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, Jonas Dobberstein, André Lollert, Andreas Hötker and Norbert Graf

Saarbrücken 2017

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.
Benchmarking Wilms’ Tumor in Multi-Sequence MRI Data:  
Why Does Current Clinical Practice Fail? 
Which Popular Segmentation Algorithms Perform Well? 

Sabine Müller  
Department of Pediatric Oncology and Hematology  
Saarland University Medical Center  
66421 Homburg, Germany  
Mathematical Image Analysis Group  
Faculty of Mathematics and Computer Science, Campus E1.7  
Saarland University, 66041 Saarbrücken, Germany  
smueller@mia.uni-saarland.de 

Iva Farag  
Department of Pediatric Oncology and Hematology  
Saarland University Medical Center  
66421 Homburg, Germany  
ivabaykova@gmail.com 

Joachim Weickert  
Mathematical Image Analysis Group  
Faculty of Mathematics and Computer Science, Campus E1.7  
Saarland University, 66041 Saarbrücken, Germany  
weickert@mia.uni-saarland.de 

Yvonne Braun  
Department of Pediatric Oncology and Hematology  
Saarland University Medical Center  
66421 Homburg, Germany  
yvonne.braun@uks.eu 

Jonas Dobberstein  
Department of Pediatric Oncology and Hematology  
Saarland University Medical Center  
66421 Homburg, Germany  
jonas.sge.92@hotmail.de
André Lollert
Department of Diagnostic and Interventional Radiology
Medical Center of the Johannes Gutenberg University
55131 Mainz, Germany
andre.lollert@unimedizin-mainz.de

Andreas Hötker
Department of Diagnostic and Interventional Radiology
Medical Center of the Johannes Gutenberg University
55131 Mainz, Germany
andreas.hoetker@uni-mainz.de

Norbert Graf
Department of Pediatric Oncology and Hematology
Saarland University Medical Center
66421 Homburg, Germany
graf@uks.eu
Abstract

Wilms’ tumor is the most frequent malignant kidney tumor 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 allowing evaluation of the quality of human or computer-based segmentations yet. As a remedy, our contributions 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 as human expert annotations.

1 Introduction

Wilms’ tumor, or nephroblastoma, accounts for 5% of all cancers in children and juveniles. It constitutes the most frequent malignant kidney tumor in childhood [40]. About 75% of all patients are younger than five years - with a peak between two and three years [13, 27]. In Europe, diagnosis and therapy follow the guidelines of the International Society of Pediatric Oncology (SIOP) [19, 26]. One of the most important characteristics of this therapy protocol is a preoperative chemotherapy. Clinicians categorize patients as high-, intermediate- or low-risk candidates according to histology, local stage and tumor volume after chemotherapy. Postoperative treatment ranges from no chemotherapy (low risk stage I) up to chemotherapy with irradiation of the tumor bed (high risk, stage II and III).

The most common histological subtypes of regressive and mixed type actually belong to the intermediate risk tumors. However, if, after chemotherapy, these tumors have a volume of more than 500 ml, they are highly malignant and the patients are treated according to the high risk group protocol [18]. In order not to expose children to unnecessary medical burden on the one hand and to maximize their chances of survival on the other, an exact determination of the tumor volume is indispensable.

Current Practices of Segmentation by Human Experts. Radiologists

\[\text{\footnotesize \textsuperscript{1}This is a revised version dated September 3, 2018.}\]
traditionally model the tumor through a time-intensive manual segmentation procedure involving the outlining of the gross tumor volume on numerous two-dimensional imaging “slices”. Alternatively, they determine the tumor volume by measuring three axes of tumor extension and assuming the nephroblastoma to have an ellipsoid shape [18]. Usually, both variants are conducted using either computed tomography (CT) or magnetic resonance imaging (MRI) data. Unfortunately, the reliability and consistent reproducibility of expert delineations of Wilms’ tumors has not been investigated so far.

**Computer-based Segmentation Algorithms.** One obvious step to avoid the reproducibility 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 and shall be discussed next.

An initial idea for segmentation is to extend user marked seed points in the tumor by region growing based on intensity thresholding [12]. A refined approach is to initialize an active contour inside the tumor and to expand the segmentation according to image intensities and gradients [12]. More recently, a more advanced energy-based method for segmentation of nephroblastoma has been proposed [35]. User-set scribbles are employed to approximate the gray value distributions of tumor and surrounding tissues. The energy is then regularized by an image metric induced by a state-of-the-art edge detection.

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. Moreover, a comparative evaluation of these algorithms is prevented by the fact that is no public benchmark available. So far the few computer-based algorithms for Wilms’ tumor segmentation have been tested on different data sets.

### 1.1 Our Contributions

The goal of our paper is to offer solutions to the before mentioned problems in a threefold way:

(i) We establish the first publically available heterogeneous benchmark
data set for Wilms’ tumors. It allows clinicians to train their segmentation abilities, and computer scientists to evaluate their algorithms. Our benchmark consists of multi-sequence MRI data before and after chemotherapy. Ground truth segmentations are approximated by consensus truth of five human experts.

(ii) Based on this benchmark, we scrutinize the widely-used ellipsoid approximation to the tumor volume as well as the interrater variability of manual delineations. Both results will reveal substantial shortcomings of the current standards.

(iii) As a second benchmark application, we evaluate seven algorithms w.r.t. their usefulness for Wilms’ tumor segmentation. Although most of these segmentation algorithms are popular and time-proven methods in the computer vision community, six of them have not been used for Wilms’ tumor segmentation yet. Our algorithms include a fully-automatic method based on Chan-Vese active contours [9], a random forest classifier [22, 7], a support vector machine [6], a k-means clustering algorithm [30], a clustering of superpixels [29], and our recent semi-automatic approach [35]. Since the Wilms’ tumor data are necessarily limited, most segmentation methods based on deep learning cannot be applied due to an insufficient amount of training data. One of the few methods that can be used is the U-Net [42], which we are also evaluating.

In computer vision, benchmarking and performance evaluation have established themselves as important triggers for scientific progress in key areas ranging from motion analysis [4, 3, 17] over optimisation algorithms [24] to segmentation methods [33]. Pure benchmarking and performance evaluation have become equally influential in medical image analysis [20, 43], e.g. in registration [38, 46] and various segmentation problems [2, 5, 25, 32, 34]. The authors of these publications typically follow the clear scientific practice not to mix benchmark data with own unpublished algorithms, since this enables a fair comparison and avoids conflicts of interests. We adhere to these standards and refrain from proposing novel algorithms: We focus on evaluating the performance of popular fully-automatic segmentation methods when being applied to Wilms’ tumor data. These well-established baseline algorithms are complemented with a recent semi-automatic approach of some of us [35]. It has produced state-of-the-art results for video segmentation and allows to incorporate medical expert knowledge through hand-drawn scribbles.
1.2 Paper Organisation

Section 2 introduces our new multi-sequence benchmark for Wilms’ tumor segmentation. We analyze interoperator variability and compare the determined volumes with volume approximations used in clinical practice. The third section evaluates human segmentations, and Section 4 is devoted to the evaluation of computer-based segmentation algorithms. Our conclusions are summarized in Section 5.

2 Benchmark Data

To describe our benchmark data set, we first present details on the acquired MRI data and the chosen method for ground truth approximation. Afterwards, we introduce our error metrics and evaluate the interoperator variability on the proposed data set. In the end, we compare volume variability among human expert raters, ground truth and ellipsoid shapes.

2.1 Data Sets

Our image data set consists of 28 multi-sequence MR scans from 17 Wilms’ tumor patients (5 male, 12 female), out of which 15 have been acquired from intermediate risk tumor (histological diagnosis: stromal predominant (2), mixed histology (6) or regressive type (7)) and 2 from high risk tumor types (histological diagnosis: blastemal predominant). For eleven patients,
we have both, data before and after chemotherapy. The remaining ones are missing either data before or after chemotherapy. Fig. 1 shows subjects’ age distribution. 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. 2.

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 vendors, varying field strength (1.5T and 3T) and implementations of the imaging sequences. The data sets used in our benchmark share the following three MRI settings:

- **T2**: $T_2$-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.
- **T1**: $T_1$-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.
- **T1c**: $T_1$-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.

The different MRI sequences were spatially co-registered on the T2 sequence using a rigid transformation.

### 2.2 Annotations by Human Experts

The images were manually annotated by five human expert raters coauthoring this publication. 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 an 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. Segmentations were performed using the MITK software\(^2\), and experts outlined tumor structures in T2-sequences in every axial slice.

### 2.3 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

\(^2\)www.mitk.org
Figure 2: Example of Wilms’ tumor (training data) before and after chemotherapy with experts’ consensus truth. From top to bottom: T2, T1, T1c.

several contexts [21, 41]. 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 [44, 45] to produce consensus segmentations. The STAPLE algorithm is based on expectation maximization. Let $D_{x,j}, j = 1,\ldots,n$ be the expert decisions and $\hat{G}$ the true consensus segmentation. The performance of each expert is estimated on the basis of his sensitivity $p_j = \Pr(D_{x,j} = 1 \mid \hat{G} = 1)$ and specificity $q_j = \Pr(D_{x,j} = 0 \mid \hat{G} = 0)$. The STAPLE algorithm iterates between estimating the conditional probability of $\hat{G}$ in relation to the expert decisions and previous estimates of the performance parameters and estimation of updated reliability parameters. Before chemotherapy, convergence is
Table 1: Estimated quality parameters of each expert before and after chemotherapy.

<table>
<thead>
<tr>
<th></th>
<th>Rater-1</th>
<th>Rater-2</th>
<th>Rater-3</th>
<th>Rater-4</th>
<th>Rater-5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre-Chemotherapy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.76</td>
<td>0.71</td>
<td>0.80</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.65</td>
<td>0.75</td>
<td>0.73</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>Post-Chemotherapy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.78</td>
<td>0.60</td>
<td>0.77</td>
<td>0.70</td>
<td>0.67</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.72</td>
<td>0.57</td>
<td>0.75</td>
<td>0.85</td>
<td>0.82</td>
</tr>
</tbody>
</table>

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. 1 and indicate high inter-rater variability.

2.4 Error Metrics

We show results in terms of the metrics suggested in [35] and compute precision and recall as

\[ P_{G,G} := \frac{|\hat{G} \cap G|}{|G|}, \quad R_{G,G} := \frac{|G \cap \hat{G}|}{|G|}, \]  

(1)

where \( \hat{G} \) is the experts’ consensus truth, and \( G \) the algorithmic prediction. The harmonic mean of precision and recall is called Dice score. It relates the area of a cluster to its overlap with the approximated ground truth. The average Dice score determines the overall segmentation accuracy.

3 Evaluation of Human Expert Segmentations

3.1 Accuracy

3.1.1 Interoperator Variability

We calculate the interoperator variability using all 28 data sets of all 17 patients. In order to do so, we compute the 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. 2 shows the interoperator 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 $0.87 \pm 0.05$ on tumor areas. After chemotherapy, when tumor tissues are barely visible, the average Dice score between human expert raters drops down to $0.78 \pm 0.11$.

### 3.1.2 Deviation from Ground Truth

As can be seen in Fig. 3, the average Dice score before chemotherapy of human experts in comparison to ground truth is $0.93 \pm 0.05$. 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.

3.2 Volume Variability

Tumor expansion after preoperative chemotherapy is an important metric used to categorize patients as high-, intermediate- or low-risk candidates. High-risk patients receive an additional postoperative chemotherapy aligned with an irradiation. Therefore, an accurate determination of tumor volume is critical. The clinical volume equals the volume information used in therapy and treatment planning. It approximates the tumor by an ellipsoid shape. It is computed as \( \text{width} \times \text{height} \times \text{depth} \times 0.524 [18] \), where \( \text{width} \), \( \text{height} \) and \( \text{depth} \) of tumor denote the maximal expansion of tumor tissue on MR images\(^3\). Starting with the assumption that the true tumor volume is found through the consensus of our five human experts, Fig. 4 compares human expert annotations and clinical volumes in terms of percental volume differences in relation to the ground truth volume before and after chemotherapy, respectively. It turns out that clinical volumes differ before chemotherapy on average by \( 22.62 \pm 16.12 \% \), and after chemotherapy by \( 35.07 \pm 41.01 \% \) from the ground truth volumes. Before and after chemotherapy, clinical volumes are on average smaller than the ground truth volume, i.e. 85.71 \% before and 92.86 \% after chemotherapy. In contrast, human experts differ before chemotherapy on average by 10.58 \pm 5.90 \% and after chemotherapy by 25.98 \pm 34.57 \% from the ground truth volume.

This shows that assuming an ellipsoid shape for Wilms’ tumors is an erroneous oversimplification, and human expert annotations are helpful to determine tumor volumes more precisely.

4 Evaluation of Segmentation Algorithms

In the following, we conduct example evaluations on our new benchmark data with six fully-automatic and one semi-automatic segmentation method:

- Chan-Vese active contours [9, 10] with two level sets.
- Entropy Rate Superpixel Segmentation [29].

\(^3\)The volume of the largest ellipsoid that fits in a cuboid is \( \pi/6 \approx 0.524 \) times the cuboid volume.
Table 2: Interoperator variability before and after chemotherapy in terms of Dice score.

<table>
<thead>
<tr>
<th></th>
<th>Rater-1</th>
<th>Rater-2</th>
<th>Rater-3</th>
<th>Rater-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Chemotherapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater-2</td>
<td>0.85 ± 0.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater-3</td>
<td>0.89 ± 0.08</td>
<td>0.89 ± 0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater-4</td>
<td>0.85 ± 0.13</td>
<td>0.90 ± 0.05</td>
<td>0.88 ± 0.08</td>
<td></td>
</tr>
<tr>
<td>Rater-5</td>
<td>0.83 ± 0.13</td>
<td>0.89 ± 0.05</td>
<td>0.87 ± 0.07</td>
<td>0.89 ± 0.05</td>
</tr>
<tr>
<td>Post-Chemotherapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater-2</td>
<td>0.63 ± 0.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater-3</td>
<td>0.83 ± 0.24</td>
<td>0.65 ± 0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater-4</td>
<td>0.84 ± 0.10</td>
<td>0.65 ± 0.36</td>
<td>0.80 ± 0.24</td>
<td></td>
</tr>
<tr>
<td>Rater-5</td>
<td>0.84 ± 0.10</td>
<td>0.64 ± 0.35</td>
<td>0.80 ± 0.24</td>
<td>0.89 ± 0.05</td>
</tr>
</tbody>
</table>

- Random-forest classification [22, 7], either with intensities or HOG-features [11].
- Segmentation with a U-Net [42].
- Our recent semi-automatic segmentation method [35].

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.

4.1 Chan-Vese Active Contours

Let us consider a cubic data domain $\Omega \subset \mathbb{R}^3$ and a volumetric data set $f : \Omega \rightarrow \mathbb{R}^3$. In our setting, the co-domain describes the different MRI modalities T2, T1, and T1c. Then a segmentation of $f$ by means of the Mumford–Shah cartoon model [36, 37] minimizes the energy functional

$$E(u, C) = \sum_i \int_{\Omega_i} \|u - f\|^2 dx + \nu \ell(C).$$

In this general formulation, the data domain $\Omega$ is split in an a priori unknown number of segments $\Omega_i$. Here the function $f$ is approximated by a piecewise
constant function $u$, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^3$, and $C$ are the segment boundaries with a (Hausdorff) length of $\ell(C)$. The first term of the energy favors homogeneity within each segment, while the second term penalizes long segment boundaries.

A reduction of this functional to binary segmentation leads to the Chan-Vese active contour model [9, 10]

$$E(u, C) = \lambda_{\text{in}} \int_{C_{\text{in}}} \|u_{\text{in}} - f\|^2 d\xi + \lambda_{\text{out}} \int_{C_{\text{out}}} \|u_{\text{out}} - f\|^2 d\xi + \nu \ell(C)$$

where $u_{\text{in}}$ and $u_{\text{out}}$ are the arithmetic means of $f$ inside and outside the segment boundaries $C$, respectively. The weights $\lambda_{\text{in}}$ and $\lambda_{\text{out}}$ control the influence of each region to the final partitioning.

4.2 K-means Clustering

K-means clustering [31, 30] is a method of vector quantization to partition $n$ 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 $f : \Omega \to D$, $D \subset \mathbb{R}^3$, $\Omega \subset \mathbb{R}^3$, k-means minimizes

$$E(D_1, D_2) = \int_{D_1} \|\xi - u_1\|^2 d\xi + \int_{D_2} \|\xi - u_2\|^2 d\xi$$

$$D = D_1 \cup D_2, \quad D_1 \cap D_2 = \emptyset,$$

where $u_1$ and $u_2$ are the arithmetic means of both classes. In this case, k-means clustering is equivalent to Otsu’s method [39, 28].

4.3 Support Vector Machine

Support Vector Machines [6] 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 image points. In our application, we use a five-fold cross validation to find optimized hyperparameters. Training was performed using MATLAB 4 and the problem was solved via Sequential Minimal Optimization [16]. Furthermore, we used radial basis functions as kernels and classification error, i.e. the weighted fraction of misclassified observations, as loss function.

4www.mathworks.com/products/matlab
4.4 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 [22, 7] are ensemble approaches for classification combining a group of decision trees. A single tree is highly sensitive to noise, while the average of many de-correlated 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 $K$ by sampling from the original training set $Y$ uniformly and with replacement. In this way, decision trees are de-correlated by using different training data. Additionally, random forests use feature bagging, i.e. features are randomly sampled for each decision tree [23]. To estimate how well the results can be generalized, we use 2-fold-cross validation, i.e. we train two sets of models.
4.5 Entropy Rate Superpixel Segmentation

The method of Liu et al. [29] 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 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.

4.6 U-Net

In many areas of medical imaging processing, deep learning and especially CNNs have proven to be very powerful tools. Within these, the U-net architecture [42] is one of the standard convolutional networks in the field of medical image segmentation. It learns segmentation in an end-to-end setting and only needs a few training examples. 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 dramatically and it is necessary to balance the classes. This is done in three steps: first we determine the connected components 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. This is then rescaled to a size of $512 \times 512$ pixels. We use the implementation presented in [1] to solve our segmentation problem and set up the network with 3 layers, batch size 5, 350 epochs, a dropout criterion of 0.5 and 32 root features.

4.7 Robust Interactive Segmentation

Our interactive segmentation approach [35] is a recent method where preliminary evaluations gave good results on Wilms’ tumor data. Since it is fairly advanced, we provide some more details to clarify the key ideas. We follow [8] and consider a minimal partitioning problem of the cubic image domain
\( \Omega \subset \mathbb{R}^3 \) into \( \Omega_1, \ldots, \Omega_n \subset \mathbb{R}^3 \) non-overlapping regions

\[
\min_{\Omega_1, \ldots, \Omega_n \subset \Omega} \frac{1}{2} \sum_{i=1}^{n} \text{Per}(\Omega_i; \Omega) + \sum_{i=1}^{n} \int_{\Omega_i} h_i(x) \, dx,
\]

s.t. \( \Omega = \bigcup_{i=1}^{n} \Omega_i \), \( \Omega_i \cap \Omega_j = \emptyset \), \( \forall \, i \neq j \) \hspace{1cm} (5)

where \( \text{Per}(\Omega_i; \Omega) \) denotes the perimeter of region \( \Omega_i \) inside \( \Omega \), and \( h_i : \mathbb{R} \to \mathbb{R}^+ \) are potential functions reflecting the cost for each pixel being assigned to a certain label \( i = 1, \ldots, n \). To align image and region boundaries, the perimeter is commonly measured in a metric induced by the underlying image \( f : \Omega \to \mathbb{R}^3 \). In this application, we weight the perimeter \( \text{Per}_g(\Omega_i; \Omega) \) of region boundaries in the metric

\[
g(x) = \exp\left(-\mathcal{E}(x)^\beta / \bar{\mathcal{E}}\right), \quad \bar{\mathcal{E}} := \frac{2}{|\Omega|} \int_{\Omega} |\mathcal{E}(x)| \, dx.
\]

Here \( \mathcal{E} : \Omega \to \mathbb{R} \) is the output of the fast structured edge detector of \([14, 15]\) and \( \beta \) is a positive parameter. Assume a (measurable) set of user-scribbles \( S_i \subset \Omega \) for each label \( i \) is given. We define the potential functions \( h_i(x) \) in (5) as the negative logarithm of

\[
\hat{h}_i(x) = \begin{cases} 
\left\{ \frac{1}{|S_i|} \int_{S_i} G_\rho \, G_\sigma \, dy \right\}_{\text{scale}}, & x \notin S_j, \\
1 - \zeta, & x \in S_j, i = j, \\
\zeta / (n-1), & x \in S_j, i \neq j,
\end{cases}
\]

and

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

Here \( \{ \text{scale} \} \) denotes linear rescaling to \([0, 1]\), \( |S_i| \) is the area occupied by \( i \)th label, \( \zeta \) is the assumed probability for a scribble being correct, and \( k_{\sigma} \) and \( k_{\rho_i} \) are Gaussians with standard deviation \( \sigma \) in intensity space and adaptive standard deviation \( \rho_i(x) = \alpha \inf_{y \in S_i} |x - y| \) 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 \( x \) to the closest scribble location.

4.8 Results

In Tab. 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice Score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre-Chemotherapy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV [9]</td>
<td>0.57</td>
<td>0.48</td>
<td>0.69</td>
</tr>
<tr>
<td>k-means [30] (INT)</td>
<td>0.53</td>
<td>0.76</td>
<td>0.41</td>
</tr>
<tr>
<td>Superpixel [29]</td>
<td>0.41</td>
<td>0.33</td>
<td>0.56</td>
</tr>
<tr>
<td>SVM [6] (INT + HOG [11])</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>RF [7] (INT + HOG [11])</td>
<td><strong>0.92</strong></td>
<td><strong>0.92</strong></td>
<td>0.91</td>
</tr>
<tr>
<td>U-net [42]</td>
<td>0.64</td>
<td>0.49</td>
<td><strong>0.94</strong></td>
</tr>
<tr>
<td>U-net + PP [42]</td>
<td>0.74</td>
<td>0.62</td>
<td>0.91</td>
</tr>
<tr>
<td>semi-automatic [35]</td>
<td>0.88</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td><strong>Post-Chemotherapy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV [9]</td>
<td>0.41</td>
<td>0.32</td>
<td>0.58</td>
</tr>
<tr>
<td>k-means [30] (INT)</td>
<td>0.35</td>
<td>0.50</td>
<td>0.27</td>
</tr>
<tr>
<td>Superpixel [29]</td>
<td>0.41</td>
<td>0.29</td>
<td>0.68</td>
</tr>
<tr>
<td>SVM [6] (INT + HOG [11])</td>
<td>0.68</td>
<td>0.69</td>
<td>0.67</td>
</tr>
<tr>
<td>RF [7] (INT + HOG [11])</td>
<td>0.81</td>
<td>0.73</td>
<td><strong>0.92</strong></td>
</tr>
<tr>
<td>U-net [42]</td>
<td>0.30</td>
<td>0.25</td>
<td>0.61</td>
</tr>
<tr>
<td>U-net + PP [42]</td>
<td>0.39</td>
<td>0.40</td>
<td>0.46</td>
</tr>
<tr>
<td>semi-automatic [35]</td>
<td><strong>0.84</strong></td>
<td><strong>0.80</strong></td>
<td>0.89</td>
</tr>
</tbody>
</table>

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 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. 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. As a remedy, we perform a post-processing step where we remove all segments with a large ratio between the boundary length and the squared segment area; see Fig. 5. This leads to a dramatic increase in mean precision.

Overall, the random forests provide the best results before chemotherapy, but are also competitive after chemotherapy. Therefore, we suggest random forests trained on HOG-features as well as intensities as the baseline fully automatic 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. These observations highlight the challenges in the data set.

The Robust Interactive Segmentation approach [35] 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 from Section 3 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 is the second best method before chemotherapy, and the leading approach after chemotherapy, yielding the highest Dice score and precision. Since the tumor volume after chemotherapy is decisive for postoperative treatment planning, it is the optimal method for this purpose. The segmentation quality of both the random forests and the semi-automatic segmentation lies within the variability of human experts. Thus, if a fully automatic approach is needed, we recommend random forests. For achieving even higher quality, one should use the semi-automatic approach which benefits from the knowledge of the expert who provides the scribbles.

5 Conclusion

We have proposed the first multi-modal benchmark for segmentation of Wilms’ tumors. In spite of the fact that the data set is only medium in size, its amount of information is rich: There are multi-sequence MRT images for all patients, and for eleven patients both pre- and post chemotherapy images. That is supplemented by manual annotations by five independent human experts, as well as histological diagnoses.

Our benchmark allows several important conclusions:

We have demonstrated that human expert annotations can reveal a large interoperator variability especially after pre-operative chemotherapy. Furthermore, we have shown that the popular tumor volume determination based on ellipsoid shapes tends to be highly erroneous.

Our data set also allowed to evaluate seven computer-based algorithms. At this time, all fully-automatic segmentations apart from random forests underestimate the tumor volume compared to human expert raters. Thus, their precision is insufficient, especially after chemotherapy. Our experiments indicate that the semi-automatic method given by the Robust Interactive Segmentation approach [35] is the most appropriate tool for 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.

In our ongoing research, we plan to include more anatomical knowledge into our segmentation models and to constantly enlarge the number of available data sets. It is our hope that our benchmark data set for segmentation of nephroblastoma will stimulate a growing interest in this research field which is challenging both from a medical and a computer vision viewpoint. Most importantly, we are confident that the resulting progress will help to maximize the survival chances of the affected children.
Acknowledgment

This work was partially funded by the EU FP7 under the project Computational Horizons in Cancer (grant agreement No 600841).

References


