29(6), 1310–1320 (Jun 2010) . . . 96

[171] Tustison, N., Wintermark, M., Durst, C., Avants, B.: Ants and Arboles. In: Proc. 2013 MICCAI BraTS Workshop. MICCAI Society, Nagoya, Japan (Sep 2013) . . . 95

[172] Cancer Research UK: Worldwide cancer statistics, https://www.cancerresearchuk.org, accessed: 2019-11-26 . . . VII, IX

[173] Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 4105–4113. Honolulu, HI, USA (Jul 2017) . . . 104

[174] Unger, M.: Convex Optimization for Image Segmentation. Ph.D. thesis, Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria (October 2012) . . . 22, 25

[175] Unger, M., Pock, T., Trobin, W., Cremers, D., Bischof, H.: TVSeg - interactive total variation based image segmentation. In: Proc. 2008 British Machine Vision Conference. pp. 1–10. British Machine Vision Association, Leeds, UK (Sep 2008) . . . 24, 79

[176] Urban, G., Bendszus, M., Hamprecht, F.A., Kleesiek, J.: Multi-modal brain tumor segmentation using deep convolutional neural networks. In: Proc. 2014 MICCAI BraTS Workshop. pp. 31–35. MICCAI Society, Boston, MA, USA (Sep 2014) . . . 94

[177] Van Den Heuvel-Eibrink, M.M., Hol, J.A., Pritchard-Jones, K., Van Tinteren, H., Furtwängler, R., Verschuur, A.C., Vujanic, G.M., Leuschner, I., Brok, J., Rübe, C., et al.: Position paper: Rationale for the treatment of Wilms' tumour in the UMBRELLA SIOP–RTSG 2016 protocol. Nature Reviews Urology 14(12), 743 (Dec 2017) . . . 34

[178] Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer Research 77(21), 104–107 (Nov 2017) 126

[179] Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation using the Mumford and Shah model. International Journal of Computer Vision 50(3), 271–293 (Dec 2002) 40

[180] Vidyaratne, L., Alam, M., Shboul, Z., Iftekharuddin, K.M.: Deep learning and texture-based semantic label fusion for brain tumor segmentation. vol. 10575, p. 105750D. International Society for Optics and Photonics, Houston, TX, USA (Feb 2018) 106

[181] Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(6), 583–598 (Jun 1991) 38

[182] Vujanić, G.M., Sandstedt, B.: The pathology of Wilms' tumor (nephroblastoma): the International Society of Paediatric Oncology approach. Journal of Clinical Pathology 63(2), 102–109 (Feb 2010) 4, 32, 133, 137

[183] Vujanić, G.M., Sandstedt, B., Harms, D., Kelsey, A., Leuschner, I., de Kraker, J.: Revised international Society of Paediatric Oncology (SIOP) working classification of renal tumors of childhood. Pediatric Blood & Cancer 38(2), 79–82 (Feb 2002) 30

[184] Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A benchmark for comparison of dental radiography analysis algorithms. Medical Image Analysis 31, 63–76 (Jul 2016) 51

[185] Wang, J.J.Y., Bensmail, H., Gao, X.: Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification. Pattern Recognition 46(12), 3249–3255 (Dec 2013) 136

[186] Wang, L., He, D.C.: Texture classification using texture spectrum. Pattern Recognition 23(8), 905–910 (Aug 1990) 45

[187] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23(7), 903–921 (Jul 2004) 57

[188] WebPathology: Webpathology - visual survey of surgical pathology, /www.webpathology.com/, accessed: 2019-05-14 32

F

[189] Weikersdorfer, D., Gossow, D., Beetz, M.: Depth-adaptive superpixels. In: Proc. 2012 IEEE International Conference on Pattern Recognition. pp. 2087–2090. IEEE, Tsukuba, Japan (Nov 2012) . . . 38

[190] Weitz, E.: http://weitz.de/sift/, accessed: 2019-09-01 . . . 47, 48

[191] Wen, P.Y., Macdonald, D.R., Reardon, D.A., Cloughesy, T.F., Sorensen, A.G., Galanis, E., DeGroot, J., Wick, W., Gilbert, M.R., Lassman, A.B., et al.: Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. Journal of Clinical Oncology 28(11), 1963–1972 (Mar 2010) . . . 94

[192] Werlberger, M., Unger, M., Pock, T., Bischof, H.: Efficient minimization of the non-local Potts model. In: Bruckstein, A.M., ter Haar Romeny, B.M., Bronstein, A.M., Bronstein, M.M. (eds.) Scale Space and Variational Methods in Computer Vision, Lecture Notes in Computer Science, vol. 6667, pp. 314–325. Springer, Berlin (Jun 2011) . . . 24, 79

[193] Wibmer, A., Hricak, H., Gondo, T., Matsumoto, K., Veeraraghavan, H., Fehr, D., Zheng, J., Goldman, D., Moskowitz, C., Fine, S.W., et al.: Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different gleason scores. European Radiology 25(10), 2840–2850 (Oct 2015) . . . 125

[194] Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D.: Understanding data augmentation for classification: when to warp? In: Proc. 2016 IEEE International Conference on Digital Image Computing: Techniques and Applications. pp. 1–6. IEEE, Gold Coast, Australia (Dec 2016) . . . 61

[195] Wu, Y., He, K.: Group normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Proc. 2018 European Conference on Computer Vision. pp. 3–19. Lecture Notes in Computer Science, Munich, Germany (Sep 2018) . . . 106

[196] Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition 90, 119–133 (Jun 2019) . . . 43

[197] Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: Proc. 2007 International Workshop on Multimedia Information Retrieval. pp. 197–206. ACM, Augsburg, Germany (Sep 2007) . . . 49, 135

[198] Zach, C., Häne, C., Pollefeys, M.: What is optimized in convex relaxations for multilabel problems: Connecting discrete and continuously inspired map inference. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(1), 157–170 (Jun 2013) . . . 40

[199] Zayed, N., Elnemr, H.A.: Statistical analysis of Haralick texture features to discriminate lung abnormalities. Journal of Biomedical Imaging 2015, 12 (Sep 2015) 121, 125

[200] Zhou, C., Chen, S., Ding, C., Tao, D.: Learning contextual and attentive information for brain tumor segmentation. In: International MICCAI Brainlesion Workshop: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 497–507. Springer, Cham (Sep 2018) 95, 106

[201] Zhou, C., Ding, C., Lu, Z., Wang, X., Tao, D.: One-pass multi-task convolutional neural networks for efficient brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 637–645. Springer (Sep 2018) 106

[202] Zhu, S.C., Yuille, A.: Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(9), 884–900 (Sep 1996) 38