A New Calibration Approach to Graph-based Semantic Segmentation

Mateus Riva

Dissertation
presented to the
Institute of Mathematics and Statistics
of the
University of São Paulo
as part of the Master's Program

Program: Computer Science

Supervisor: Prof. Dr. Roberto M. Cesar Junior

São Paulo, March 2018

Resumo

RIVA, M. **Uma Nova Técnica de Calibração para Segmentação Semântica Baseada em Grafos**. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.[1]

Apresentamos um método de calibração da segmentação semântica de imagens baseada em Grafos Estatísticos-Relacionais (GERs), com um foco particular em Imagens de Ressonância Magnética (IRM) pediátricas. O GER provê uma representação de uma cena estruturada, descrevendo tanto os atributos de cada objeto de interesse quanto a natureza de seus relacionamentos, por exemplo a posição relativa no espaço. Cada vértice no grafo representa um objeto de interesse e cada aresta representa um relacionamento entre dois objetos. A segmentação semântica pode ser feita realizando um casamento entre um GER construído a partir de uma imagem observada com um GER modelo previamente construído. Nós desenvolvemos um método de calibração para verificar a qualidade da segmentação baseada em GER dado um conjunto de parâmetros, assim como uma exploração de diversos conjuntos de parâmetros aplicados à segmentação de IRM. Apresentamos a validade e utilidade da técnica de calibração, junto de resultados preliminares de segmentação de dados IRM reais. Adicionalmente, discutimos trabalhos futuros na melhoria de segmentação de dados reais utilizando GERs.

Palavras-chave: métodos estruturais, visão computacional, imageamento de ressonância magnética, segmentação baseada em grafos.

Abstract

RIVA, M. **A New Calibration Approach to Graph-based Semantic Segmentation**. Dissertation (Master's) - Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2018.[2]

We introduce a calibration method for semantic segmentation of images utilizing statistical-relational graphs (SRGs), with a particular focus on pediatric Magnetic Resonance Imaging (MRI). The SRG provides a representation of a structured scene, describing both the attributes of each object of interest and the nature of their relationships, such as relative position in space. Each vertex in the graph represents an object of interest and each edge represents the relationship between two objects. Semantic segmentation can thus be performed by matching an SRG built from an observed image to a previously-built model SRG. We develop a calibration method for assessing the quality of SRG segmentation given a set of parameters, as well as an exploration of several sets of parameters applied to MRI. We present the validity and usefulness of the calibration technique, along with preliminary results on real MRI data segmentation. We additionally discuss future work on improving real data SRG-based segmentation.

Keywords: structural methods, computer vision, magnetic resonance imaging, graph-based segmentation.

Summary

List of Abbreviations

SRG	Statistical-Relational Graph
CT	Computarized Tomography
MRI	Magnetic Resonance Image(ing)
ME	Multi-echo
CV	Computer Vision
USP	University of São Paulo
ICr	Children's Institute (*Instituto da Criança*)
FM	Faculty of Medicine of the University of São Paulo
EM-MRF	Expectation Maximization and Markov Random Field

Chapter 1

Introduction

1.1 Motivation

With the recent advances in medical imaging technology, an immense amount of patient data is now available to doctors, researchers and health-care professionals. Modern non-invasive techniques for acquiring data from inside the body, such as computerized tomography (CT) scan and magnetic resonance imaging (MRI), provide three-dimensional, high-resolution models of organs, bones and other body structures. However, an increase in data volume means that the time- and work-intensive method of manual medical image analysis is no longer feasible, requiring the development of faster, computer-assisted, automatic analysis methods.

Computer-based analysis of MRI data has recently been used in Alzheimer's disease detection [WZL+16], breast cancer subtype identification [GZM15], glioma segmentation [BZS+15], left ventricle segmentation [WPC+15], and many other applications. Methods for automatic segmentation of regions of interest in MRI are many and vary according to their goal: segmentation of multiple sclerosis lesions [GLFN+13], segmentation of brain tumors [FHHGM01], and detection and segmentation of necrotic femoral heads [ZSN+04], for example. A large amount of academic work in this area aims to segment the tissues in a brain MRI (typically but not exclusively white matter, gray matter and cerebrospinal fluid) [RGC+97, SSLS+01, FSB+02, ZBS01].

Despite significant interest in the area of automatic MRI analysis over the last two decades, most of the papers published deal with the analysis of standard adults. Thus, cases that diverge from this norm, such as very young children [DCSA15], whose developing bodies present many differences to the commonly studied adult body [DCSA15, NPD+13, HWK+98], are often overlooked. Additionally, most segmentation methods focus on healthy patients, or on a single anomaly (such as a specific type of tumor) [HAB17].

There is a demand for methods that are capable of dealing with the large diversity of cases available – developing body structures, extraneous elements such as tumors or missing elements such as organs removed by surgery – in the corporal structure without significant retraining or adjustment. The analysis of images from such individuals would benefit immensely from the development of a framework capable of detecting and adapting to these cases.

Recent scientific activity points to various applications of brain MRI to pediatric neuro-oncology, such as diagnosis and treatment of acute lymphoblastic leukemia [BEG+15], brain tumor analysis [GR17, GMJ+13], the damage of pediatric radiation treatment to neural tissue [MEBT+16, SLM+13] and early diagnosis of brainstem glioma [PKV+11]. This activity further demonstrates the necessity of works focusing on methods for automatic pediatric MRI analysis.

1

One potential approach to address the aforementioned aspects relies on the Statistical-Relational Graph (SRG) [Gra12, GCB14], which is based on combining the supervised recognition of primitive structural elements of a given object and their spatial relationships. In her thesis, Graciano applies this approach to the segmentation of abdominal organs in normal, adult patient MRIs, utilizing classic probabilistic pattern recognition techniques to attempt to segment organ tissues while simultaneously considering the spatial relationship between distinct organs in order to improve segmentation quality. We believe this method may be further developed for the semantic segmentation of liver divisions in pediatric MRIs. Semantic segmentation is the process of finding the distinct regions of an image, and labeling them, and is explained in further detail in Section 2.1.3.

Applications for pediatric abdominal organ segmentation, with a focus on segmentation of livers in MRI images, are of particular interest to FM-USP, as this is an important ongoing project. This application would aid doctors at Instituto da Criança (Children's Institute, ICr-FM), which require both an automatic method for diagnosing high iron content in children's livers without resorting to biopsies and an automatic method for assessing liver volumetry for organ transplants.

Children with hemoglobinopathies, that is, diseases affecting the red blood cells, are required to undergo several blood transfusions. Due to the transfusions, a large number of these children develop iron buildup in hepatic tissue, a condition which is immediately benign but paves the way for further complications along the patient's life. However, standard techniques for accurately measuring iron content in the liver often require liver biopsies, which are both invasive exams, and may suffer from sampling errors (as only a small sample of the liver is measured, and the iron buildup may be located elsewhere in the organ) [KGSH90].

Modern techniques allow for an estimate of liver iron content through the use of a specific MRI acquisition sequence; however, this does require the segmentation of the liver regions by the radiologist, which is a time-intensive task. Not only this technique has, due to recent advances, achieved accuracy similar to the biopsies, but it also eliminates the two complications described above: MRI is completely non-invasive, and it can also map the iron buildup in the entire liver at once. Additionally, liver transplants often benefit – or even outright require – an estimate of the volume of the liver subdivision being transplanted. Naturally, such a measure may be extracted directly from MRI data (as it does contain the volume of each voxel in its metadata); however, this once again would require the time-intensive task of manual segmentation of the liver subdivision by the radiologist.

1.2   Goals

The goal of this work is to develop a set of methods for the calibration of statistical-relational graph-based semantic segmentation of medical images. We propose a new structural method for model learning and application to medical image semantic segmentation, based on previous work by [Gra12]. We then propose a new calibration approach to this method, and demonstrate its applicability in the exploration of parameter sets for SRG-based medical image segmentation.

1.3   Contributions

In this project, we aim to contribute with mathematical models and associated algorithms for the representation and understanding of features and structural information in relevant data. In particular, we intend to focus on applications for pediatric abdominal imaging, where the volume and iron content of liver substructures is an important information for the physician, but expensive

to manually obtain. The project is being developed in a collaboration with Professors Isabelle Bloch (Telecom ParisTech), Carlos Moreira Filho, Lisa Suzuki and Marcelo Straus Takahashi (FM - USP).

The process of semantic segmentation of medical images based on the Statistical-Relational Graph is highly complex in nature, possessing a large amount of variables and parameters. Keeping this is mind, a calibration technique for the process was devised. This calibration technique intends to be the cornerstone of all future work in this project involving SRG-based semantic segmentation. It will both guide the practical aspects of implementation, debugging, visualization and testing, and the theoretical aspects of proper parameter configuration, analysis of the discrimination of attributes. Additionally, it will provide a test-bed for all graph-based segmentation ideas, algorithms, and improvements designed.

The calibration of SRG segmentation is fundamentally centered on providing the full segmentation pipeline in a controlled environment. By generating synthetic data for use in both training and testing of SRG techniques, with known and controlled noise profiles, and with good quality measures and knowledge of the expected results, the calibration enables the exploration of the parameters involved in medical image SRG-based segmentation. It provides measures of each parameter's overall estimated quality and of their impact in the quality of the process, both absolute and relative to one another.

1.4 Organization

This document is organized as follows. Chapter 2, outlines the basic concepts involved in our project, and reviews related papers on these subjects. Chapter 3 describes the methods and techniques used to pursue the goals of this dissertation. Chapter 4 provides the setup for experiments conducted, as well as their results and discussion. Finally, Chapter 5 concludes this document, its results, and discusses related future work.

Chapter 2

Concepts and Bibliography

In this Chapter we cover important topics related to our project. Specifically, we will cover the automatic analysis of medical images utilizing computer vision techniques in Section 2.1, and structural-based or hybrid techniques for representation and segmentation of images in Section 2.3.

2.1 Computer-assisted Medical Image Analysis

Computer-assisted medical image analysis is a significant field in computer science, spanning more than four decades [SRE12] of work and still extremely active [HLO⁺18, LKB⁺17]. As such, a complete and comprehensive review of the field is out of the scope of this work. This section will describe the basic concepts involved and focus on papers related to the project's task, that of pediatric Magnetic Resonance Imaging segmentation.

2.1.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging, or MRI, is a medical image acquisition technique, which is based on nuclear magnetic resonance. Through the controlled emission of radio waves inside a strong magnetic field, the nuclei of hydrogen atoms are stimulated, resulting in the emission of measurable radio frequency (that is, electromagnetic waves with frequency in the megahertz [MHz] range) which in turn is used to reconstruct images of the interior of the body, with the intensity of each pixel being dependent on the tissue represented. Different sequences of radio wave emission are able to produce different contrasts between tissues, by exploiting the properties of nuclear magnetic resonance.

Through the emission and absorption of radio waves, a measurement of the concentration of an element's protons on a given tissue is obtained. Many elements may be measured by leveraging their nuclei magnetic properties – though elements with even atomic number (i.e. number of protons in an atom) and mass number (i.e. number of protons and neutrons in an atom) may not, as the angular momentum of neutrons and protons cancel each other out. Hydrogen atoms, however, have an odd atomic number (1), well-known magnetic properties, and are the most abundant in the body (as the body is composed primarily of water, which itself is composed of two hydrogen and one oxygen atom). Thus, medical MRI generally measures the hydrogen content of tissues.

Protons inside atomic nuclei rotate around their axes of orientation, a property known as "nuclear spin". As protons are electrically charged, this rotation generates magnetic momentum perpendicular to the axis. This generates a microscopic magnetic field around each particle. Due to the

random orientation of each particle, however, the macroscopic ensemble of particles tends to have no significant magnetic field.

In order to obtain this magnetic momentum in a macroscopic context, the orientation of the protons may be forcibly aligned by a strong external magnetic field. While under the influence of a strong external field with a vector of magnetic force $\vec{B_0}$, the resultant spin of all protons tend to align in the same direction as $\vec{B_0}$, in one of two orientations: parallel (with the same orientation as $\vec{B_0}$) and anti-parallel (with the opposite orientation of $\vec{B_0}$). The parallel protons are in the "state of minimum energy", while the anti-parallel are in the "state of maximum energy".

As these protons are now aligned, their magnetic momenta coalesces. This net magnetization, defined by M, is in theory directly proportional to the concentration of each specific particle in a given region and, thus, we are interested in its measurement. Measurement of magnetic fields is often done through magnetic induction, which requires an alternating magnetic field. However, if the protons are perfectly aligned to $\vec{B_0}$, M is unchanging. By adding precession to the protons, we will generate an alternating magnetic field which can in turn be measured.

Radio-frequency pulses are then used to generate a magnetic field $\vec{B_{xy}}$, orthogonal to $\vec{B_0}$. All protons possessed of spin are precess around the axis of $\vec{B_0}$ with a specific frequency ω_0, given by Larmor's Frequency: $\omega_0 = -\gamma|\vec{B_0}|$, where γ is a given particle's gyromagnetic ratio (for an hydrogen atom ^1H, $\gamma = 42.58 MHz/T$). By alternating $\vec{B_{xy}}$ at a given particle's Larmor frequency, we obtain resonance, inducing a sinusoidal current at frequency ω_0, whose signal M we measure. By keeping $\vec{B_{xy}}$ active, a greater flip angle (that is, the angle between the precession axis and the particle's axis) may be achieved. The highest measured signal, naturally, occurs when the flip angle is $90°$.

When $\vec{B_{xy}}$ is stopped, the proton rotates back to its original position, spinning around the axis of $\vec{B_0}$. As it does so, the measured signal decays (as the field no longer alternates), eventually reverting to zero. This decay is called *T1-relaxation*. Another effect is also measured: the protons are all precessing in-phase while $\vec{B_{xy}}$ is active (if they were not, then the measured signal's amplitude would be too small to measure). When $\vec{B_{xy}}$ is stopped, the proton's spins go slowly out of phase; this dephasing process is called *T2-relaxation*.

As the protons dephase, the transverse component of the measured magnetic field, M_{xy}, decays according to the formula $M_{xy}(t) = M_{xy0}e^{-\frac{t}{T_2}}$, where M_{xy0} is the maximum value of the transverse component (when all protons are in phase). The time taken for M_{xy} to be equal to 37% of M_{xy0} is called T_2 or T2, and is highly tissue-dependent, as it is directly related to the environment surrounding a particle.

As the flip angle decreases, the longitudinal component of the measured magnetic field, M_z, increases according to the formula $M_z(t) = M_{z0}(1 - e^{-\frac{t}{T_1}})$, where M_{z0} is the maximum value of the longitudinal component (when the flip angle is $0°$). The time taken for M_z to be equal to 63% of M_{z0} is called T_1 or T1, and is related to the viscosity of a given tissue.

Thus, given a specific external field and knowledge of the expected concentrations of hydrogen in each tissue (often acquired empirically), each tissue in a magnetic resonance image will have specific T1 and T2 values.

In order to measure these times, a sequence of radio-frequency pulses is applied. The most common of these sequences is the "spin-echo" sequence, where a $90°$ pulse is applied to flip the protons, and after a given amount of time t, an $180°$ pulse (the inversion pulse) is emitted to focus spins that went out-of-phase due to local field discrepancies. After an equal amount of time t, the signal peaks in an "echo" which allows for measurement of T2 and, as the protons' spins return to the longitudinal axis, of T1.

Depending on the times between the pulses and the measurements, an image may be "T1-weighted" or "T2-weighted". These are often treated in much the same way as distinct contrasts for X-ray, which tissues having which specific intensities depending on the nature of the acquisition. Generally, T1 images have high values for fat and slow-flowing blood, while having lower values for tissues with high water content; T2 images have high values for tissues with high water content. T1- and T2-weighted images of the same region of a single patient can be seen in Figure 2.1.
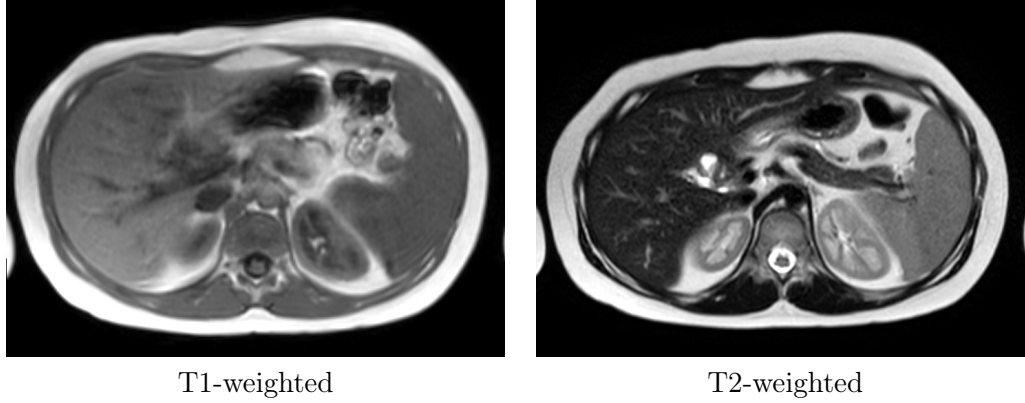

T1-weighted   T2-weighted

Figure 2.1: *Comparison of T1- and T2-weighted images. Both slices are from the same region of the same patient. Images courtesy of FM-USP.*

A stronger magnetic field (that is, with a greater magnitude for $\vec{B}_0$) tends to have a higher ratio of parallel to anti-parallel protons. As the magnetic momenta generated by anti-parallel protons cancels out part of the momenta generated by the parallel protons, a stronger external magnetic field results in an increase of the acquired images' quality. However, there is a practical limit to this: stronger magnetic fields often require a higher amount of resources to maintain, and the equipment required for maintaining such a field may overheat the acquisition area above the limits of the human body.

In addition to the "spin-echo" sequence described above, a great number of pulse sequences have been proposed, for a wide array of applications. Many are designed to put into evidence an specific type of anatomical structure, such as hyper-intensifying multiple sclerosis plaques [GPU$^+$05]. Yet others are used to obtain useful, quantifiable information from the composition of the tissues, such as iron content [HCH$^+$05].

Through the utilization of a gradient magnetic field, distinct "slices" of a patient may be measured in turn, which are then reconstituted into a 3D image. Due to the nature of this process, MR images tend to have much higher resolution on the plane of capture than on the others, which are reconstituted from the acquired slices. Three planes are generally used: the *axial*, which runs perpendicular to the axis that runs from head to feet and is the most commonly used; the *sagittal*, which runs perpendicular to the axis that runs from left to right; and the *coronal*, which runs perpendicular to the axis that runs from front to back.

With an accurate and non-invasive view of the interior of the body, MRI is often used in diagnostic medicine. The investigation of a patient's anatomy, which may be inferred from the medical image based on the contrasts between tissues, allows for the discovery and analysis of a diverse spectrum of pathologies and anomalies, and serves as guidance for many applications, ranging from pharmacological treatment to surgery planning [Sue02].

Figure 2.2 shows a T2-weighted MRI axial slice of a patient's lower abdomen, where the liver, kidneys, spleen, pancreas and gastrointestinal tract can be seen.
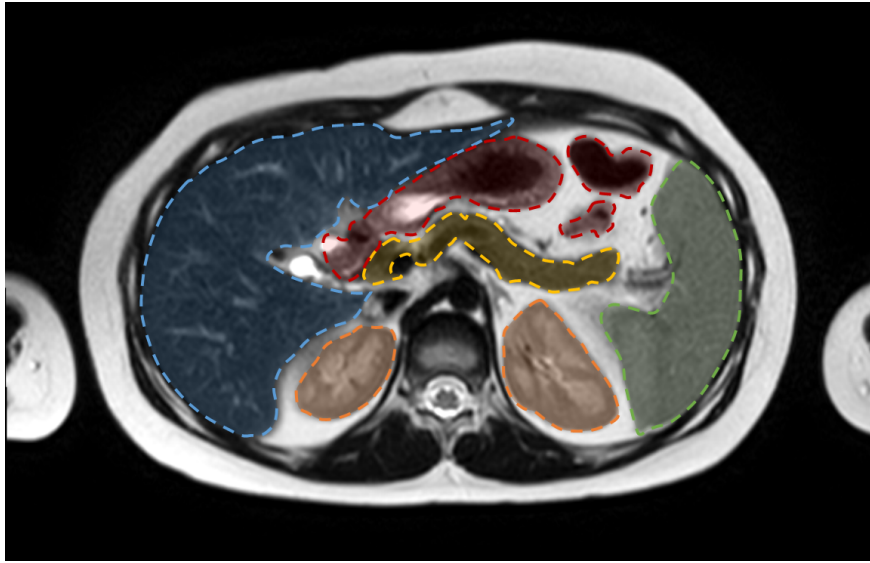
Figure 2.2: *Example of a MR image. Some structures are highlighted: the liver in blue; the kidneys in orange; the pancreas in yellow, the gastrointestinal tract in red and the spleen in green.*

2.1.2   Pediatric Magnetic Resonance Imaging

The developing bodies of babies and children present specific challenges to the analysis of medical images acquired from them. A common hurdle in the analysis of neonatal brain MRI is the inversion of the contrasts in white and gray matter, which happens due to the process of myelination that children undergo in their first few years of life [DCSA15]. Figure 2.3 compares a MRI of the brain of a neonate and of an adult. A general difficulty in the acquisition of pediatric MRI are patient motion artifacts [DCSA15], that is, image artifacts and noise generated by the patient moving during said image's acquisition, which is much more common in infants and small children. Examples of ringing artifacts caused by motion are presented in Figure 2.4. Finally, the high variability of the developmental process compounds the intrinsic inter-patient variability, producing significant corporal variation even among individuals with small age differences [HWK+98, NPD+13].

2.1.3   Semantic Segmentation of Medical Images

In computer vision, segmentation is the process of extracting one or more objects in an image from the background, and recognition is the process of obtaining useful descriptors or labels for these objects. Recognition is also called "classification", as the process of attributing classes to the objects of interest. We thus define semantic segmentation as the process of segmenting and labeling object pixels of interest in a given image [MNA16].

For example, given a slice of a Magnetic Resonance Image of the lower abdomen of a person, we may want to acquire the left and right kidneys. In order to do so, first we must extract our "objects of interest" – in our case, the kidneys – from the "background" – in our case, the rest of the image, including the remainder of the body and the black region. In addition to extracting, we also want to determine which kidney is the left one and which is the right one, labeling them as such. Figure 2.5 shows the results of these steps.

A great number of tasks in medical image analysis require the segmentation of one or more regions of interest. The traditional method is manual segmentation, which in many cases is an "easy" task for an expert – that is to say, while it may be time-consuming, it will rarely be considered a challenge. However, manual segmentation is both time-intensive and subjective and, as such, is

Figure 2.3: *Comparison between T1 sagittal brain MR images of (a) a 1-day old neonate; (b) an adult. Source: [DCSA15]*

Figure 2.4: *Examples of ringing artifacts caused by patient motion during capture. (a) T1 sagittal brain MRI of a 2-day old neonate; (b) T1 axial abdominal MRI of a 12-year old child. Source for (a): [DCSA15]*
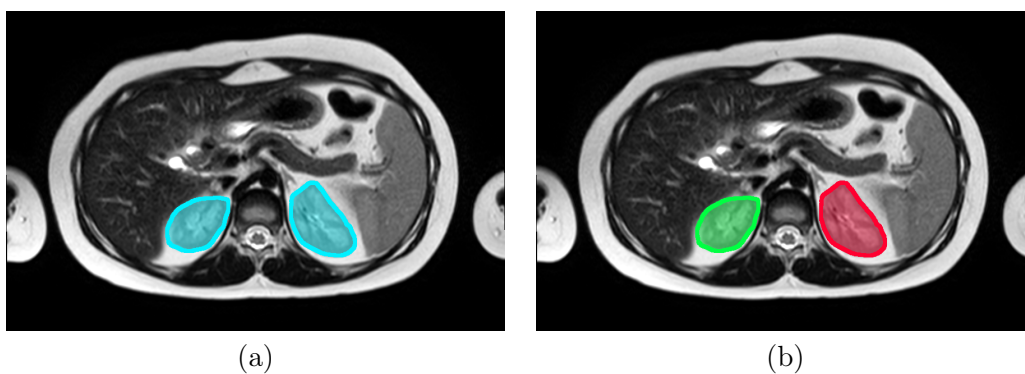

Figure 2.5: *(a) segmentation of the kidneys, distinguishing which pixels represent kidneys and which represent the background, where the "blue" pixels are the kidneys. (b) recognition and attribution of semantic information to the kidneys, where "green" is the left kidney and "red" is the right kidney.*

affected by often significant inter- and even intra-operator variability [MAT$^+$16]. Thus, there is a demand for automatic and semi-automatic methods for the segmentation of regions of interest in medical images. This is reflected by the large amount of work focused on this task, of which we present some of the most relevant.

Lesage et al. [LAFB16] utilize a method based on particle filtering for coronary segmentation in three-dimensional CT angiograms. The extraction of fetal envelopes from ultrasound imaging is explored by Dahdouh et al. [DAGB15], using a segmentation method based on shape priors. Reliable methods for generating models for dosimetry studies, including organs of interest, are proposed in [DVO$^+$16], where infants' 3D MRI data were automatically segmented using the framework detailed in [DB15].

Jaware et al. [JKZ16] explored a combination of classifiers, including k-Nearest Neighbors, Multi-Kernel Support Vector Machine and Neural Networks to segment and classify neonatal brain MRI. K-means segmentation of neonatal brain MRI is explored in Udyakumar et al [USG$^+$16]. Hareendranathan et al. [HZM$^+$16] utilize a prior-based Random Walker segmentation formulation for infant MRI hip segmentation. Moeskops et al. [MVM$^+$16] propose the use of a deep Convolutional Neural Network for infant brain MRI segmentation. Weisenfeld and Warfield [WW09] utilize a classifier fusion-based approach to supervised classification of newborn brain tissues in MRI. In his thesis, Moeskops [Moe16] proposes three layers of classifiers for infant brain MRI segmentation, utilizing k-Nearest Neighbors and SVM classifiers. His work mentions the need for proper segmentation of white matter, gray matter and cerebrospinal fluid in infants' MRI.

Xue et al. [XSJ$^+$07] utilize Expectation Maximization and Markov Random Fields (EM-MRF) to solve the problem of infant brain segmentation. Mislabeling detection and local segmentation thresholds are utilized to minimize errors in the process. The paper mentions many of the problems with segmenting said images, as the brain undergoes development until the child's second year and, as such, any atlas or models must consider age of capture as a fourth dimension. Based on [XSJ$^+$07], Kuklisova et al. [KMAS$^+$11] built a 4D (age-specific) atlas for fetuses in the third trimester of pregnancy. Serag et al. [SAB$^+$12] propose a 4D extension of Seghers et al.'s method [SDM$^+$04], using kernel regression, for building a 4D atlas in the same age range. A subject-specific atlas approach was proposed by Shi et al. [SFT$^+$10], where later images of a subject were used to build an atlas for segmentation of the same subject's neonatal brain MR images. Kim et al. [KFD$^+$13] propose the use of an Intensity Growth Map, created by establishing voxel-wise correspondence between manually segmented paired MR images taken at 1 and 2 years old (each pair belonging to a single subject). The Intensity Growth Map is then used to enhance Expectation Maximization for 1 year old brain MRIs of different subjects. Cardoso et al. [CMK$^+$13] utilize an algorithm based on maximum *a posteriori* Expectation Maximization, Markov Random Fields and intensity and mislabeling correction to segment infant brain MRI. Their work also devotes significant focus on correctly segmenting abnormal individuals (such as children with ventriculomegaly).

A comprehensive review of liver CT segmentation techniques, before the popularization of deep learning techniques, may be found in [HVGS$^+$09]. This work focuses on some of the techniques applied to the SLIVER 2007 Challenge, a MICCAI challenge on liver segmentation which provided an annotated dataset of liver CT images for participants.

Many works attempt to leverage the greater amount of available data, along with data augmentation techniques, to implement and train neural networks-based segmentation techniques. Due to a great availability of data, several papers have applied convolutional neural networks to the segmentation of Computerized Tomography abdominal images [LWH$^+$17, BCDK$^+$16, CEE$^+$16], as well as brain MRI [ZLD$^+$15, XGB17, HDWF$^+$17]. Milletari et al. [MNA16] have developed a fully convolutional network for volumetric prostate segmentation in MRI, utilizing data augmentation techniques to enhance a dataset of lower-abdominal MRIs. Virzi et al. [VGM$^+$18] explore

the use of patch-based pseudo-3D deep learning techniques for the segmentation of vessels, where distinct neighbouring slices are combined as color channels and fed into an ImageNet-trained neural network. Avendi et al. [AKJ16] utilizes stacked auto-encoders to infer shapes for deformable model-based segmentation of the left ventricle in cardiac MRI.

2.2   Fundamental Concepts from Graph Theory

A graph G is defined as a tuple $G = (V, E)$, where V is a set of *vertices* (sometimes also called *nodes*) and E is a set of *edges*. An edge $e \in E$ connects two vertices $v \in V$ and, as such, $E \subseteq V \times V$. Graphs where edges connect more than two vertices exist, and are called *hypergraphs*. If V and E are finite sets, then G is a finite graph. We will use the notations $|V|$ and $|E|$ to indicate the cardinality of the V and E sets, respectively. $|V|$ is also called the *order* of G, and $|E|$ the *size* of G.

Each edge $e \in E$ is associated with a pair of vertices of G and defined as $e = (v_i, v_j)$, where $v_i, v_j \in V$. G is called a directed graph, or digraph, when the pair (v_i, v_j) is ordered, with v_i representing the start and v_j the end of an edge. A digraph is symmetrical if, for each edge (v_i, v_j), there is an edge (v_j, v_i). G is called an undirected graph when the pair (v_i, v_j) is not ordered, that is, $(v_i, v_j) = (v_j, v_i)$.

Two vertices v_i and v_j of a graph are called adjacent, connected, or neighbours if and only if they are extremities of an edge of G, that is, if $(v_i, v_j) \in E$ or $(v_j, v_i) \in E$. The set N_E of vertices adjacent to a given vertex $v \in V$ is defined as $N_E(v) = \{v_j \in V : (v, v_j) \in E \text{ or } (v_j, v) \in E\}$. If all vertices of G share an edge with all other vertices, then G is called fully-connected.

In this work, all graphs presented are finite, directed and symmetrical.

2.3   Structural-based Techniques for Semantic Image Segmentation

When faced with a structured scene – that is, an image representing one or more objects whose position (or size, or color, or other property) in relation to other objects and the background are important – humans tend to unconsciously utilize structural information in order to guide the process of semantic segmentation. In addition, linguistic descriptions of a scene often include said structural relations: for a related example, "the left kidney is **located to the left** of the right kidney, and both are **equally spaced** from the vertebral column".

As such, the use of structural relationships to aid in automatic semantic segmentation of images representing structured scenes are a natural step to explore, and over the years has been shown to produce interesting results [Vie97, HAB08, Ede99, CR08].

2.3.1   Structural-based Medical Image Analysis

Due to the nature of human anatomy, most medical images are representations of structured scenes. In fact, structural and relational information is oftentimes essential in the understanding and analysis of anatomical structures [Jac01]. Thus, there is potential for the development of structural-based medical image segmentation techniques.

Colliot et al. [CCB06] utilize fuzzy sets to represent structural relations (specifically, spatial

relations), arguing that said relations are intrinsically imprecise and thus are best represented by techniques which allow for a degree of uncertainty; the structural information is then utilized to constrain a deformable model as an external force. Khotanlou et al. [KCAB09] applied this method to tumor segmentation in 3D brain MRI. A more refined approach based on [CCB06] was proposed by Fouquier et al. [FAB12] for the segmentation of structures in brain MRI, which automatically optimizes the sequence of steps to be taken for the best segmentation result possible. Morel et al. [MXV$^+$16] utilize a max-tree representation to perform semi-automatic segmentation of diffuse excessive high signal intensity (DEHSI) of the white matter of newborn and premature infants. These regions of interest are represented as hyper-intense zones inside the white matter, and their method extracts said regions from within white matter regions.

Graciano's work [Gra12] is the main basis for this project, formally proposing the use of a graph for representation of both statistical attributes of the structures of interest and their relationships in an image, the Statistical-Relational Graph (SRG). An instance of this graph is built in order to describe and represent the kidneys, liver, stomach and spine in a 2D coronal slice of an abdominal MRI, and its robustness and quality of representation is demonstrated.

Chapter 3

Materials and Methods

3.1 The Statistical-Relational Graph

In order to fulfill the project's goals of semantic segmentation of structures of interest in medical images, there is a need for the design of a suitable graph-based structural model for segmentation and parts recognition. In order to accurately model and represent the knowledge found in MRI data, proper understanding of the structural components of the images is required. As shown by [CCB06, FAB12, MXV$^+$16, NAB13], structural knowledge is of primary importance for the interpretation of images where intensity and shape may not be sufficient data for an adequate analysis, such as abdominal and brain MRIs.

To mitigate this deficiency, we propose the usage of graphs and hypergraphs for the representation of anatomical structures, where each vertex represents a region or structure of interest and the (hyper)edges represent relations between these (such as relative spatial positioning). An schematic of such a graph can be found in Figure 3.1. Our challenge is the application of this knowledge in a way that guides and enhances the image interpretation process.

[Gra12]'s work on structural-statistical representation provides the core method which we implement and explore in this work: the Statistical-Relational Graph, or SRG, which describes objects in a structured scene using a combination of statistical attributes extracted from each individual object and attributes extracted from their relationships. We expand the definition of the SRG in Definition 3.1.1.

An overview of the process of SRG-based semantic segmentation is provided. Initially, a model SRG is built to represent the template of the scene which we are interested in segmenting; this model SRG, denoted by G_M, is defined in section 3.1.1. When input data is observed, a super-segmentation algorithm is applied to it, and the resulting regions are used to assemble a super-observation SRG, denoted by G_{super} and defined in Section 3.1.2. Then, a matching solution between G_{super} and G_M is obtained, denoted by S and defined in Section 3.1.2. S is used to map G_{super} to an observation graph, denoted by G_{obs} and defined in Section 3.1.2. G_{obs} represents the semantic segmentation of the observed data, according to the model provided. Visual representations of each element are provided in Figures 3.1, 3.2, 3.3 and 3.4.

3.1.1   Model Description

In order to perform our tasks of segmentation and recognition, we must first build a model capable of extracting and properly representing the structures and objects of interest. This model must be able to discriminate precisely between structures, using both statistic attributes from the structures and the relationships between them.

Definition 3.1.1. *Let us formally define the SRG as a sextuplet* $G = (V, E, A_V, \Theta_V, A_E, \Theta_E)$ *where:*

- V *is the set of n vertices, $V = \{v_1, v_2, ..., v_n\}$ in G. Each vertex represents a single part of the object of interest;*

- E *is the set of m edges, $E = \{e_1, e_2, ..., e_m\}$ in G. Each edge represents a relationship between a pair of parts of the object of interest. In the case of hypergraphs, the edges represent relations among sets of vertices, possibly more than two;*

- A_V *is the set of n random attribute vectors, $A_V = \{a_{1v}, a_{2v}, ..., a_{nv}\}$, where each random vector a_{iv} is obtaining by mapping the j attributes of the corresponding vertex v_i. Thus, $a_{iv} : v_i \mapsto \mathbb{R}^j$;*

- Θ_V *is the set of the n distribution parameters for the random vectors A_V, $\Theta_V = \{\theta_{1v}, \theta_{2v}, ..., \theta_{nv}\}$, where each set of parameters θ_{iv} corresponds to the random vector a_{iv};*

- A_E *is the set of m random attribute vectors, $A_E = \{a_{1e}, a_{2e}, ..., a_{me}\}$, where each random vector a_{ie} is obtaining by mapping the k attributes of the corresponding edge e_i. Thus, $a_{ie} : e_i \mapsto \mathbb{R}^k$;*

- Θ_E *is the set of the m distribution parameters for the random vectors A_E, $\Theta_E = \{\theta_{1e}, \theta_{2e}, ..., \theta_{ne}\}$, where each set of parameters θ_{ie} corresponds to the random vector a_{ie}.*

When building a SRG for an application, the attributes chosen for extraction for both structures (represented by the vertices) and their relationships (represented by the edges) must be capable of properly discriminating that which they represent. Thus, the task of attribute definition is often heavily dependant on domain, as each particular application may be benefited by specific sets of attributes.

Expert knowledge is often useful in attribute definition, as the task of semantic segmentation often resembles the experts' reasoning process; if, for example, a radiologist usually determines that an hepatic vein is the "brighter" region "inside" the liver, then we can infer that intensity, position, relative position and contrast between structures are useful attributes for this given problem.

Once the attributes are defined, a model must be built in order to represent the "expected" scene. This model graph acts much like a template, and future observations of unknown data will be compared against it. Proper model building is important for achieving the task of segmentation and recognition using the SRG, but is not in the scope of this dissertation. This task will be addressed in more detail in future works in this project. In this work, a model is learned through the extraction of relevant attributes from a manually-annotated image. This process is schematized in Figure 3.1.

Definition 3.1.2. *Let us formally define the model graph as a Statistical-Relational Graph $G_M = (V_M, E_M, A_{VM}, \Theta_{VM}, A_{EM}, \Theta_{EM})$ where:*

1. *each vertex in V_M represents an object of interest in the represented scene;*

2. *each edge in E_M represents a relationship between two objects of interest in the represented scene;*

3. *each random attribute vector in A_{VM} and corresponding distribution parameters in Θ_{VM} represents the template or expected value of the attributes of each object in the represented scene;*

4. *each random attribute vector in A_{EM} and corresponding distribution parameters in Θ_{EM} represents the template or expected value of the attributes of each relationship between two objects in the represented scene.*
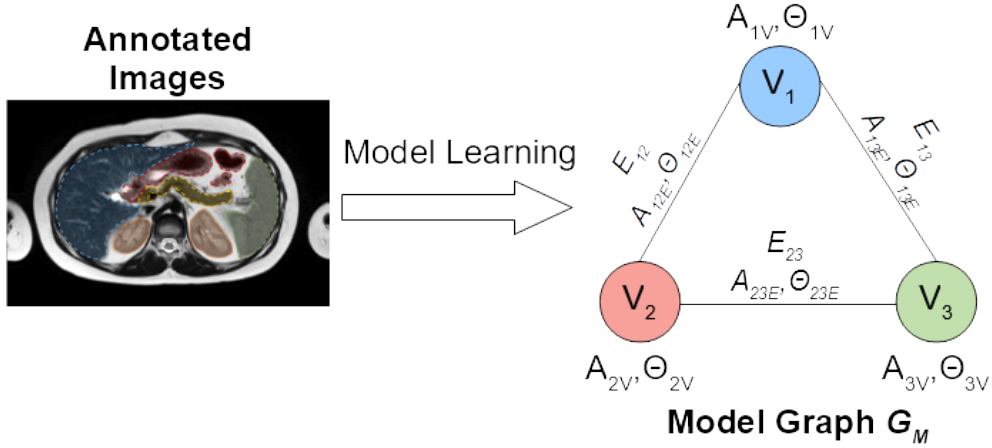

Figure 3.1: *Schematic representation of the learning and construction of the model graph G_M from a set of annotated images.*

3.1.2   Semantic Segmentation of Observed Data

Observation data is acquired in the form of an image to be segmented. It is in our interest to utilize the SRG in order to achieve proper and accurate semantic segmentation of the structures of interest in this data. We do this by building an observation SRG G_{obs} based on the input image, and attempting to match this new graph with the model SRG G_M.

Super-segmentation

To generate an observation graph, we first apply an automatic super-segmentation method, such as the *watershed*, which will divide the image into several sub-regions. This method will often split the actual regions of interest into several sub-regions, but it should contain the actual boundaries between regions as well. By extracting the proper features, we can build a SRG from this super-segmentation, which we will call the super-observation graph or G_{super}. Thus, the problem of matching the observation to the model is transformed into the problem of mapping which vertices in G_{super} should be merged.

Definition 3.1.3. *Let us formally define the super-observation graph as a Statistical-Relational Graph $G_{super} = (V_S, E_S, A_{VS}, \Theta_{VS}, A_{ES}, \Theta_{ES})$ where:*

1. *each vertex in V_S represents a subregion of the scene, as determined by the super-segmentation algorithm;*

2. *each edge in E_S represents a relationship between two of the subregions in V_S;*

3. *each random attribute vector in A_{VS} and corresponding distribution parameters in Θ_{VS} represents the attributes of a given subregion V_iS;*

4. *each random attribute vector in A_{ES} and corresponding distribution parameters in Θ_{ES} represents the attributes of a given relationship between two subregions E_iS.*
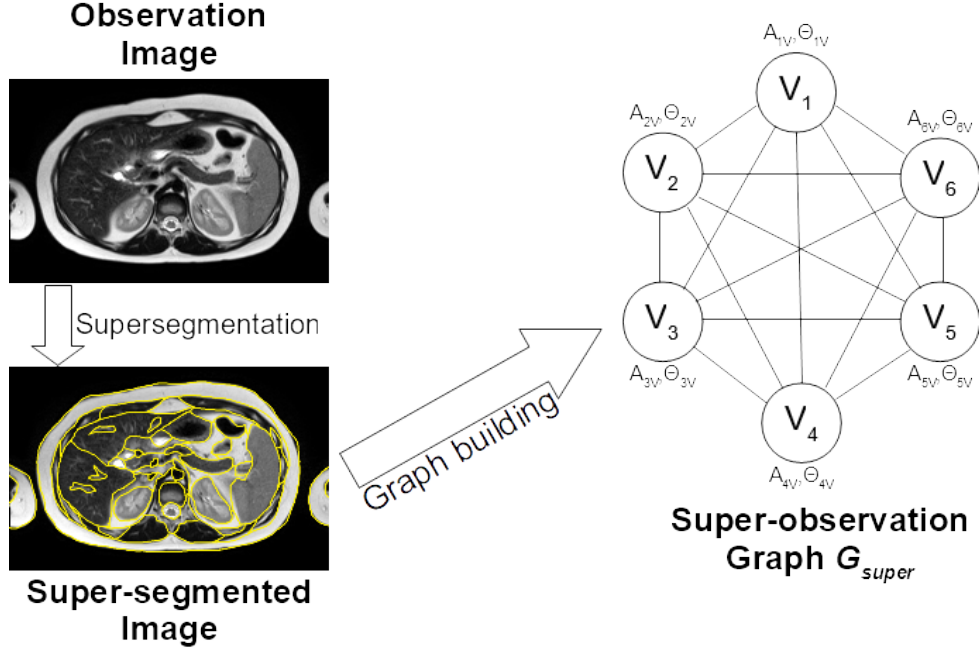

Figure 3.2: *Schematic representation of the super-segmentation of an observed image and construction of the super-observation graph G_{super}. Edge information (name E_{ij}, random vector A_{ijE} and distribution parameters Θ_{ijE}) are omitted from the figure for clarity.*

Observation solution

In order to compare the observed image with the model graph G_M, we must build an observation graph G_{obs} with the same number of vertices as the model. This G_{obs} is representative of our segmentation and its quality will be measured by direct comparison with G_M, through the use of some sort of cost function. The observation graph is built by matching each vertex of G_{super} with a vertex in G_M – producing a matching solution S – and joining all vertices matched to the same label. The matching solution should, optimally, result in an observation graph similar to the model graph. This process is schematized in Figure 3.3.

Definition 3.1.4. *Given a super-observation graph G_{super} with $|V_{super}| = n$, we can define a solution S which maps every vertex of the G_{super} to an observation graph G_{obs} with $|V_O| = |V_M| = m$ as:*

$$S = [s_1, s_2, ..., s_n], s_i \in \{1, m\} \tag{3.1}$$

Each vertex v_{iO} in the observation graph represents the union of all subregions s_j of G_{super} where $s_j = i$.

Definition 3.1.5. *Given a super-observation graph G_{super} and a solution S, let us formally define the observation graph as a Statistical-Relational Graph $G_{obs} = (V_O, E_O, A_{VO}, \Theta_{VO}, A_{EO}, \Theta_{EO})$ where:*

1. *each vertex in V_O represents an approximation of an object of interest in the represented scene;*

2. *each edge in E_O represents a relationship between the approximations of two objects of interest in the represented scene;*

3. *each random attribute vector in A_{VO} and corresponding distribution parameters in Θ_{VO} represents the estimate of the attributes of each approximate object in the scene;*

4. *each random attribute vector in A_{EO} and corresponding distribution parameters in Θ_{EO} represents the estimate of the attributes of each relationship between the approximations of two objects of interest in the scene.*
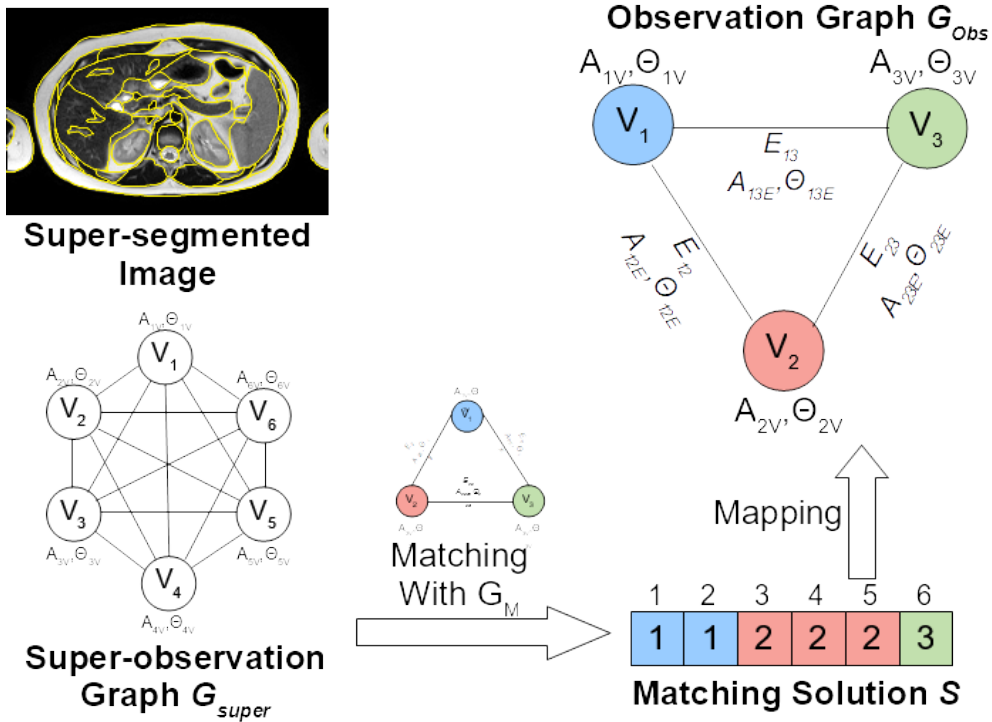
Figure 3.3: *Schematic representation of the construction of a solution S which matches G_{super} to the model graph G_M, and from S and G_{super} builds the observation graph G_{obs}.*

Matching Optimization Algorithm

S must be initialized with a given set of values, or predictions. These may be further modified by improvement algorithms, which will attempt to replace predictions in S in order to find a better solution (i.e., one with a lower matching cost). As mentioned above, optimally, we wish to find the solution S which generates an observation graph G_{obs} as close to the model graph G_M as possible, that is, whose random attribute vector distributions are as close to the model's distributions as possible.

Thus, we may phrase the problem of searching for the optimal S as finding the G_{obs} with the highest probability of being equal to G_M, or of maximizing $P(G_{obs} = G_M)$. However, measuring this probability directly requires knowledge of Θ_V and Θ_E, the parameters that control the distribution of the attributes. The stochastic learning of these parameters is currently the subject of other works related to this project. As such, lacking this knowledge, the value of $P(G_{obs} = G_M)$ must be estimated through other means, such as the computation of a "matching cost" [CJBBL05].

Given the properties of the SRG presented so far, we can infer that the quality of the process of semantic segmentation of an image using the SRG is dependent on two major steps: the quality of the super-segmentation algorithm, as the final regional borders are directly dependent on the super-segmentation subregions borders; and on the solution S, which fundamentally joins the subregions into the expected regions, producing an observation graph G_{obs} which should both describe the observed scene correctly, and be as close to the model graph G_M as possible. Thus, the step of determining and improving the matching of the distinct graphs G_{super} and G_M – represented by the solution S, and whose quality is measured by the matching cost $C(S)$ – is of utmost importance.

Several graph matching algorithms have been proposed in the literature, aiming to find a suitable matching solution between distinct graphs based on some sort of objective function (such as the matching cost). [CJBBL05], in particular, explores and compares several sets of algorithms, such as tree search algorithms (where each node of a search tree represents a possible match between a super-observation vertex and a model vertex); genetic algorithms, where individuals represent potential solutions and are evolved against the objective function; and estimation of distribution algorithms, stochastic search strategies not unlike genetic algorithms, which perform evolution through estimation of the probability distribution of the fittest individuals.

However, while of great interest to the overall project, such techniques and algorithms are beyond the scope of the current work, which aimed to implement a minimally-viable version of the SRG semantic segmentation process. As such, a brute-force greedy substitution algorithm was proposed, designed and implemented for the application of semantic segmentation in MRI and calibration approaches. This algorithm and its properties are described and discussed in Section 3.3.6.

Matching Cost

To evaluate the quality of a solution S, a cost function $C(S)$ must be established. This cost function must be able to properly measure the similarity between G_{obs} and G_M: it should be zero when $G_{obs} = G_M$, and it should be lower the more similar G_{obs} is to G_M. The cost function is utilized as a proxy to reflect $P(G_{obs} = G_M)$; while this may appear incongruent at first, higher costs represent lower probabilities of similarity and vice-versa (with $C(S) = 0$ representing $P(G_M = G_{obs}) = 1$) and, as such, the estimation of the probability is given by $P(G_M = G_{obs}) \sim \beta C(S)^{-1}$, where β is a scale factor. This is schematized in Figure 3.4.

The adopted cost of a solution S is defined as

$$C(S) = \alpha \frac{1}{n} \sum_{j=1}^{n} c_V(s_j) + (1 - \alpha) \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} c_E(s_j, s_k) \tag{3.2}$$

where $\alpha \in [0, 1]$ is the weight of the vertex cost; the costs between vertices s_j in G_{obs} and s_j in G_M, $c_V(s_j)$, is defined as

$$c_V(s_j) = \sum_{a \in A_V} \alpha_a d_a(s_{j_a}) \tag{3.3}$$

and the cost between the edges $\{s_j, s_k\}$ in G_{obs} and $\{s_j, s_k\}$ in G_M, $c_E(s_j, s_k)$, is defined as

$$c_E(s_j, s_k) = \sum_{a \in A_E} \alpha_a d_a(s_{j_a}, s_{ka}) \tag{3.4}$$

with d_a being the distance function corresponding to each attribute a and α_a being a customizable parameter, used for tuning the weight of the distance between each attribute a in the final cost. Note that $\sum_{a \in A_E} \alpha_a = \sum_{a \in A_V} \alpha_a = 1$. Experiments in Chapter 4 utilize the Euclidean distance as d for all attributes, unless explicitly stated otherwise.

It is important to mention that this cost requires the application of some normalization technique, in order to avoid attributes with a higher dynamic range dominating those with smaller values and variability. This is further detailed in Section 3.3.3.
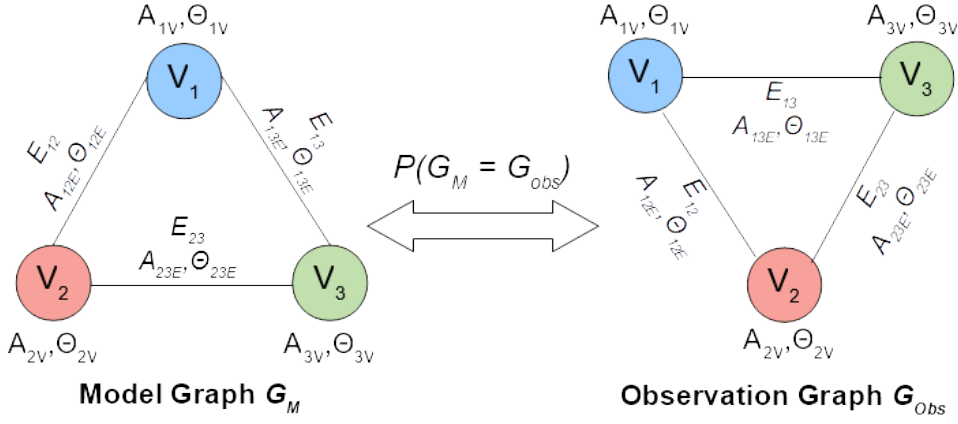

Figure 3.4: *Schematic representation of the assessment of matching solution S, through the estimation of the probability that G_{obs} is equivalent to G_M.*

3.2   Calibration of the SRG

In Section 3.1, we describe the Statistical-Relation Graph and its many components. Given its inherent complexity, a set of parameters must be set for its proper utilization. Those are:

- which attributes should be extracted for A_V and A_E;

- which super-segmentation algorithm to utilize, and its own parameters;

- which distance function d_a to utilize for each attribute $a \in A_V \frown A_E$;

- which algorithms to use to both initialize and improve S.

Given the many degrees of freedom available and the lack of literature in using the SRG for the task of medical image segmentation, a method for calibrating these parameters or, at least, assessing the quality of a set of parameters, was deemed necessary.

The utilization of synthetic data, with known truth values and controlled noise, allows for the determination of the quality of a set of parameters, measured as the difference between the prediction and the ground truth (both qualitatively, with visualizations of the results and the truth, and quantitatively using measures such as the similarity index, detailed in Section 3.2.1). This may be further developed to a method akin to a musical tuner, which both assesses quality and provides guidance on how to improve the parameter choice. An example of synthetic data, both clean and noisy, simulating an organ and a vein inside a body, is presented in Figure 3.5.

Additionally, the use of controlled data with known expected results is a great aid in the testing and debugging of the implementation of a SRG-based segmentation system.

3.2.1   The Calibration Pipeline

In order to obtain useful results from the calibration, a pipeline must be observed. First, a proper synthetic phantom must be generated, capable of simulating the real-world scenario but
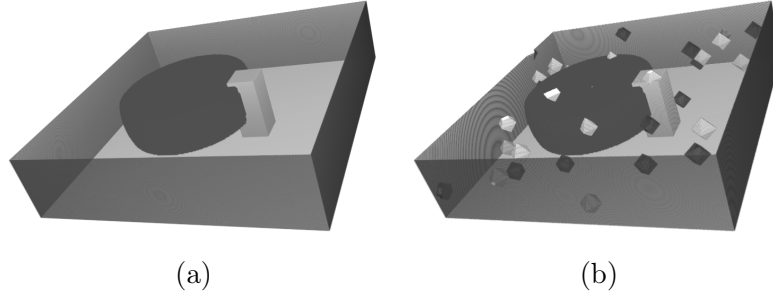
Figure 3.5: *Synthetic liver phantoms, used in the calibration of the SRG, simulating a darkened organ and a bright vein. The background has been omitted for visualization purposes. (a) clean; (b) with noise.*

simple enough for easy analysis. The phantom's true labels must also be generated. Second, a noise profile must be determined, and the generated noise may be added to the phantom. This aims to simulate the difficulties in the acquisition of real images, often fraught with noise and artifacts. Third, a model SRG must be learned, using a model learning technique. Fourth, we apply the process of segmenting and matching the observation graph. Finally, we assess the quality of the results found. These steps are detailed in the following sections.

Phantom Generation

The generation of a phantom consists of building a three-dimensional matrix of values, or simply "image", simulating but simplifying the conditions found in real-world MRI data. In addition, a set of annotations must also be generated, correctly labeling each voxel as belonging to the represented class.

In keeping with the project's applications, a phantom representing the liver and the Vena Cava inside a "body", itself inside a black background, was generated to simulate a T2-weighted MRI. All elements are homogeneous in intensity (that is, "color"). This phantom is depicted in Figure 3.5 (a). Its details may be found in Table 3.1.

Table 3.1: *Details of the liver phantom, generated for the calibration. See Figure 3.5 (a).*

Structure	Shape	Dimensions	Center	Intensity
Background	Rectangular prism	$300 \times 300 \times 50$	$(150, 150, 25)$	0.0
Body	Rectangular prism	$200 \times 200 \times 50$	$(150, 150, 25)$	0.5
Liver	Truncated sphere	$101 \times 101 \times 30$	$(150, 130, 28)$	0.3
Vena Cava	Rectangular prism	$20 \times 20 \times 45$	$(150, 180, 22.5)$	0.8

When utilizing positional attributes, such as the centroid of a region, it is important to ensure that said attributes are discriminant for each class. In order to do so, enveloping structures such as the background or the body (which envelop the body and the organs, respectively) are often split into several classes. For calibration as related to MRI segmentation, we split the background into two distinct regions separated by the sagittal (that is, X, Z) plane, resulting in an anterior background and a posterior background; we also split the body into eight distinct regions utilizing all planes centered on the body's centroid. This process of class splitting can be seen in Figure 3.6.
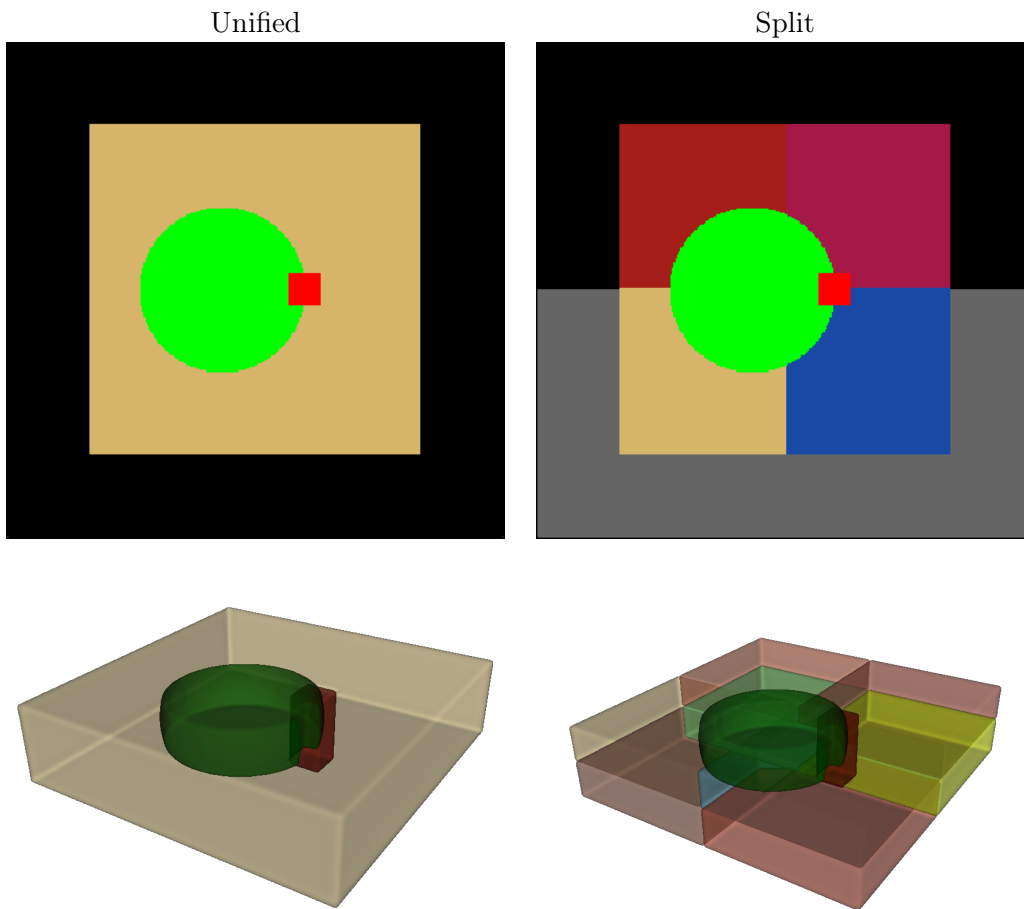
Figure 3.6: *Representation of the body and background splitting process. The upper images are the central axial slices of the phantom. The leftmost images show the background (in black) and the body (in beige); the rightmost images show the background split in two (in black and grey) and the body split in eight. The background is omitted from the 3D visualizations.*
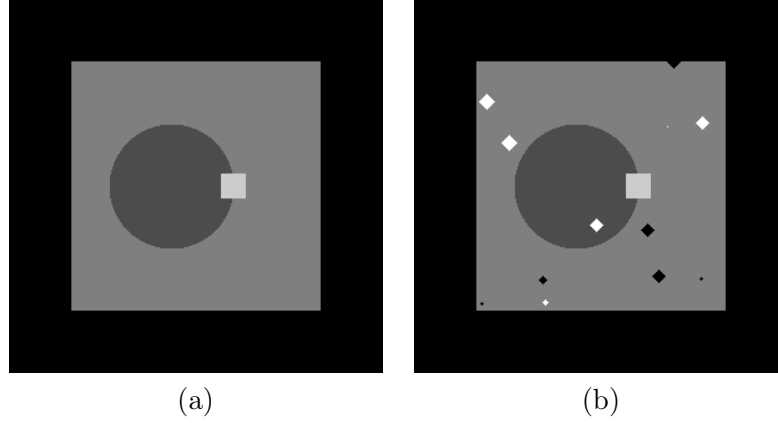
(a)   (b)

Figure 3.7: *Application of the FSP noise profile in a image. (a) clean slice; (b) noisy slice.*

Noise generation

In order to approach the diverse conditions found in real MRI data, the calibration includes the generation and application of controlled noise to the phantom, which enables stress-testing the SRG segmentation process. Noisy phantoms may be used both as model images – to calibrate model learning techniques and algorithms – and as observation images – to calibrate the observation, segmentation and recognition process.

With the intention of simulating one of the most daunting aspects of abdominal MRI segmentation, the presence of a great number of extraneous elements inside both the background (that is, the many organs that are not of interest comprising the rest of the body) and the foreground (that is, the diverse structures, such as veins, inside the region of interest), the "Fat Salt & Pepper" (FSP) noise profile was designed.

To generate FSP, we first define the "amount" of voxels to become new, noisy elements, which we call p; and the "radius" of all noisy elements, which we call r. We then convert p randomly chosen voxels to intensity 1.0 (the white "salt"), and other p randomly chosen voxels to intensity 0.0 (the black "pepper"). Finally, we apply morphological dilation with a 4-neighborhood to both salt and pepper voxels r times (the "fattening"). The end result is a series of black and white "bubbles" in the whole image, as can be seen in Figure 3.5 (b). Most of our applications additionally opt to remove any bubbles left in the black background, as large artifacts in real MRI backgrounds are quite rare. The results of this process on a 2D image can be seen in Figure 3.7.

Additional noise profiles, to be used separately or in conjunction, have been imagined, but left as future work. These include Gaussian noise profiles, to simulate acquisition noise in MR machines with smaller magnetic fields; and noise profiles which aim to simulate the Gibbs ringing artifact [BZ15], utilizing operations in the Fourier space.

Model Learning

The learning of a proper model graph, capable of accurately describing the objects of interest in a wide array of distinct images, is an essential step for the application of the SRG in the segmentation and recognition of real images. Thus, we must be able to assess the quality of any model learning techniques developed. However, as stated in Section 3.1.1, model learning is beyond the scope of this dissertation and will be addressed by future works in the project. As such, currently the only method utilized in calibration is the extraction the model from the clean phantom, utilizing its true labeling as basis.

However, the definition of the model also includes the definition of the attributes to be considered, both A_V and A_E. Different sets of attributes have different capabilities for description of the scene, its objects and their relationships. As such, the set of attributes to be used is an important parameter, whose quality the calibration process aids in assessing. Once defined, all desired attributes are computed and the model graph is generated.

Observation and Segmentation

Observation, segmentation and recognition of images are the steps containing the greatest number of parameters and, as such, tend to be the main focus of calibration. Observation starts with the observation image – typically, the phantom generated in the first step, with the addition of noise from the profile generated in the second step.

Then, the chosen super-segmentation technique is applied. This is a significant calibration parameter, as improper super-segmentation may complicate the graph matching, or even render it unfeasible. Through the use of different amounts of noise, and by utilizing the ground truth to guide the matching step, we may specifically assess the quality of the super-segmentation technique, by measuring how close the borders found match the ground truth borders.

In sequence, we must define two parameters simultaneously: the attribute distance function d_a for each attribute a, and the algorithms for initialization and improvement of the matching solution S. These are the core of the segmentation step, as it is expected that after initialization and improvement, S will contain an accurate segmentation of the observed image, based on the previously established model.

Quality Assessment

The final step, and arguably the main goal of the calibration, is the assessment of the quality of the segmentation result. Given the output of the entire pipeline, we are interested in examining its accuracy and applicability both qualitatively (by providing visualization capabilities) and, most importantly, quantitatively. In order to do so, we utilize the similarity index:

$$S_l = \frac{2|T_l \cap P_l|}{|T_l| + |P_l|}, l \in \{1, ..., |V_{G_M}|\} \tag{3.5}$$

where T_l is the true binary mapping of the single label l, P_l is the predicted binary mapping of the same label, and $|\cdot|$ is the cardinal. While not a true distance metric (as it does not satisfy the triangle inequality), it is nonetheless commonly used to measure the quality of a given label's prediction in the context of image segmentation [CCB06].

3.3   Semantic Segmentation of MRI using the SRG

As shown by [Gra12, GCB14], the SRG is capable of properly describing distinct structures of interest in abdominal MRI, along with their spatial relationships. Thus, by utilizing the process described in Section 3.1 with an adequate set of parameters and a proper model, unknown MRI images may be automatically segmented.

3.3.1   MRI Data

The main source of data for this project is the Instituto da Criança (ICr-FM), with support from Professors Lisa Suzuki and Marcelo Takahashi. Abdominal MRI of twelve patients, ranging from four to eighteen years old, were acquired for use in this project. All acquisitions were performed using ICr's Philips Ingenia (1.5T) machine. Three sequences were acquired:

- A **Dixon** sequence [Dix84], which is a T1-weighted sequence that produces four complementary mappings; this sequence has a resolution of 1.21 x 1.21mm per slice, 2.5mm inter-slices;

- A **T2** sequence; this sequence has a resolution of 0.72 x 0.72mm per slice, 5mm inter-slices;

- A **multi-echo** (ME) sequence, which contains twelve captures of a few (four to five) slices; this sequence will be used for determining iron content in the liver.

The acquisitions from patient #4, a healthy, 11-years old female patient, were chosen for use in this dissertation. This decision stems from the clearer images, with visually lower amount of noise and artifacts. We focus on the T2 sequence (with 42 slices), with a dynamic range of 1500 values, as it is one of the most commonly used in the literature. The sequence was annotated and verified by medical experts.

3.3.2   Conceptual Modeling

In order to attempt segmentation and recognition of the structures of interest in MRI, the following attributes of each structure were considered:

- Intensity;

- Volume;

- Absolute position in the body.

Additionally, the relationships between a structure and all others must also be taken into consideration, with the following attributes being most relevant:

- Relative distance and position;

- Differences in volume;

- Differences in intensity.

We believe the combination of these statistical and relational attributes represent an accurate approximation of the discriminant factors required for proper segmentation and recognition of the relevant structures. Previous works have shown successful results produced by such attributes [CCB06, GCB14, MBCJ16].

3.3.3   Computational Modeling

Considering the conceptual model proposed in Section 3.3.2, we then define the set of attributes, both statistical (that is, represented in the vertices) and relational (that is, represented in the edges), to be extracted.

Three vertex attributes are computed for each vertex. Considering the structure they represent, these attributes are **i)** the average of the intensity values in the structure; **ii)** the total volume of the structure; and **iii)** the position of the centroid of the structure, given as three distinct attributes, x, y and z.

These attributes may be computed on the data space, that is, voxel-wise, with centroids given as the voxel's indexes in the 3D volume and volume as the count of voxels in the region of interest; or they may be computed in real space, with centroids given as real-valued positions, utilizing the MRI's metadata to determine the real coordinates of a given voxel, and volume computed using the actual volume of each voxel in the region of interest. We have chosen to compute attributes using the real-valued positions and volume. It is important to note that, as patients may position themselves in different heights of the scanning table, a position normalization method may be required for model learning and observation with multiple patients.

Three edge attributes are computed for each edge. Considering the structures represented by the connected edges, these attributes are **i)** the vector between the centroids of each structure,given as three distinct attributes, a, b and c where $\vec{v} = (a, b, c)$; **ii)** the absolute difference in volume between the structures; and **iii)** the contrast (that is, difference) between the average intensities of each structure.

Edge attributes are derived directly from the vertex attributes, but are explicitly defined and stored as distinct elements. The reason for this explicit definition is threefold; first, doing so allows us to optimize the computation of edge distances; second, these values are used frequently in graph matching and description and as such, storing them is more efficient than constantly recomputing them; and finally, we may design other edge attributes which are not derived from the vertex attributes.

After computing all attributes, a normalization operation was applied to each attribute in turn, transforming them into distributions with a mean of 0 and a standard deviation of 1. The values used in this transformation – that is, the mean and deviation of the model – are stored to be used in normalizing the observation graph, by applying the same transformation. Thus, observation attributes may not have a mean of 0 and a deviation of 1, but they are in the equivalent dynamic range of the model attributes.

Figure 3.8 shows an example of a simplified SRG, with three fully connected vertices, representing Segment I (SI), Segment V (SV) and the Vena Cava (VC); vertex attributes: centroid and mean intensity; and a single edge attribute: distance.


```
                    [Mean intensity: 108.3]
                    [Centroid: (246,215,13)]
                             SI
  {Distance:∓(22,30,3)}            {Distance:∓(24,97,7)}

         VC        {Distance:∓(2,77,4)}        SV
  [Mean intensity: 343.5]              [Mean intensity: 113.2]
  [Centroid: (224,185,10)]            [Centroid: (222,118,6)]
```
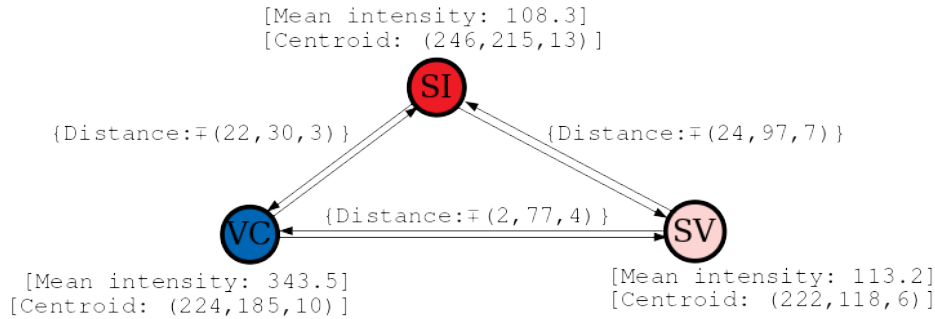
Figure 3.8: *Schema of a simple, example SRG for MRI description. Vertex attributes are between square brackets. Edge attributes are between curly braces.*

Experiments in Chapter 4 utilize this set of attributes, unless explicitly stated otherwise.

As model learning from multiple sources was not included in the scope of this dissertation, we
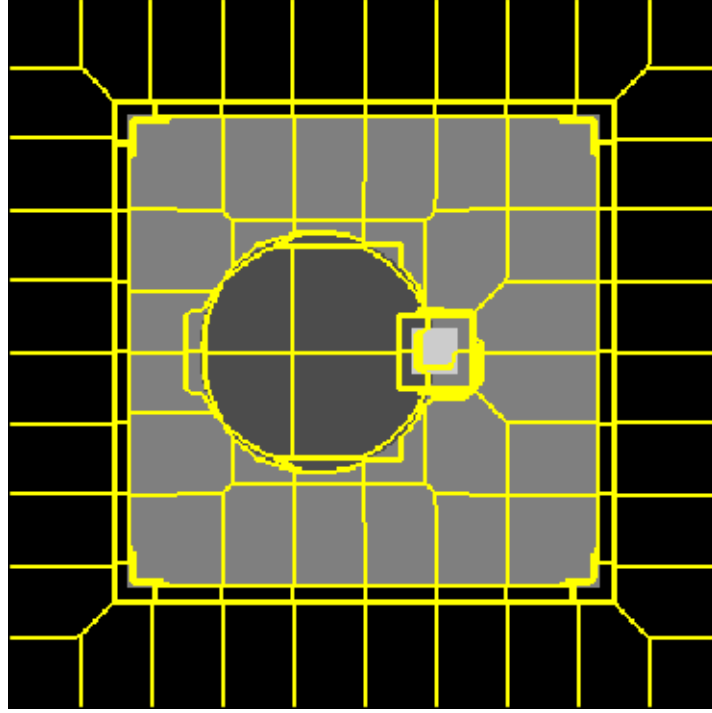
Figure 3.9: *Example of the regions produced by watershed techniques on the body phantom. The yellow lines denote the borders between super-segmented regions.*

opted to build our model from a single annotated MRI volume. The same image would be used as observation input; this allows for the investigation of the observation graph building and matching process without concerning ourselves with general model learning.

3.3.4   Super-segmentation

To build the observation graph G_{obs}, we must first build the super-observation graph G_{super} from the output of a super-segmentation algorithm applied over the observed image. This algorithm should ideally be able to discern or approximate the true borders of the elements of interest; the SRG matching should then eliminate the superfluous borders, obtaining the true segmentation or an approximation thereof.

The classic algorithm for automatic super-segmentation is the *watershed*, applied over the gradient of the image. The watershed needs a set of markers, which represent initial basins and may be thought of as region seeds. These markers may be distributed equally spaced, or automatically determined (e.g., as the local minima of the gradient). An example of watershed segmentation results is presented in Figure 3.9.

Other algorithms, such as the Simple Linear Iterative Clustering (SLIC) [ASS+12] or the compact watershed [NP14], may be used to achieve super-segmentation. In Section 4.1.2, we conduct experiments to explore and compare these super-segmentation algorithms. We find that the compact watershed more closely approaches the borders of the liver, while the SLIC performs better when taking the average of the quality of the borders for all classes.

3.3.5   Initial Matching

As the matching algorithm is dependent on a fully-defined cost function, we first define the distance function d_a as the Euclidean distance for all attributes $a \in A_V \frown A_E$. Experiments in Chapter 4 utilize this distance function, unless explicitly stated otherwise.

We subsequently define an algorithm for initializing S based on the nearest model vertex for each vertex in G_{super}. This initialization cannot take into consideration the "size" attribute, however, as the super-segmented regions are of smaller size than the model regions; thus we use the vertex cost function c_V^*, which is the cost function c_V but without taking into consideration the size attribute. This algorithm is displayed in Algorithm 1.

Algorithm 1 Solution Initializer

1: **procedure** INITIALIZER(V_{super}, V_M)   ▷ Closest vertex solution
2:      S is an empty vector of size $|V_{super}|$
3:      **for all** vertex v **in** V_{super} **do**   ▷ c_V^* does not include size attribute
4:          $C \leftarrow [c_V^*(v, v_{M0}), c_V^*(v, v_{M1}), ...]$ where v_{Mi} is the i-eth element of V_M
5:          $S[i] \leftarrow argmin(C)$
6:      **end for**
7:      **return** S
8: **end procedure**

After the initialization, it is possible that some noncontiguous vertices are assigned to the same model vertex. As the structures of interest which we deal with are all contiguous, we apply Algorithm 2 to guarantee the contiguity of each region in G_{obs}. To achieve this, we need the adjacency graph of all regions in the super-segmentation, adj.

Algorithm 2 Solution Contiguity Guarantee

1: **procedure** CONTIGUITY(V_{super}, V_M, S, adj)   ▷ Contiguity guarantee
2:      **for** $i \leftarrow 1, |V_M|$ **do**
3:          $T^* \leftarrow arg_x(S[x] = i)$ ▷ T^* is the list of regions assigned to i
4:          $T \leftarrow connected\_components(T^*, adj)$ ▷ T joins all adjacent regions in T^*
5:          $best \leftarrow argmin_j(c_V(T_j, V_{Mi}))$ ▷ Find best connected region as "true"
6:          **for all** j **in** T **and** $j \neq best$ **do**
7:              $S[j] \leftarrow UPDATE$ ▷ Other regions are flagged for updating
8:          **end for**
9:      **end for**
10:     **for all** j **in** S **where** $S[j] = UPDATE$ **do**
11:         $S \leftarrow improvement(j, G_{super}, G_M, S, adj)$ ▷ IMPROVEMENT is defined in Algorithm 3
12:     **end for**
13:     **return** S
14: **end procedure**

3.3.6   Solution Improvement

The proposed algorithm for initialization of S does not take into consideration the relational aspect of the SRG, being based exclusively on vertex similarity. However, as stated, there are significant gains to be achieved by taking structural information into consideration. As such, we designed an algorithm for improving upon S through exhaustive consideration of all attributes, both statistical and relational. This algorithm is described in Algorithm 3. This algorithm depends

on the adjacency graph of all regions in the super-segmentation, *adj*, as it maintains the contiguity of each region in G_{obs}.

Of particular note is the consideration that, as this is a greedy algorithm, the choice of which vertex(ices) to improve – that is, the choice of the set of *j*s to be passed as an argument to the function – is a significant one. In this work, we have found it better and intuitive to improve the vertices in order of initial matching cost, that is, $min(C)$ of C in line 4 of Algorithm 1, going from costliest to cheapest. Depending on the experiment, a cutoff may be introduced, such as a cost threshold or an absolute number of vertices to improve.

Experiments in Chapter 4 utilize these graph matching algorithms, unless explicitly stated otherwise.

Algorithm 3 Solution Improvement

1: **procedure** IMPROVEMENT(j, G_{super}, G_M, S, adj)   ▷ Improving vertex j from G_{super}
2:     $candidates \leftarrow adj[j] \frown j$ ▷ Contiguity guarantee
3:     $costs$ is an empty vector of size $|candidates|$
4:     **for all** c **in** $candidates$ **do**
5:         $S^* \leftarrow copy(S)$
6:         $S^*[j] \leftarrow c$
7:         $G_{obs} \leftarrow build\_srg(G_{super}, S)$
8:         $costs$(index of c) $\leftarrow C(G_{obs}, G_M)$ ▷ Compute cost of each candidate
9:     **end for**
10:    $best \leftarrow candidates[argmin(costs)]$ ▷ Get cheapest candidate
11:    $S[j] \leftarrow best$
12:    **return** S
13: **end procedure**

Chapter 4

Experimental Results and Discussions

A set of experiments were designed with the dual goals of determining whether the calibration method is capable of properly assessing the quality of a SRG segmentation method, helping to set its parameters, and of attempting segmentation of real data with the SRG. We demonstrate the correlation between matching cost and accuracy, reinforcing the results obtained by [Gra12] in the context of the calibration pipeline; we then leverage said pipeline to explore and assess the relative quality of several super-segmentation methods and attribute weight sets. Finally, we perform a preliminary experiment of real MRI data segmentation and recognition, in preparation for future work.

All of the code utilized in both the implementation of our version of the SRG, as well as the executable scripts of the experiments here described and their results, may be found in the project's Git repository[1].

4.1 Experimental Setup

4.1.1 Correlation of Cost and Accuracy

In order to utilize the SRG as a segmentation and recognition technique, we must demonstrate both its robustness and the correlation between the cost of matching two SRGs and the similarity of the represented objects. To do so, we must first show that the SRGs of two identical representations are identical themselves, and that the cost of matching two identical SRGs is zero. This is proven mathematically.

We then demonstrate that deviations from an accurate prediction result both in decreased accuracy and increased costs. To do so, we proceeded as follows:

1. Acquire the calibration phantom as our model and observation image;

2. Build the model graph G_M from the phantom's ground truth (we split the body and background labels);

3. Build the super-observation graph G_{super} utilizing watershed with $x \times y \times z$ initial markers (distributed evenly along each axis), plus 5 markers to ensure the vena cava's segmentation;

---

[1] https://github.com/MarEe0/STAP-liver

4. Build the matching solution S as follows: each G_{super} vertex is matched to the G_M vertex whose label is predominant in the super-segmentation region it represents. This results in the optimum S;

5. Compute cost and accuracy for optimum S;

6. Repeat k times:

   (a) Randomly replace p predictions in S with a different label;

   (b) Compute cost and accuracy for current S.

We measure the accuracy using two distinct values: the similarity index of the predicted liver region and the ground truth liver region, and as the average of all similarity indexes for all labels in the phantom. We compute cost utilizing even-weighted Euclidean distance, as described in Section 3.3.5. We explore values of $(x, y, z) = \{(6, 6, 4), (7, 7, 5), (8, 8, 6)\}$. For each set of initial markers, $p = \lfloor \frac{x \times y \times z}{20} \rfloor$ and $k = \lfloor \frac{x \times y \times z}{10} \rfloor$. Each experiment is repeated 100 times to smooth out the random factor in the disturbance of the solution.

4.1.2   Exploration of Super-Segmentation Techniques

As proper super-segmentation (that is, which approaches all true borders) is essential to SRG segmentation, it is in our best interests to ascertain which method has the greatest performance. Thus, we devise an experiment utilizing part of the calibration pipeline in order to explore the relative gains between different super-segmentation techniques. To ascertain each technique's robustness to noise, different noise profiles were investigated. For a given super-segmentation algorithm A, we:

1. Acquire the calibration phantom as our model;

2. Build the model graph G_M from the phantom's ground truth (we **do not** split the body and background labels);

3. Generate a set of m noise profiles $F = \{f_0, .., f_m\}$;

4. For each $f_i \in F$:

   (a) Apply f_i to the phantom to generate the observation;

   (b) Build the super-observation graph G_{super} utilizing A;

   (c) Build the matching solution S as follows: each G_{super} vertex is matched to the G_M vertex whose label is predominant in the super-segmentation region it represents. This results in the optimum S;

   (d) Compute accuracy for optimum S.

We measure the accuracy using two distinct values: the similarity index of the predicted liver region and the ground truth liver region, and as the average of all similarity indexes for all labels in the phantom.

We utilize the set of noise profiles $F = \{f_0, f_1, f_2, f_4\}$, where f_i is a FSP profile (see Section 3.2.1) with $r = 7$ and $p = |D|i10^{-5}$, $|D|$ being the number of voxels in the image. For the calibration phantom, $|D| = 300 \times 300 \times 50 = 4.500.000$ voxels, and thus the set of p for each f in F is $P = \{0, 45, 90, 180\}$.

We explore the following super-segmentation algorithms:

1. Traditional watershed, three-dimensional, with both gradient local minima and $4{\times}4{\times}4$ points distributed evenly across each axis as initial markers;

2. Compact watershed [NP14], three-dimensional, with $8{\times}8{\times}6$ initial markers distributed evenly across each axis;

3. Compact watershed, three-dimensional, with $10 \times 10 \times 8$ initial markers distributed evenly across each axis;

4. Single-channel SLIC Superpixel [ASS$^+$12], three-dimensional, with anisotropic voxel spacing $(6 \times 6 \times 1)$ and 400 regions;

5. Single-channel SLIC Superpixel, three-dimensional, with anisotropic voxel spacing $(6 \times 6 \times 1)$ and 600 regions.

All algorithms are prefaced by preprocessing with an anisotropic Gaussian filter, with $\sigma = (5, 5, 1)$. Each experiment is repeated 100 times to smooth out the random factor in the generation of noise.

4.1.3   Weight Calibration

As stated in Section 3.1.2, when defining the cost function, several weights must be set:

- the relative weighting between vertex cost and edge cost, in order to increase the impact of the statistical or the relational attributes;

- the relative weighting between each attribute, in order to increase the impact of some attributes over the others.

In order to improve the quality of the segmentation, we must determine a set of optimal weights. The calibration pipeline allows us to assess the quality of a given set of weights. To ascertain each set's robustness to noise, different noise profiles are investigated. For a given set of weights W and super-segmentation algorithm A, we:

1. Acquire the calibration phantom as our model;

2. Build the model graph G_M from the phantom's ground truth (we split the body and background labels);

3. Generate a set of m noise profiles $F = \{f_0, .., f_m\}$;

4. For each $f_i \in F$:

   (a) Apply f_i to the phantom to generate the observation;
   (b) Build the super-observation graph G_{super} utilizing A;
   (c) Build the matching solution S utilizing the initialization and improvement algorithms described in Section 3.3.5;
   (d) Compute cost and accuracy for S.

We measure the accuracy using two distinct values: the similarity index of the predicted liver region and the ground truth liver region, and as the average of all similarity indexes for all labels in the phantom. We compute cost utilizing W-weighted Euclidean distance, as described in Section 3.3.5.

Table 4.1: *Weight sets for the weight calibration experiment. Each weight set emphasizes a specific attribute.*

$\mathbf{W_i}$	Emphasized attribute
W_0	All weights are equal
W_1	Region centroid x coordinate (vertex attribute)
W_2	Region centroid y coordinate (vertex attribute)
W_3	Region centroid z coordinate (vertex attribute)
W_4	Mean region intensity (vertex attribute)
W_5	Total region size (vertex attribute)
W_6	Relative position vector x coordinate (edge attribute)
W_7	Relative position vector y coordinate (edge attribute)
W_8	Relative position vector z coordinate (edge attribute)
W_9	Mean region intensity contrast (edge attribute)
W_{10}	Region size difference (edge attribute)

We utilize the same set of noise profiles $F = \{f_0, f_1, f_2, f_4\}$ described in Section 4.1.2.

We utilize the *Compact Watershed* $10 \times 10 \times 8$ super-segmentation algorithm as A, chosen because of its high performance in obtaining the true liver borders, as shown and discussed in Section 4.2.2. We utilize the improvement algorithms discussed in Section 3.3.6, attempting to improve the costliest half of the vertices in the G_{super}.

We explore eleven sets of weights, as follows: vertex and edge costs are even-weighted (that is, $\alpha = 0.5$ in Equation 3.2); the first set of weights W_0 is even-weighted for all attributes; all subsequent sets of weights W_i are even-weighted except for the i-th attribute, which has double the weight of the other attributes. Table 4.1 shows which attribute (as described in Section 3.3.3) is emphasized for each W_i.

Thus, the set W_1 has vertex weights $\alpha_V = (\mathbf{0.\overline{3}}, 0.1\overline{6}, 0.1\overline{6}, 0.1\overline{6}, 0.1\overline{6})$ and edge weights $\alpha_E = (0.2, 0.2, 0.2, 0.2, 0.2)$; the set W_2 has vertex weights $\alpha_V = (0.1\overline{6}, \mathbf{0.\overline{3}}, 0.1\overline{6}, 0.1\overline{6}, 0.1\overline{6})$ and edge weights $\alpha_E = (0.2, 0.2, 0.2, 0.2, 0.2)$; the set W_6 has vertex weights $\alpha_V = (0.2, 0.2, 0.2, 0.2, 0.2)$ and edge weights $\alpha_E = (\mathbf{0.\overline{3}}, 0.1\overline{6}, 0.1\overline{6}, 0.1\overline{6}, 0.1\overline{6})$; and so on. Each experiment was repeated 100 times to smooth out the random factor in the generation of noise.

4.1.4   Preliminary Application on Real Data

After building a segmentation pipeline and assessing its quality through calibration, the next natural step is to apply said pipeline to real MRI data. The purpose of this is twofold: first, to measure the progress towards the project's stated goal of accurate segmentation of MRIs; and second, to analyze potential deficiencies of the calibration method. This analysis allows us to improve the calibration method with better phantoms and noise profiles, which in turn allows us to improve real MRI segmentation in a positive feedback loop. Thus, given a super-segmentation algorithm A and a set of weights W, we:

1. Acquire annotated real MRI data as both model and observation image;

2. Build the model graph G_M from the image's ground truth (we split the body and background labels);

3. Build the super-observation graph G_{super} utilizing A;

4. Build the matching solution S utilizing the initialization and improvement algorithms described in Section 3.3.5;

5. Compute cost and accuracy for S.

We measure the accuracy using two distinct values: the similarity index of the predicted liver region and the ground truth liver region, and as the average of all similarity indexes for all labels in the phantom. We compute cost utilizing W-weighted Euclidean distance, as described in Section 3.3.5.

The weight set W chosen, based on the results of the previous experiment, as presented in Section 4.2.3, is $W = 4$, which emphasizes the vertex attribute *mean region intensity*. We utilize the *Compact Watershed* $10 \times 10 \times 8$ super-segmentation algorithm as A, chosen because of its high performance in obtaining the true liver borders, as shown and discussed in Section 4.2.2. Due to the nature of real MRI data, the anisotropic Gaussian smoothing was performed with $\sigma = (5, 5, 3)$. We utilize the improvement algorithms discussed in Section 3.3.6, attempting to improve the costliest half of the vertices in the G_{super}.

4.2 Results and Discussion

4.2.1 Correlation of Cost and Accuracy

Let us prove that two identical objects generate identical SRGs.

Lemma 4.2.1. *Given two identical objects, x and y, and a deterministic SRG generation function F, then the graphs generated by x and y, G_x and G_y respectively, are equal.*

Proof. Let us assume the function for the generation of an SRG from an image, called F, is deterministic and surjective. This is the function proposed in Section 3.1.1, and instantiated in Section 3.3.3. Thus, $G_x = F(x)$ and $G_y = F(y)$.

If F is deterministic and surjective, then $x = y \implies F(x) = F(y) \implies G_x = G_y$. □

Let us prove that two identical SRGs have a cost of zero.

Lemma 4.2.2. *Given two identical graphs, G_x and G_y, and a cost function C whose atomic distance functions are zero for identical objects, then $C(G_x, G_y) = 0$.*

Proof. Let us assume a cost function C as described in Section 3.1.2. Let us assume that every distance function $d_a \forall a$ equals zero when measuring the distance between two identical elements, e.g. $d_a(x, x) = 0 \forall a, x$. The Euclidean distance utilized in this implementation of the SRG obeys this assumption.

Equation 3.2 shows that the total matching cost of two SRGs is given by a weighted mean of both partial costs C_V and C_E, described in Equations 3.3 and 3.4 respectively. As both partial costs are sums of the distances, then if we are measuring the cost between two identical graphs G_x and G_y, $G_x = G_y \implies d_a = 0 \forall a \implies C_V = C_E = 0 \implies C(G_x, G_y) = 0$. □

Taken together, these two lemmas show that the matching cost for graphs derived from identical objects is zero, as we intended.

We report the results of the correlation experiment in Figure 4.1. Due to the random nature of the experiment, 100 repetitions of each experiment were executed, and we report both the mean of the results and the 95% confidence interval for 100 repetitions.

Discussion

Figure 4.1 shows that, as disturbances are introduced (represented by the x-axis), the solution naturally steps away from the optimum. This process is demonstrated by the lowering of the solution accuracy, as estimated by both the similarity index of the solution liver and the true liver (represented by the red lines in the plots) and the average similarity indexes between the prediction of all classes and their respective ground truths (represented by the green lines in the plots).

Meanwhile, as disturbances increase and the accuracy decreases, the cost of matching the observation graph G_{obs} to the model graph G_M (represented by the blue lines) steadily increases, inversely proportional to the estimated accuracy. It is evident upon visualization that the matching cost is inversely proportional to the accuracy of the prediction. As stated in Section 3.1.2, the matching cost should also inversely proportional to $P(G_M = G_{obs})$, which is directly related to the similarity of the predicted regions and the ground truth. Thus, the evidence backs up the assumption that the inverse of the cost is a good estimator of solution quality.

The shaded areas in the plots are the 95% confidence intervals for each value, given by the repetition of the experiments 100 times. As the tight intervals show, the nature of the disturbances and their progressions (as determined randomly each execution) does not significantly impact the results obtained; that is, despite the difference in the nature of the disturbances, accuracy steadily decreases and cost steadily increases, thus demonstrating the robustness of the hypothesis of the correlation between cost and accuracy.

Additionally, these results can be observed across all three sets of initial markers. Regardless of the nature of the super-segmentation (and of its myriad subregions), as disturbances increase and similarity decreases, the cost steadily increases, thus demonstrating that this correlation does not depend on the nature of the super-segmentation and of the disturbances, and instead is a property intrinsic to the relationship between similarity and matching cost. It is worth noting that the average SI is lower than the liver SI in the $7 \times 7 \times 5$ experiment; this is a result of improper super-segmentation of the other classes. However, the observations made above are still valid.

This correlation strengthens the hypothesis that a SRG with the attributes specified in Section 3.3.3 is capable of properly representing and describing the phantom's objects and their relationships. As the phantom is a simulation of real medical MRI data, this points to robustness of representing real MRI data with such a SRG. This corroborates results found in [Gra12] and is the fundamental assumption behind SRG segmentation and recognition of images.

One thing must be kept in mind about the SRG matching costs: the comparison of costs between different executions of SRG-based semantic segmentation is only valid when costs obtained from the same model SRG are compared. SRGs with fundamental differences (such as distinct attributes, distinct cost functions or even distinct attribute weights) may possess different matching costs for the same observation image. This derives from Lemma 4.2.1, as identical objects with distinct SRG configurations will not necessarily generate identical SRGs.

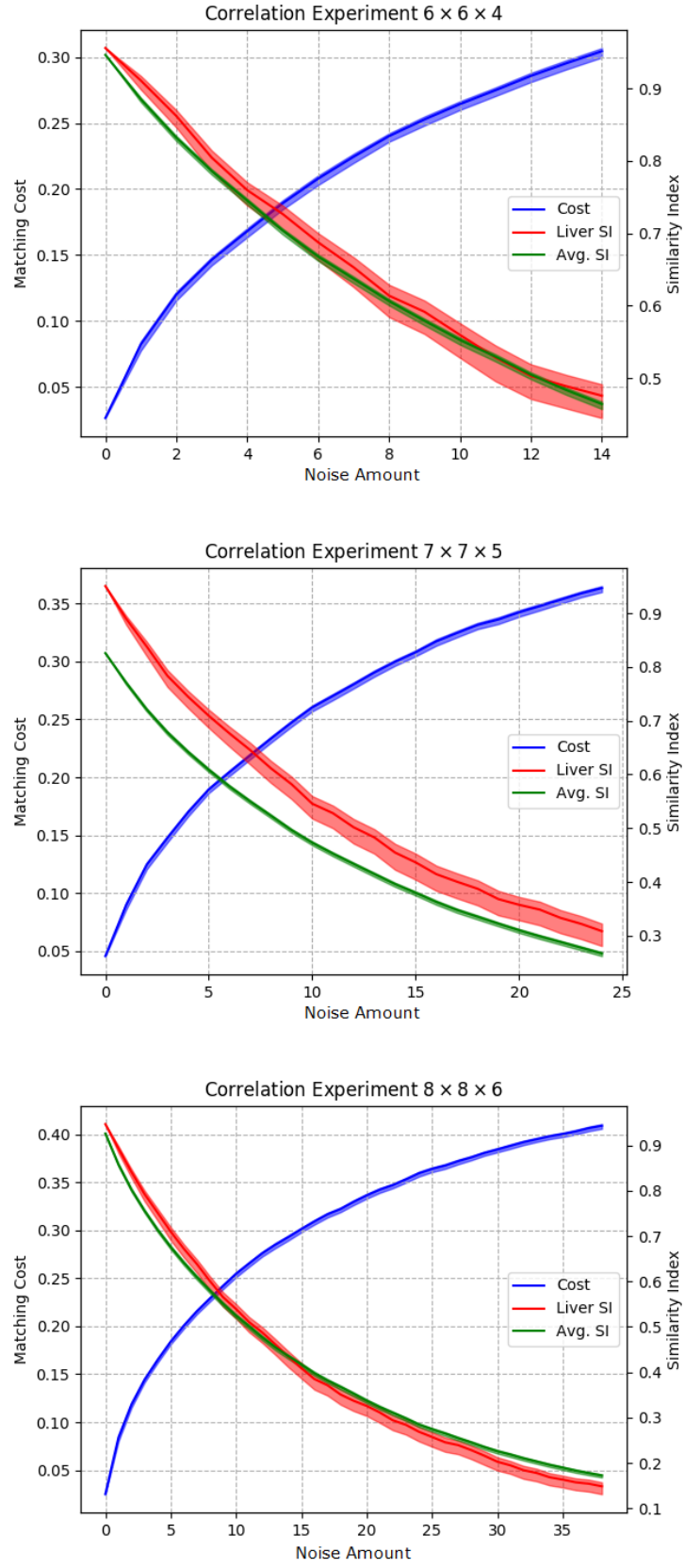**Figure 4.1:** *Results for the correlation of cost and accuracy experiments, for different distributions of initial markers. We present the results for $(6, 6, 4)$, for $(7, 7, 5)$, and for $(8, 8, 6)$ markers. Shaded areas represent the 95% confidence interval. We report the matching cost (Equation 3.2), the Similarity Index, or SI (Equation 3.5), of the liver and the average of the SIs of all labels.*

4.2.2 Exploration of Super-Segmentation Techniques

We report the results of the super-segmentation experiment on Table 4.2. The best value for each category is highlighted. Due to the random nature of the experiment, 100 repetitions of each experiment were executed, and we report both the mean of the results and the 95% confidence interval for 100 repetitions.

Table 4.2: *Results for the super-segmentation experiments. Average similarity index (SI, Equation 3.5) of all classes and liver similarity index were reported for each algorithm, on each noise profile. The 95% confidence interval is reported.*

Noise	Algorithm A	Average SI	Liver SI
f_0	Traditional Watershed	0.6523 ± 0	0.9410 ± 0
	Compact Watershed $8 \times 8 \times 6$	0.8345 ± 0	0.9354 ± 0
	Compact Watershed $10 \times 10 \times 8$	0.8408 ± 0	$\mathbf{0.9593 \pm 0}$
	Slic, 400 regions	0.8979 ± 0	0.8688 ± 0
	Slic, 600 regions	$\mathbf{0.9057 \pm 0}$	0.8782 ± 0
f_1	Traditional Watershed	0.6504 ± 0.0007	0.9395 ± 0.0008
	Compact Watershed $8 \times 8 \times 6$	0.8221 ± 0.0056	0.9309 ± 0.0010
	Compact Watershed $10 \times 10 \times 8$	0.8359 ± 0.0047	$\mathbf{0.9548 \pm 0.0009}$
	Slic, 400 regions	0.8910 ± 0.0020	0.8592 ± 0.0045
	Slic, 600 regions	$\mathbf{0.8991 \pm 0.0017}$	0.8720 ± 0.0048
f_2	Traditional Watershed	0.6498 ± 0.0008	0.9418 ± 0.0009
	Compact Watershed $8 \times 8 \times 6$	0.8124 ± 0.0077	0.9273 ± 0.0012
	Compact Watershed $10 \times 10 \times 8$	0.8272 ± 0.0068	$\mathbf{0.9504 \pm 0.0011}$
	Slic, 400 regions	0.8900 ± 0.0027	0.8603 ± 0.0052
	Slic, 600 regions	$\mathbf{0.8968 \pm 0.0018}$	0.8618 ± 0.0059
f_4	Traditional Watershed	0.6489 ± 0.001	$\mathbf{0.9453 \pm 0.0008}$
	Compact Watershed $8 \times 8 \times 6$	0.7716 ± 0.0133	0.9189 ± 0.0015
	Compact Watershed $10 \times 10 \times 8$	0.8106 ± 0.0097	0.9428 ± 0.0013
	Slic, 400 regions	0.8869 ± 0.0032	0.8570 ± 0.0047
	Slic, 600 regions	$\mathbf{0.8905 \pm 0.0022}$	0.8490 ± 0.0050

We display some of the super-segmentation results for the *Slic, 600 regions* and the *Compact Watershed* $10 \times 10 \times 8$ algorithms in Figures 4.2 and 4.3. For one instance of each noise profile, we display the central axial and coronal slices, with yellow lines indicating the border between adjacent regions.

Discussion

As Table 4.2 shows, the super-segmentation algorithm which consistently obtains the best approaches to the true liver borders is the *Compact Watershed* $10 \times 10 \times 8$, while the one with the best overall approximation of the all classes' borders is the *Slic, 600 regions*. As Figures 4.2 and 4.3 reveals, however, many regions in the SLIC segmentation are not properly split across the sagittal/coronal axis. Given the results displayed, the watershed algorithm is the best suited for the task at hand, even when given a relatively low number of starting regions.

Overall, the watershed algorithms had better accuracy for the liver borders, and the SLIC algorithms had better average accuracy overall. As can be seen from the images provided, SLIC segmentation tends to utilize a large number of thin regions in the borders; however, these regions are often heterogeneous and, as such, envelop the actual true border. Additionally, noisy regions near the border of the liver would result in significant disturbances to the real border, as can be
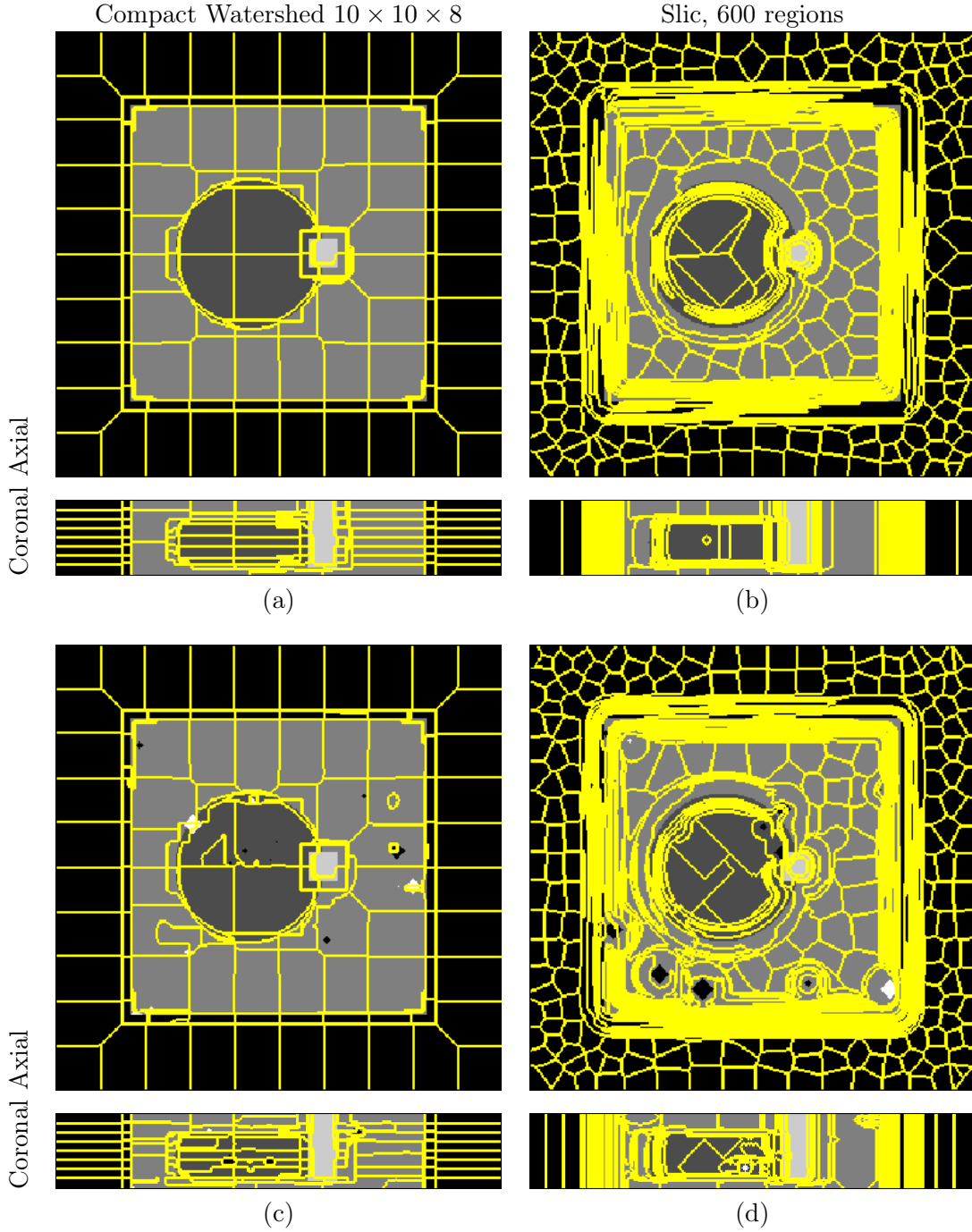
Compact Watershed $10 \times 10 \times 8$ Slic, 600 regions

(a)   (b)

(c)   (d)

Figure 4.2: *Selected results for the super-segmentation experiments. (a) and (b) are results for the noise profile f_0; (c) and (d) are results for the noise profile f_1. The upper image in each pair is the central axial slice; the lower image is the central coronal slice. Yellow lines denote the border between adjacent super-segmented regions. Results for f_2 and f_4 are found in Figure 4.3.*

seen, for example, in Figure 4.2 (d).

However, the SLIC does provide a small improvement over the watershed to the average overall similarity index. This is due to a better representation of both background and venal borders, specially when faced with noise surrounding these regions. An example of improper vein super-segmentation by the watershed algorithm can be seen in Figure 4.2 (a), as the venal regions contains a significant amount of voxels from the body. As our goal is accurate liver segmentation, though, the results present by the watershed algorithms are of greater interest.
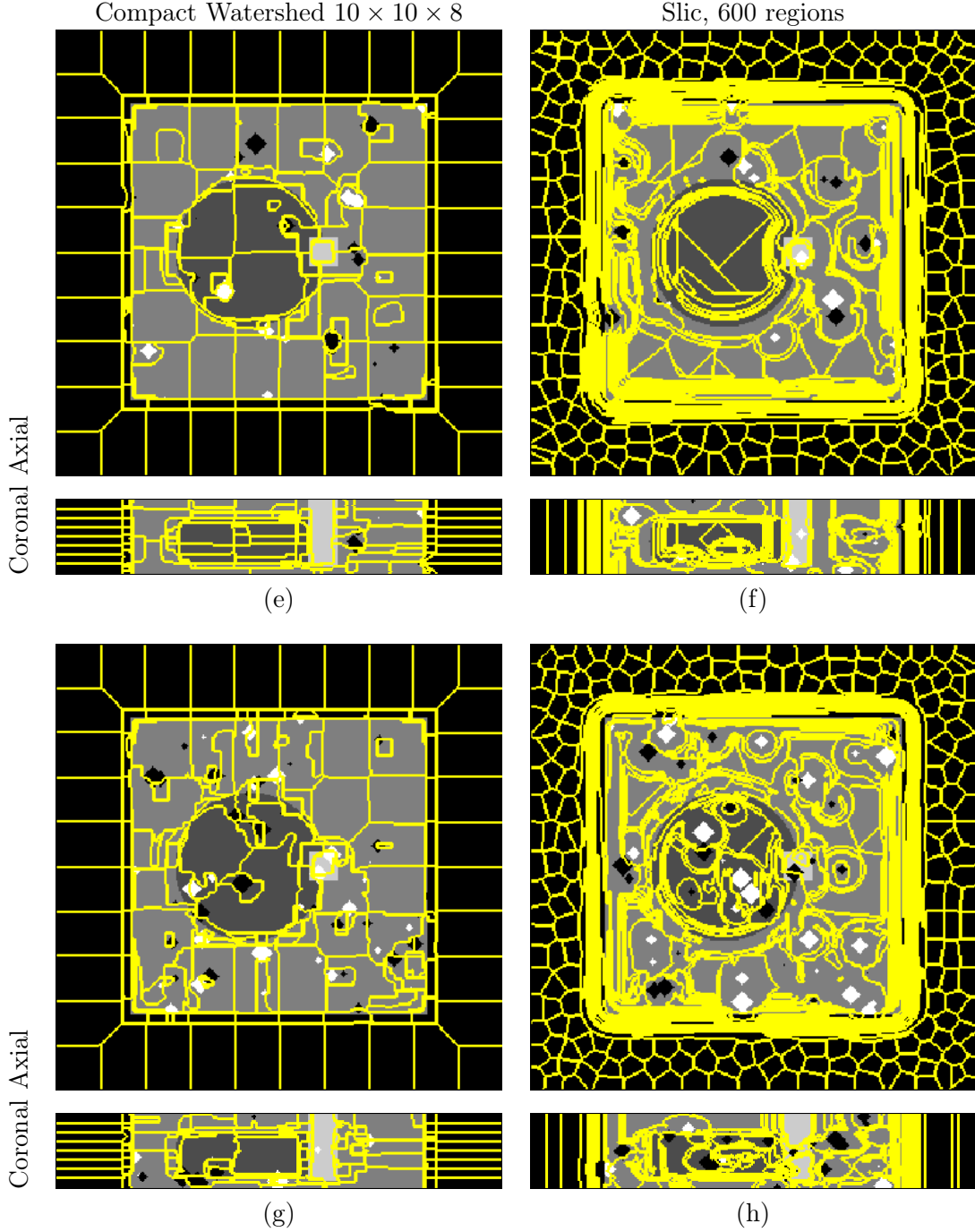
Figure 4.3: *Selected results for the super-segmentation experiments. (e) and (f) are results for the noise profile f_2; (g) and (h) are results for the noise profile f_4. The upper image in each pair is the central axial slice; the lower image is the central coronal slice. Yellow lines denote the border between adjacent super-segmented regions. Results for f_0 and f_1 are found in Figure 4.2.*

As expected, a larger amount of initial markers – and, consequently, a larger amount of subregions – results in a higher chance of the true borders being represented. However, while not the focus of this experiment or its analysis, practical concerns may arise, as a large number of regions may significantly increase the running time of high-complexity matching algorithms.

All of the results' 95% confidence interval have bounds separated by no more than 2% of the mean, pointing to said results' robustness to noise. Thus, we have reason to believe that each algorithm will perform as expected when faced with any new observation generated by these noise profiles. As such, we decide to utilize the *Compact Watershed* $10 \times 10 \times 8$ algorithm for further

experimentation, such as in Experiment 4.1.3.

Future implementations of SLIC, potentially reinforcing the anisotropic aspect of the 3D algorithm, could be explored as an alternative to those presented here. As the number of regions may adversely affect the matching algorithms, there is interest in future work that attempts to achieve precise borders with a small amount of regions.

4.2.3   Weight Calibration

We report the results of the weight calibration experiment in Table 4.3 and in Figures 4.4 and 4.5, considering the weight sets presented in Table 4.1. Due to the random nature of the experiment, 100 repetitions of each experiment were executed, and we report both the mean of the results and the 95% confidence interval for 100 repetitions.

We also provide visualizations of the predicted liver (in green) and vein (in red) for selected weight sets, with the accompanying ground truth in dark green and dark red for liver and vein, respectively, in Figures 4.6-4.11.

Discussion

Figures 4.4 and 4.5, along with Table 4.3, put into evidence several particular aspects of the effects of each attribute in the process of segmentation proposed, while further demonstrating the usefulness and potential of the calibration technique in the tuning of the SRG segmentation pipeline's parameters. Indeed, the results produced by this experiment are the fundamental example of use of the calibration approach proposed in this work.

It is important to reiterate the conclusions found in Section 4.2.1, and in particular the discussion on how the range of costs changes when the SRG context is different. With this in mind, the costs of one weight set are not necessarily directly comparable to those of another weight set; and while some of the lower costs are related to the best accuracies (as seen in Figure 4.4 (a) for $W = 4$) and some of the highest costs are related to the worst accuracies (as seen consistently across all results for the weight set $W = 2$), there is not a consistent cost-accuracy correlation across different weight sets; costs should only be directly compared within the same weight set.

The objective of generating different noise profiles, with increasing amounts of noise, was to verify the capabilities of a given SRG configuration in a controlled environment, and to test said configuration's robustness to the different levels of noise and artifacts that may be found in real data. As the results of this experiment show, even as the amount of noise increases, following the progression of the charts from (a) to (d), the quality of the results (as measured by the similarity index) remains consistent throughout. As expected, costs tend to increase as noise increases, as even the optimal noisy solution performs worse than the solution for the clean model. Even so, high accuracy scores are produced even when faced with great noise. The high noise profile, f_4, shows a drop of only 5% in liver SI for the highest scoring case when compared with the zero-noise profile, f_0.

Consistently across all noise profiles, the weight set with the best accuracy results – both liver and overall average – is $W = 4$, that is, the set which emphasizes the *vertex intensity* attribute. This may be due to the relative homogeneity of the intensities of the calibration phantom in the current setup, where each structure of interest has a relatively specific intensity, even with the FSP noise. This may also be due to the relative simplicity of the liver segmentation task in the given context, where in fact a human expert might default to utilizing intensity in assessing the nature

Table 4.3: *Results for the weight calibration experiments, for the weight sets defined in Table 4.1. Cost (Equation 3.2, average similarity index (SI, Equation 3.5) of all classes and liver SI were reported for each of the weight sets presented in Table 4.1, on each noise profile. The 95% confidence interval is reported.*

| Noise	Weight Set	Cost	Average SI	Liver SI
f_0	W_0	0.0850 ± 0.0000	0.8394 ± 0.0000	0.9012 ± 0.0000
	W_1	0.0603 ± 0.0000	0.8612 ± 0.0000	0.9340 ± 0.0000
	W_2	0.2365 ± 0.0000	0.6015 ± 0.0000	0.7352 ± 0.0000
	W_3	0.1513 ± 0.0000	0.6621 ± 0.0000	0.5890 ± 0.0000
	W_4	0.0306 ± 0.0000	$\mathbf{0.9188 \pm 0.0000}$	$\mathbf{0.9680 \pm 0.0000}$
	W_5	0.0854 ± 0.0000	0.8394 ± 0.0000	0.9012 ± 0.0000
	W_6	0.1700 ± 0.0000	0.7218 ± 0.0000	0.9012 ± 0.0000
	W_7	0.1679 ± 0.0000	0.7194 ± 0.0000	0.9012 ± 0.0000
	W_8	0.1677 ± 0.0000	0.7218 ± 0.0000	0.9012 ± 0.0000
	W_9	0.1852 ± 0.0000	0.7194 ± 0.0000	0.9012 ± 0.0000
	W_{10}	0.1313 ± 0.0000	0.7805 ± 0.0000	0.9012 ± 0.0000
f_1	W_0	0.1696 ± 0.0608	0.8157 ± 0.0046	0.8955 ± 0.0017
	W_1	0.3600 ± 0.4014	0.8317 ± 0.0077	0.9136 ± 0.0062
	W_2	0.6556 ± 0.1760	0.5773 ± 0.0049	0.7329 ± 0.0058
	W_3	1.1305 ± 0.4085	0.6577 ± 0.0047	0.6064 ± 0.0082
	W_4	0.1215 ± 0.0908	$\mathbf{0.8997 \pm 0.0016}$	$\mathbf{0.9510 \pm 0.0012}$
	W_5	0.1700 ± 0.0608	0.8168 ± 0.0041	0.8954 ± 0.0017
	W_6	0.2402 ± 0.0501	0.7028 ± 0.0038	0.8955 ± 0.0017
	W_7	0.2372 ± 0.0511	0.7031 ± 0.0039	0.8957 ± 0.0016
	W_8	0.2383 ± 0.0501	0.7028 ± 0.0038	0.8955 ± 0.0017
	W_9	0.2534 ± 0.0507	0.7030 ± 0.0038	0.8957 ± 0.0016
	W_{10}	0.3007 ± 0.1097	0.7242 ± 0.0093	0.8955 ± 0.0017
f_2	W_0	0.5575 ± 0.2214	0.7809 ± 0.0089	0.8864 ± 0.0032
	W_1	0.8943 ± 0.7127	0.8084 ± 0.0093	0.9012 ± 0.0068
	W_2	1.7681 ± 0.3493	0.5425 ± 0.0069	0.7351 ± 0.0076
	W_3	2.5906 ± 0.6798	0.6385 ± 0.0076	0.6286 ± 0.0130
	W_4	0.5974 ± 0.2901	$\mathbf{0.8703 \pm 0.0042}$	$\mathbf{0.9367 \pm 0.0017}$
	W_5	0.5579 ± 0.2215	0.7821 ± 0.0085	0.8844 ± 0.0041
	W_6	0.5628 ± 0.1838	0.6734 ± 0.0072	0.8848 ± 0.0049
	W_7	0.5614 ± 0.1852	0.6755 ± 0.0074	0.8854 ± 0.0042
	W_8	0.5610 ± 0.1838	0.6734 ± 0.0072	0.8848 ± 0.0049
	W_9	0.5756 ± 0.1845	0.6754 ± 0.0075	0.8854 ± 0.0042
	W_{10}	0.9387 ± 0.3694	0.6958 ± 0.0112	0.8868 ± 0.0033
f_4	W_0	1.5742 ± 0.4751	0.7342 ± 0.0099	0.8770 ± 0.0032
	W_1	1.9123 ± 1.2918	0.4417 ± 0.0737	0.5183 ± 0.0847
	W_2	4.1689 ± 0.7545	0.4910 ± 0.0081	0.7305 ± 0.0093
	W_3	4.2040 ± 0.7979	0.6095 ± 0.0077	0.6315 ± 0.0143
	W_4	1.8234 ± 0.4904	$\mathbf{0.8188 \pm 0.0057}$	$\mathbf{0.9111 \pm 0.0025}$
	W_5	1.5748 ± 0.4751	0.7357 ± 0.0096	0.8754 ± 0.0038
	W_6	1.4084 ± 0.3955	0.6327 ± 0.0083	0.8750 ± 0.0052
	W_7	1.4134 ± 0.3970	0.6336 ± 0.0085	0.8758 ± 0.0043
	W_8	1.4065 ± 0.3955	0.6327 ± 0.0083	0.8750 ± 0.0052
	W_9	1.4246 ± 0.3964	0.6331 ± 0.0085	0.8758 ± 0.0043
	W_{10}	2.6395 ± 0.7909	0.6370 ± 0.0101	0.8737 ± 0.0059

(a)

(b)
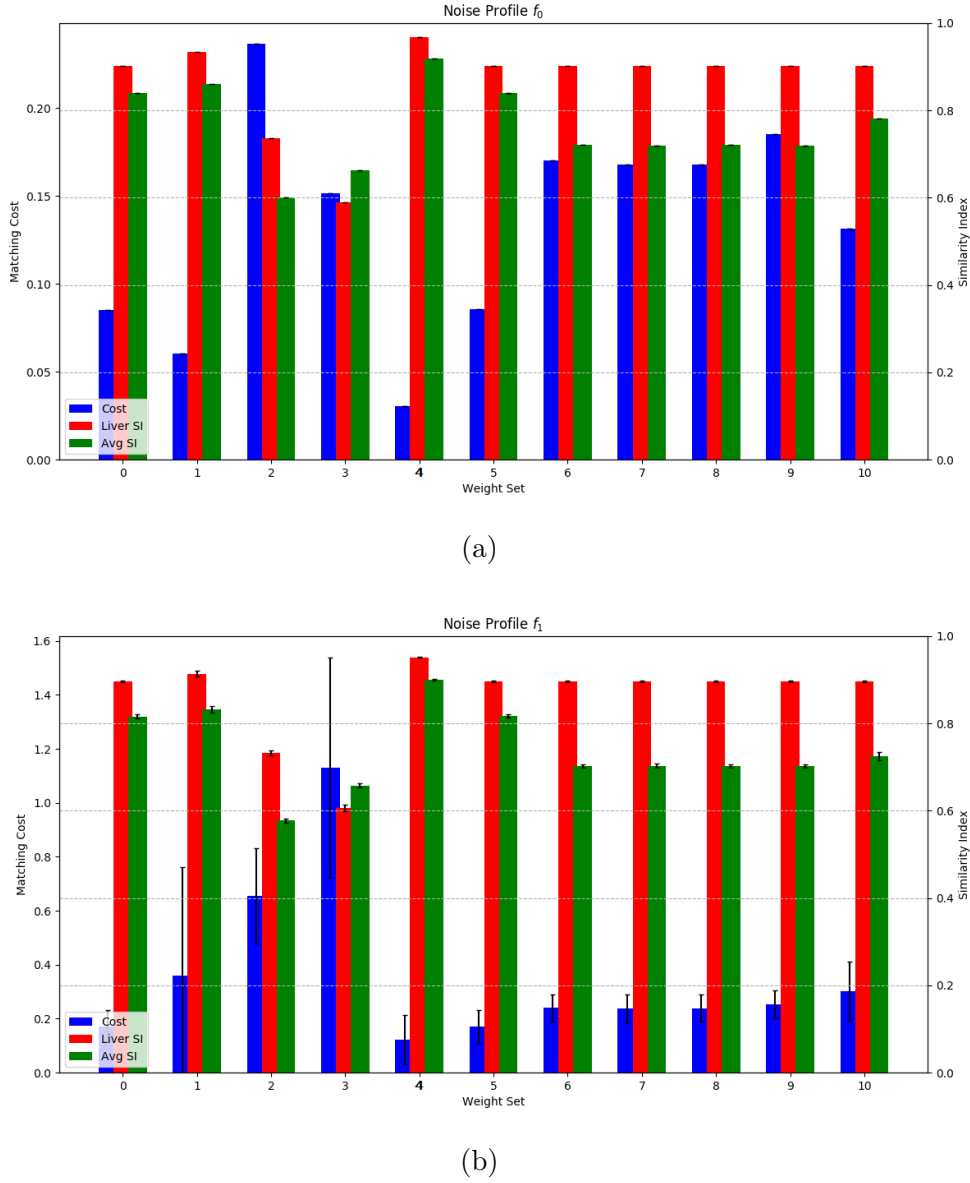
Figure 4.4: *Results for the weight calibration experiments, for the weight sets defined in Table 4.1. (a) are the results for the noise profile f_0; (b) are the results for the noise profile f_1. The black bars are the 95% confidence interval for each result. We report the matching cost (Equation 3.2), the Similarity Index, or SI (Equation 3.5), of the liver and the average of the SIs of all labels.*

of each region in the phantom. In fact, future calibration experiments can measure the impact each individual attribute has in the final matching cost and, through that, automatically provide a reasoning for a decision, such as "this region is liver because it has the intensity of liver, despite not having exactly the position of liver".

Figures 4.10 and 4.11 display the results of $W = 4$ segmentation on a sample set of noise profiles. The results for the liver region (represented in bright green) are encouraging and highly accurate, even in the high noise case (noise profile f_4). The borders are maintained quite close to the optimum (as limited by the super-segmentation algorithm) and to the ground truth. The vein is also accurately segmented, especially in comparison to the other visualizations; however, it is noteworthy to point out that this weight set completely mislabels the vein with the noise profile f_4.

The even weight set $W = 0$ also presented consistently good performance across all noise profiles.
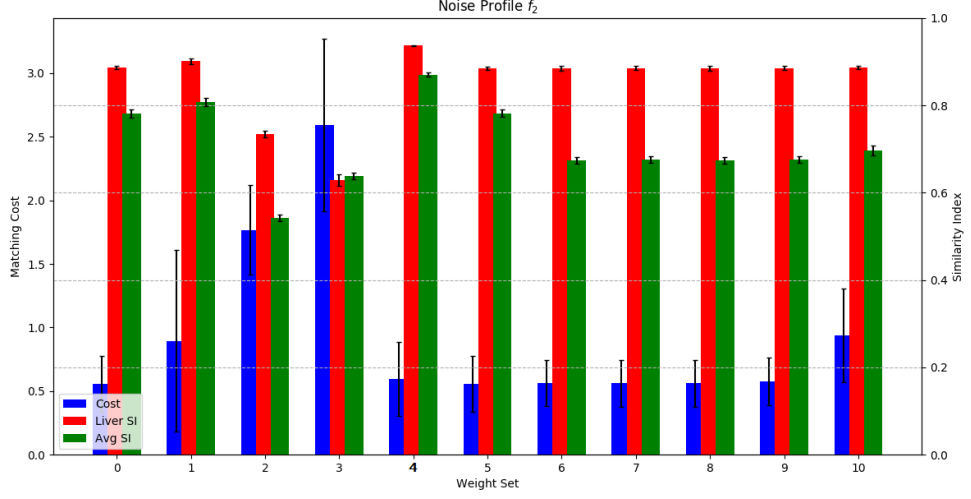
(c)
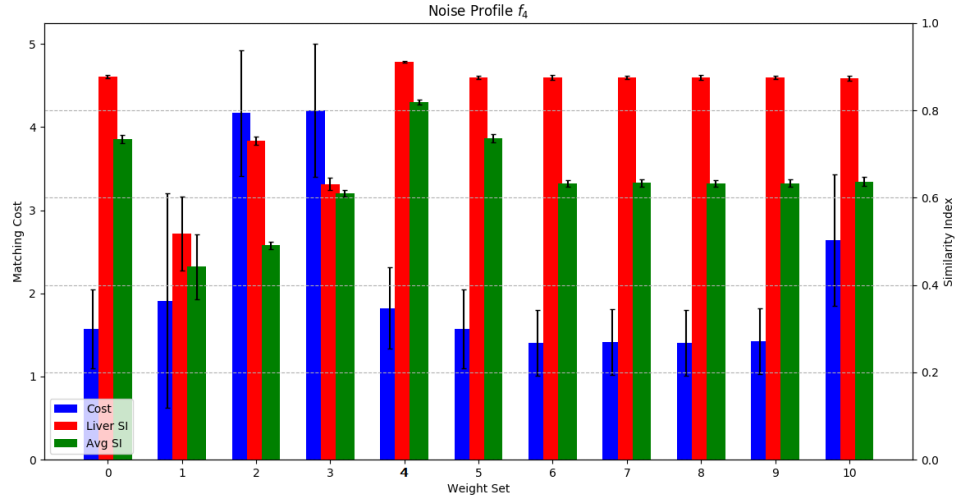
(d)

Figure 4.5: *Results for the weight calibration experiments, for the weight sets defined in Table 4.1. (c) are the results for the noise profile f_2; (d) are the results for the noise profile f_4. The black bars are the 95% confidence interval for each result. We report the matching cost (Equation 3.2), the Similarity Index, or SI (Equation 3.5), of the liver and the average of the SIs of all labels.*
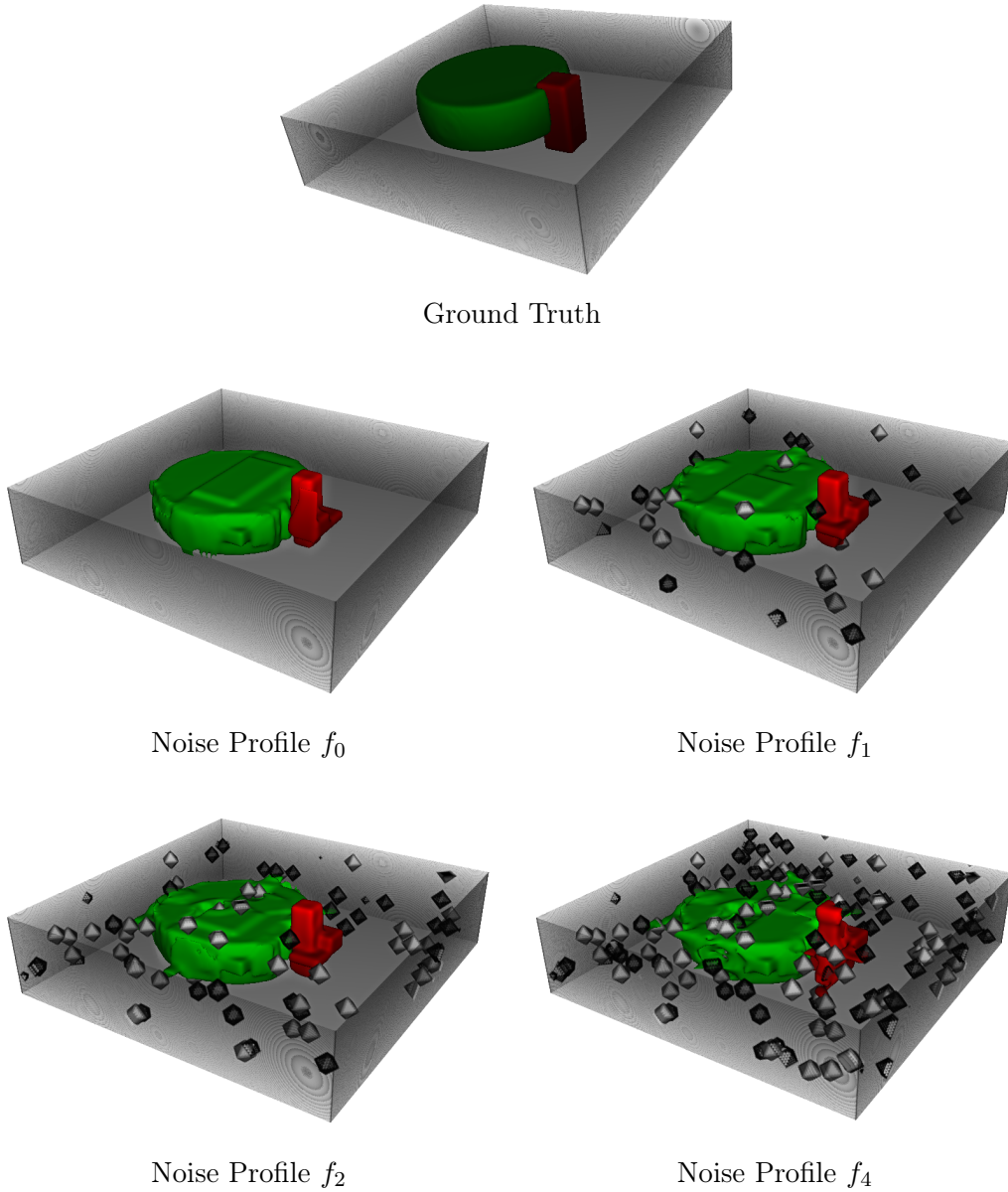
Ground Truth

Noise Profile f_0

Noise Profile f_1

Noise Profile f_2

Noise Profile f_4

Figure 4.6: *3D visualizations of the results for the $W = 0$ calibration experiment. The predicted liver region is shown as the green volume, and the predicted vein region is shown as the red volume. The observation volume's body and noise are also represented. The ground truth is the topmost image. Noise was generated with a fixed seed of 1.*

This is an encouraging result, as it does demonstrate the value of utilizing the ensemble of the attributes devised in order to properly represent each segment visualized. This also showcases the "semantic" nature of the segmentation, as a representation of what attributes go into the reasoning for a given segmentation. As discussed above, relative comparison of the impact of each attribute in a given cost – specially when the even-weighted set provides good and accurate results – may provide further guidance for the calibration process.

Figures 4.6 and 4.7 display the results of $W = 0$ segmentation on a sample set of noise profiles. The segmentation of the liver region (represented in bright green) shows good accuracy, which nonetheless declines more with the increase in noise than the weight set $W = 4$; with the results for the noise profile f_4 displaying a significant dip in quality. The segmentation of the vein region also displays good accuracy, declining as noise rises. As with $W = 4$, most of the inaccuracies take
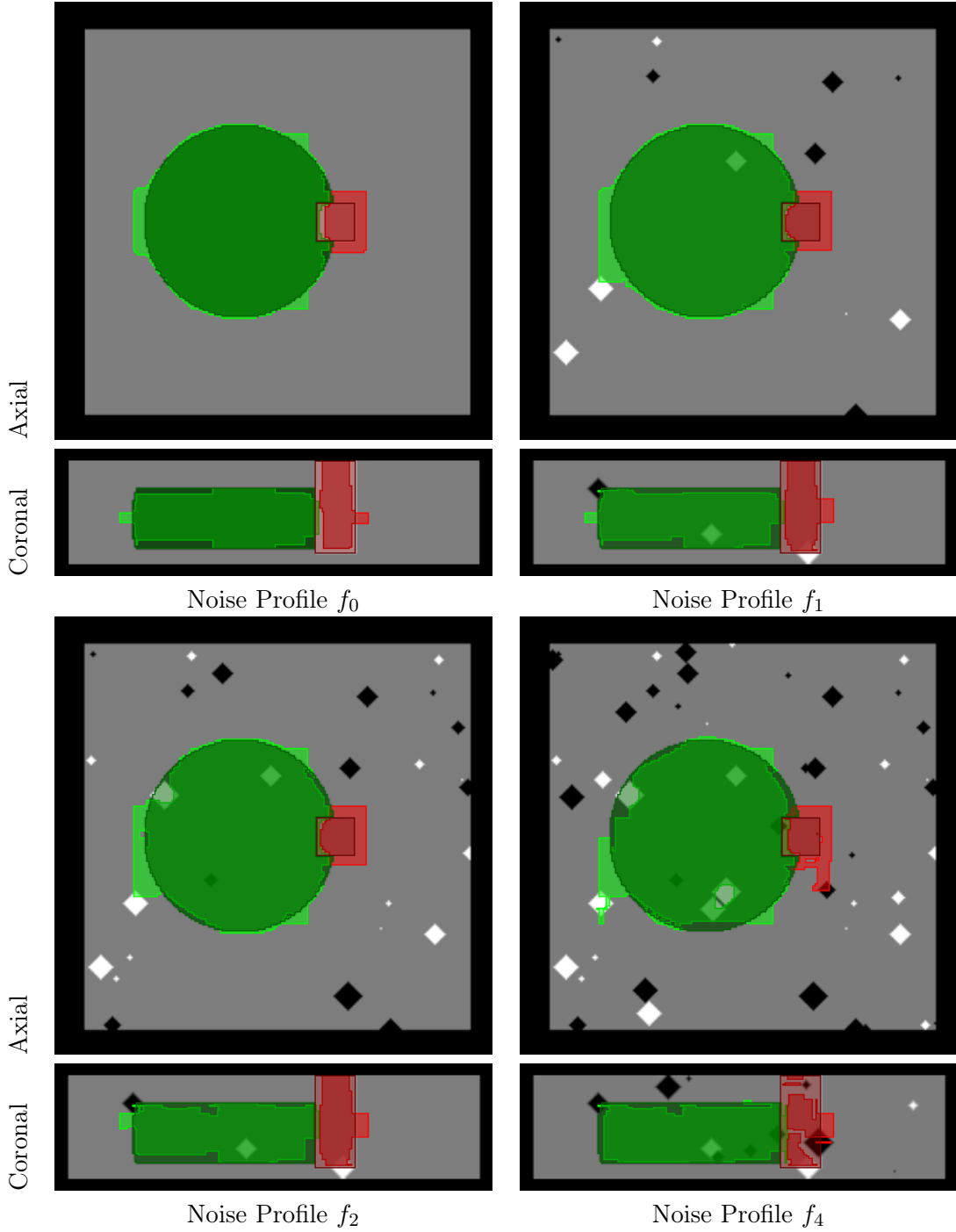
Figure 4.7: *2D visualizations of the central slices of the results for the $W = 0$ calibration experiment. The upper image in each pair is the central axial slice; the lower image is the central coronal slice. The predicted liver region is shown as the bright green area, and the predicted vein region is shown as the bright red area. The ground truth is shown as the dark green and dark red area for the true liver and true vein, respectively. Noise was generated with a fixed seed of 1.*

the form of false-positives, for both structures.

The results for weight sets $W = 1, 2, 3$ provide useful insight in one of the most intuitive attributes, the region centroid. As can be expected given the nature of the calibration phantom – which aligns all relevant structures on the x-axis – the weight sets $W = 2$ and $W = 3$, representing focus on the centroid's y and z attributes, respectively, present the worst results across all noise profiles. In fact, they are most vulnerable to being affected by noise, as shown by the loose confidence
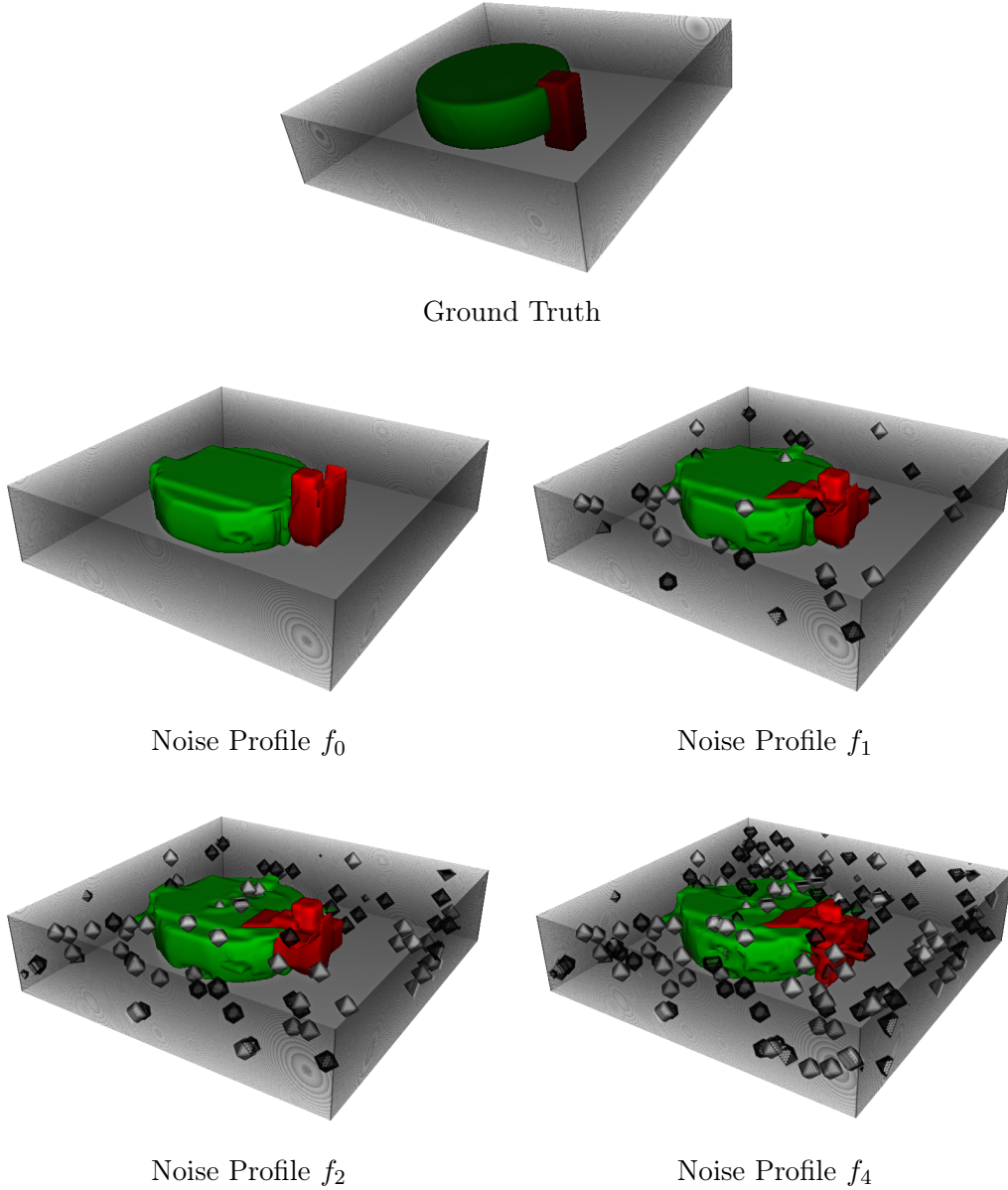
Ground Truth

Noise Profile f_0 Noise Profile f_1

Noise Profile f_2 Noise Profile f_4

Figure 4.8: *3D visualizations of the results for the $W = 1$ calibration experiment. The predicted liver region is shown as the green volume, and the predicted vein region is shown as the red volume. The observation volume's body and noise are also represented. The ground truth is the topmost image. Noise was generated with a fixed seed of 1.*

intervals. These factors point to a lower discrimination capacity for these attributes, which, given the phantom's nature (and an actual human liver's nature, as well), is not unexpected.

Conversely, the weight set which emphasizes the centroid's x coordinate, $W = 1$, displays the second-best results. This may be explained by the same mechanisms which explain the poor performance of the other coordinates: the nature of the phantom is such that the position of a region alongside the x-axis is highly relevant. Thus, along with intensity, we can infer that the centroid's x coordinate is a vertex attribute with high capability for discrimination. In a sense, these results are providing a semantic interpretation of the observation image, by stating that "in order to recognize the structures in this, we can focus on their intensity and their position in the x-axis". Indeed, when examining the phantom (or even, indeed, a real MR image), a natural and intuitive observation is "the liver is the dark-grey region on the center-left side of the image".
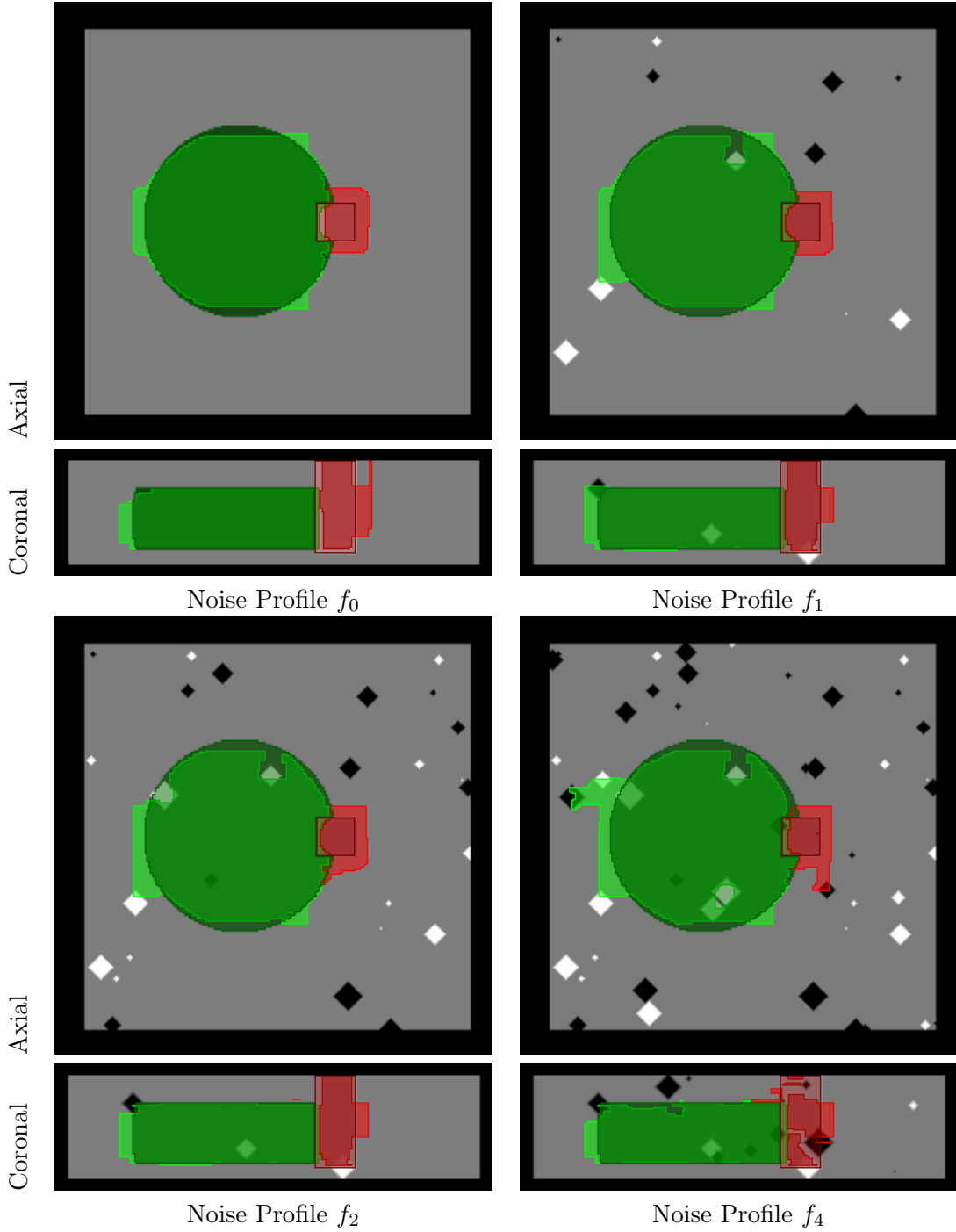
Figure 4.9: *2D visualizations of the central slices of the results for the $W = 1$ calibration experiment. The upper image in each pair is the central axial slice; the lower image is the central coronal slice. The predicted liver region is shown as the bright green area, and the predicted vein region is shown as the bright red area. The ground truth is shown as the dark green and dark red area for the true liver and true vein, respectively. Noise was generated with a fixed seed of 1.*

However, when faced with a high amount of noise in the noise profile f_4, the centroid's x coordinate becomes a much less discriminant attribute, demonstrating much lower mean accuracies and a greater variability in both cost and accuracy, consistent with low-discriminative attributes in other noise profiles. While the visualization present in Figures 4.8 and 4.9 do not showcase a particularly inaccurate prediction, we can notice large mislabeled regions for both vein and liver in comparison to the other weight sets, and in particular a much greater amount of false-positives. In Figures 4.2 and 4.3, we also see larger regions in the f_4 noise profile results, whose centroid
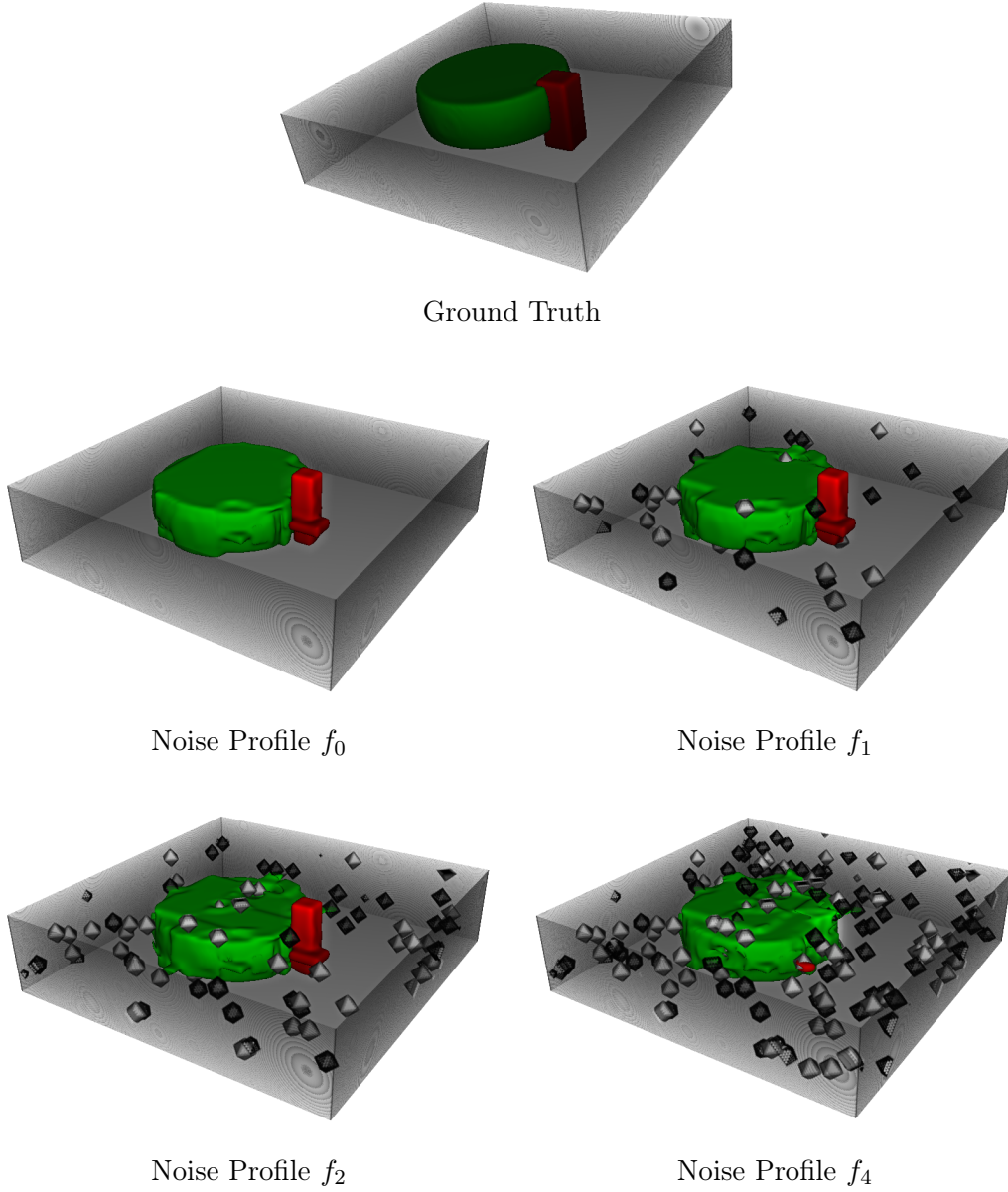
Ground Truth

Noise Profile f_0

Noise Profile f_1

Noise Profile f_2

Noise Profile f_4

Figure 4.10: *3D visualizations of the results for the $W = 4$ calibration experiment. The predicted liver region is shown as the green volume, and the predicted vein region is shown as the red volume. The observation volume's body and noise are also represented. The ground truth is the topmost image. Noise was generated with a fixed seed of 1.*

coordinates do not correlate neatly with the true region they represent. We may thus conclude that the centroid's x attribute is vulnerable to high amounts of noise.

Figures 4.8 and 4.9 display the results of $W = 1$ segmentation on a sample set of noise profiles. The liver regions (represented as bright green) display somewhat good accuracy, with false-positives to the left of the true liver and false-negatives in the posterior and anterior (that is, top and bottom of the axial slice, respectively) of the true liver. This, of course, is to be expected when the centroid's x is the emphasized attribute. Of note is the increased susceptibility to noise of both the liver and the vein prediction, when compared to the visualizations of the other weight sets.

The weight sets $W = 6, 7, 8, 9, 10$, by comparison, seem to have a much smaller impact on the results than the others, specially when compared to the even-weighted set $W = 0$. As all of these
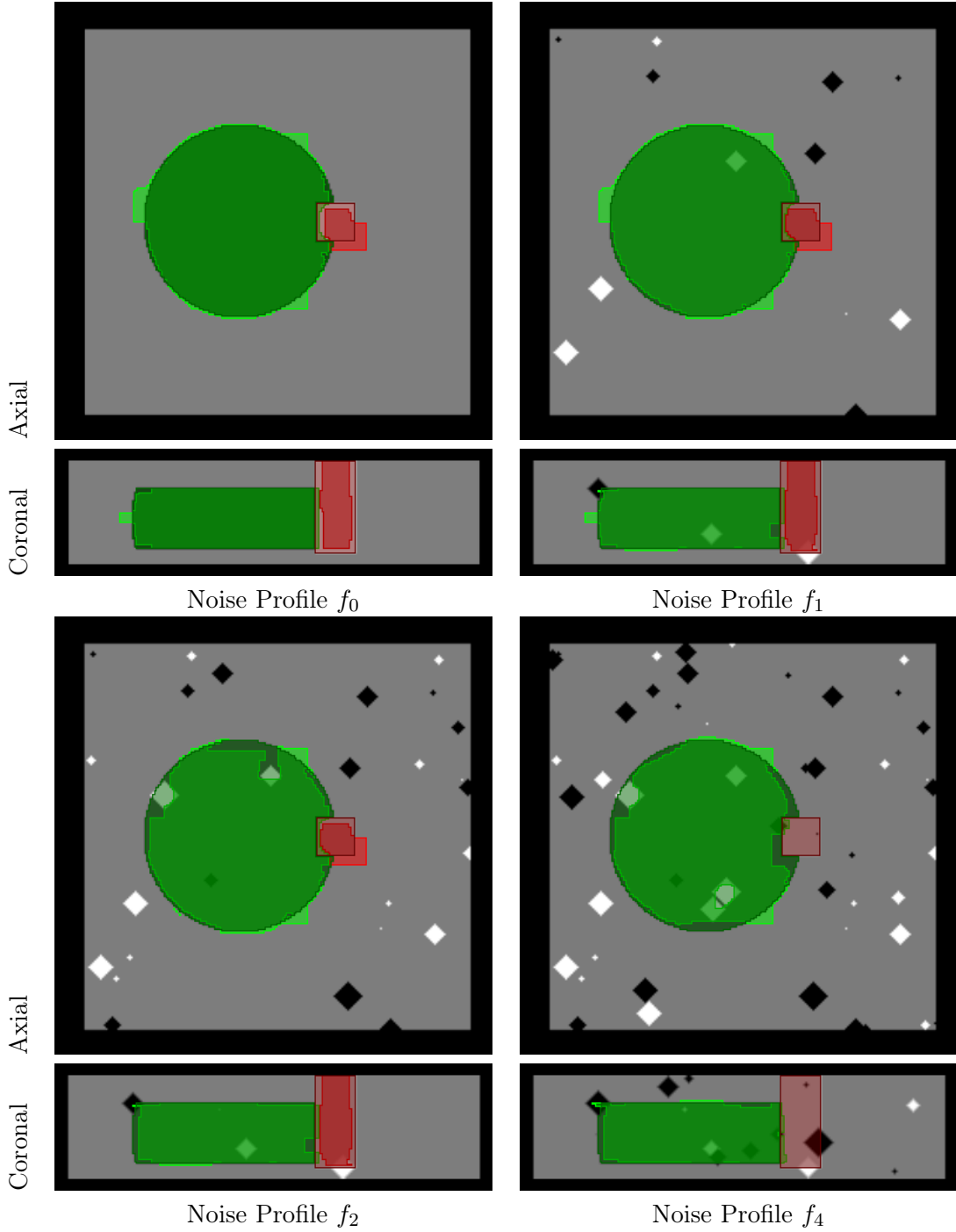
Figure 4.11: *2D visualizations of the central slices of the results for the $W = 4$ calibration experiment. The upper image in each pair is the central axial slice; the lower image is the central coronal slice. The predicted liver region is shown as the bright green area, and the predicted vein region is shown as the bright red area. The ground truth is shown as the dark green and dark red area for the true liver and true vein, respectively. Noise was generated with a fixed seed of 1.*

weight sets are those emphasizing the edge attributes, we can infer that the edge attributes have a much smaller impact on the solution than the vertex attributes. Average accuracy does tend to take a hit, likely because each body division may be inaccurately segmented but still maintain their structural relationships. This is not particularly surprising, as the structural information is most useful as a guide in improving the recognition of the individual structures than as the cornerstone of the segmentation process. However, further experiments which emphasize the cost of the edges over the cost of the vertex might produce more significant differences in the final result.

4.2.4 Preliminary Application on Real Data

We report the results of the real data application experiment on Table 4.4, as well as provide visualizations of the resulting segmentation in Figures 4.12 and 4.13. We also provide images of some slices displaying the super-segmented regions in Figure 4.14. An in-depth analysis of the results regarding the false-positives in the spine are shown in Figure 4.15.

Table 4.4: *Results for the preliminary application of the SRG segmentation in real data. We report matching cost, average similarity index (SI) between the prediction and truth of all classes and SI of the prediction and truth of the liver.*

Matching Cost	27.44
Average SI	0.5253
Liver SI	0.5618

Anterior view

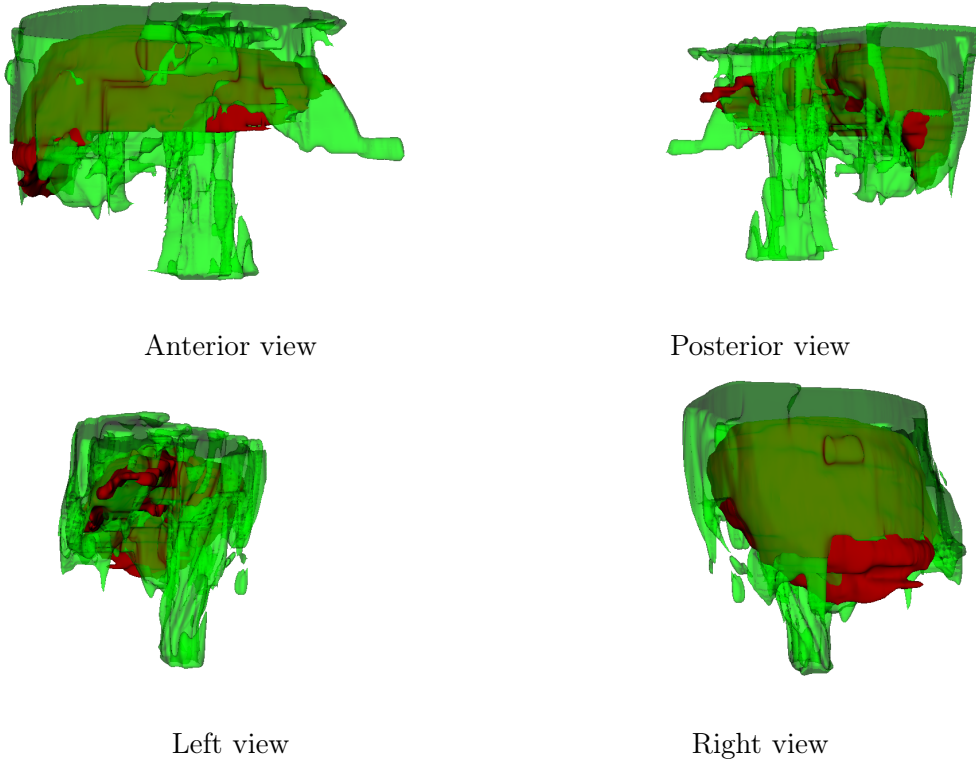Posterior view

Left view

Right view

Figure 4.12: *Anterior, posterior, left and right views of the 3D visualization of the real MRI data segmentation results. The green region is the predicted liver; the red region is the true liver.*

Discussion

Table 4.4 displays the quantitative results for the segmentation of real data using the SRG. As expected from a preliminary experiment, the results shown display insufficient accuracy for real medical application. However, both the reported similarity index and the visualizations showcase interesting results; for instance, most of the errors are due to false-positives, and not false-negatives – thus, it is possible to apply further operations to increase accuracy using only the segmented region. Additionally, the rough liver shape was acquired, and some of the most glaring mistakes are similar to rookie human mistakes in manual segmentation (such as having trouble telling the lung apart from the liver, as seen in Figure 4.13 Slice 40).
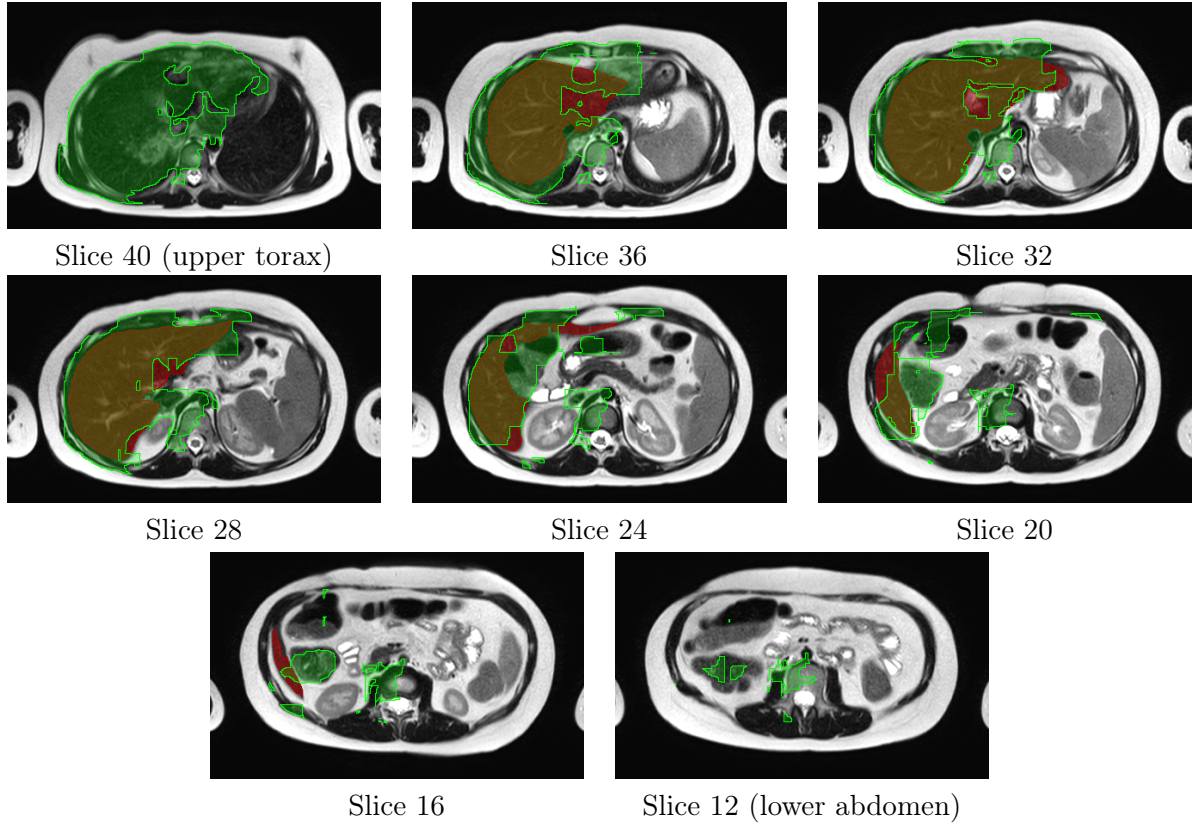
| Slice 40 (upper torax) | Slice 36 | Slice 32 |

| Slice 28 | Slice 24 | Slice 20 |

| Slice 16 | Slice 12 (lower abdomen) |

Figure 4.13: *Axial slices of the real MRI data, with the liver prediction and the true liver overlayed (in green and red, respectively). Higher-numbered slices are closer to the head of the patient.*

These preliminary results point to the potential of the SRG pipeline for the accurate segmentation of structures of interest in real MRI data and, as such, provide motivation towards both this work's and the project's stated goal of developing a SRG-based technique for semantic segmentation of MRI data.

As Figures 4.12 and 4.13 point out, the vast majority of liver segmentation errors are false-positives, that is, the segmentation algorithm labeled a non-liver region as liver. Some false-positives of note: the lungs, located above the liver (visible in Figure 4.13 Slice 40), which have very similar intensities to the liver in the image; the spine (emphasized in Figure 4.15); and the ribcage (visible in Figure 4.13, Slices 28-36), which borders the liver. A few false-negatives are present, such as in the Segment I of the liver, the central region which surrounds the vena cava (visible in Figure 4.13, Slices 28-36) and the Segment III of the liver, the rightmost region of the inferior liver (visible in Figure 4.13 Slice 24).

Figure 4.14 put into evidence the fact that some of the errors noted in the final segmentation are the result of improper super-segmentation. Of immediate notice is the fact that both the spine joints and the nerves are super-segmented as a single region each, while the vertebrae are segmented as a small set of large, monolithic sub-regions (this is visible in the spine sagittal slice). Thus, the mislabeling of one of these large regions, due to a local improvement of the matching cost, results in a large segmentation error. Additionally, the upper abdomen axial slice reveals that the ribcage (above and to the left of the liver) is inside the same subregion as the liver and, as such, it is impossible to properly segment it from the liver. This showcases the need for a solid, adequate super-segmentation algorithm, and feeds back into the calibration method in the form of putting into evidence complications that the current phantom and noise profiles do not foresee.

Despite this, the choice of the weight set $w = 4$ produced interesting results, with relatively few
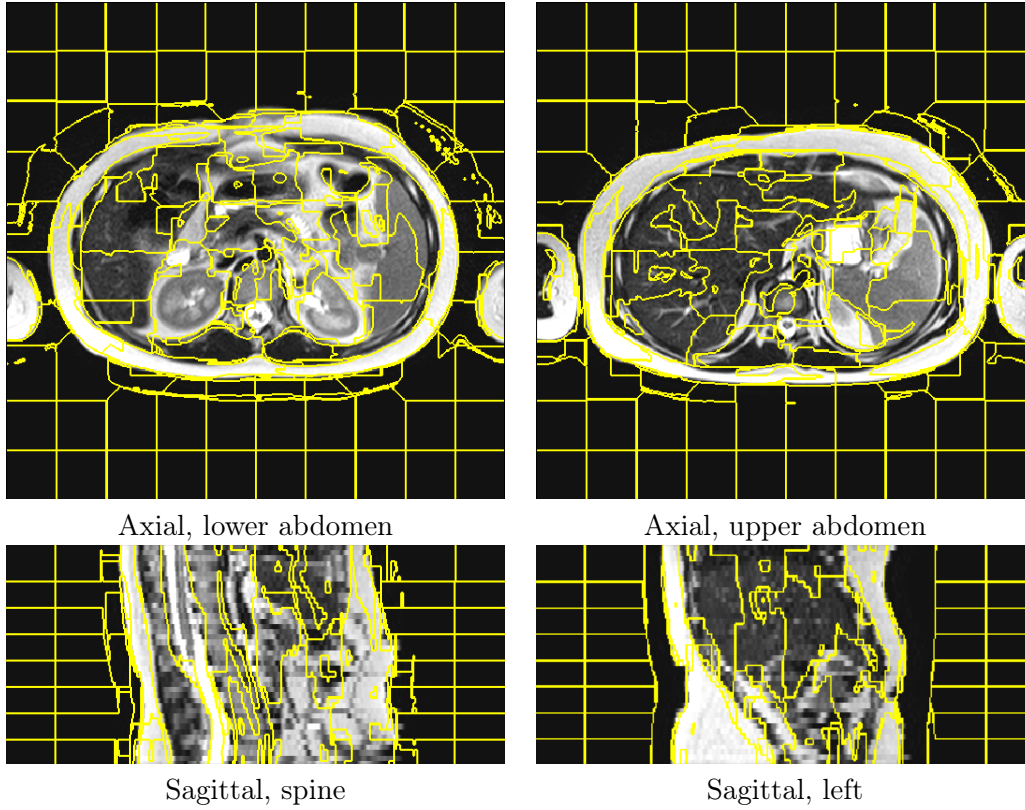
Axial, lower abdomen   Axial, upper abdomen

Sagittal, spine   Sagittal, left

Figure 4.14: *Selected representations of the super-segmentation of the real MRI data. Yellow lines denote the border between adjacent super-segmented regions.*

false-negatives; as the liver in real MRI data is quite distinguishable from most of its neighbours by intensity alone. Some regions, however, are affected negatively by this focus on intensity, with the lungs being the most obvious example. Intensity, however, does have particular challenges in a real set, such as different dynamic ranges for different images – even if acquired from the same machine – as external and environmental circumstances, such as disturbances in the magnetic field, may affect the intensity reported for a given voxel. In fact, there is support for the use of the gradient of the intensity as opposed to the true intensity as the image to be segmented.

These results showcase how to best improve the calibration technique. As the contrast between the results of both the super-segmentation and the segmentation pipeline between the real data and the calibration attest, there is a need to improve the representations of the real data difficulties in the calibration technique. Perhaps most notable is the inhomogeneity of the real body when compared to the calibration phantom's body; as two of the greatest challenges put into evidence are the distinguishing of regions with very similar intensity (such as the lung and the liver) and of other macro-regions in the body (such as the spine).

The lack of accuracy in the final results, coupled with the high matching cost, point out to an additional weak point in the current SRG pipeline: the improvement algorithm, which follows a greedy brute-force approach and is thus vulnerable to local minima. Even after a large number of improvement steps, the final results were subpar; better matching and solution improvement algorithms may help improve these results. An exploration of different improvement algorithms may be the subject of future work.

Further calibration phantoms of the abdomen may benefit from the simulation of other macro structures (such as the spleen or the spine), and of less distinguishable regions (such as the lungs). Additionally, other noise profiles may be devised to further approach the calibration to the real data, by simulating artifacts common in real data such as Gibbs ringing.
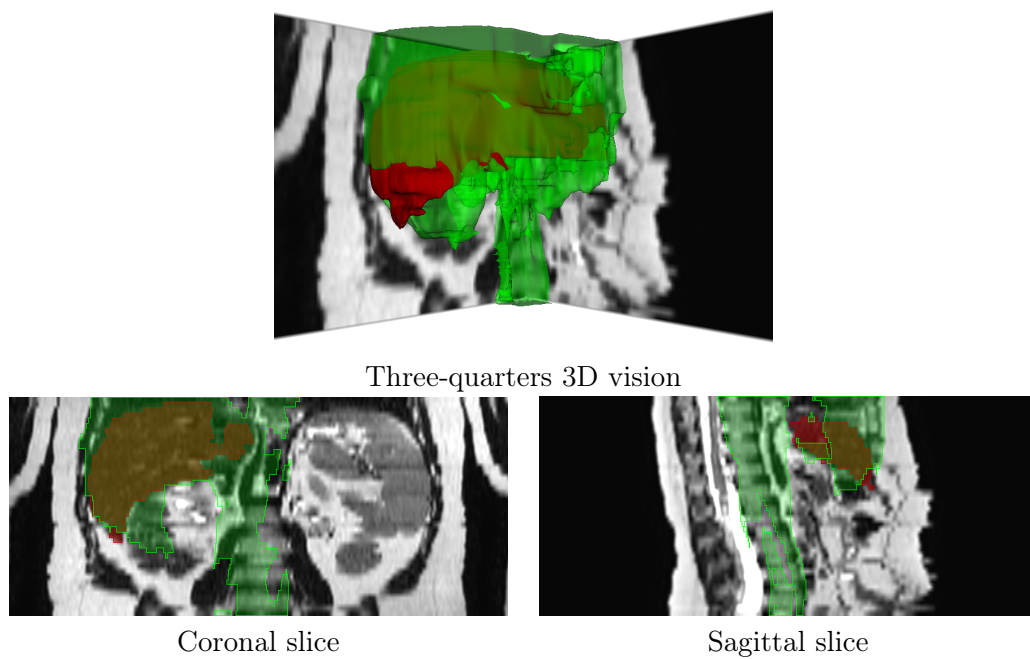
Three-quarters 3D vision

Coronal slice   Sagittal slice

Figure 4.15: *Selected representations of the real MRI data segmentation, emphasizing the false-positives in the spine. The top image is a three-quarters view of a 3D visualization of the results; the slices intersecting the 3D visualization are those shown in the bottom images. The green region is the predicted liver; the red region is the true liver.*

Chapter 5

Concluding Remarks

This work proposes the use of a Structural-Relational Graph-based approach to the semantic segmentation of 3D Magnetic Resonance Images; along with a pipeline for the calibration of its parameters. We explain the process of SRG segmentation, the effect and importance of each of its parameters, and how to apply it to a set of data. We also explain the calibration process, its use in the assessment of the quality of a given set of parameters, and how to apply it to a set of parameters. We propose and perform experiments to demonstrate the usefulness and validity of the calibration technique, and to open future avenues of research in the area.

In order to perform SRG-based semantic segmentation, a set of attributes, related to either the structures of interest in the image or to the relationships between structures, must be chosen. These attributes must be capable of adequately discriminating the structures of interest in the image. A model graph must then be learned and built from annotated data; future observations will be matched using this model as template. Each vertex in the graph corresponds to a single structure of interest in the scene; each edge corresponds to the relationship between structures. Observed data is super-segmented and a super-observation graph is built, with each vertex in this graph representing a subregion of the observed data. The super-observation graph is then matched to the model graph in order to build an observation graph, which represents the segmentation of the image that best approaches the model. This representation offers a semantic segmentation of the observed data.

The calibration technique begins with the generation of a predictable, controllable "phantom" of the structured scene, which is a simplified simulation of the real scene. A set of SRG parameters, including attributes, model learning technique, super-segmentation algorithm, cost functions and improvement algorithms, is chosen for assessment. Controlled and known noise may be added to this "phantom", in order to test the robustness of the chosen parameter set to noise. Observation data, with known ground truth values, is generated and segmented with the SRG pipeline. The quality of the segmentation is assessed both quantitatively and qualitatively.

Experiments were designed in order to test the validity and usefulness of both the implementation of SRG segmentation and of the calibration technique. A correlation between the cost of matching an observation graph with the model and the quality of the segmentation produced was demonstrated. This supports our claim that the minimization of the matching cost is a good technique for maximizing the probability that the observation corresponds to the model. A comparative study of super-segmentation algorithms in the context of calibration was conducted. A comparative study of different weights for the cost of matching distinct attributes was conducted, demonstrating the usefulness of the calibration technique in assessing the quality of a given set of parameters for the SRG-based segmentation. A preliminary application of SRG semantic segmentation in real

data was explored, with its results displaying further avenues for the improvement of both the segmentation parameters and the calibration technique.

An open-source implementation of the segmentation and calibration software was produced and is available on GitHub[1]. This software is the initial step towards building an application capable of attending to the goals of the project into which this work is inserted, providing a tool for medical experts to use in the segmentation of livers in pediatric MRI, for both assessment of iron liver content and volumetry of liver subdivisions for guiding transplants and surgery.

5.1   Future Work

The conclusion of this work is the first step in a larger project, which aims to provide a framework for fast and accurate segmentation of medical images. Future works in this project may tackle several issues raised in this work.

Providing a more in-depth analysis of the matching cost in the calibration, by showcasing the relative costs of vertex matching, edge matching, and the effect of each individual attribute in both the final and partial matching cost, may help in guiding the choice and adjustment of the parameter set chosen for the segmentation process, thus increasing the overall usefulness of the calibration technique.

The stochastic learning of the parameters that control the distribution of attributes in a given domain is of utmost importance to both model learning and to the determination of the probability that the observation graph is equal to the model graph. As stated in Section 3.1.2, in this work we have indirectly measured this probability through the use of the matching cost. We suspect that the matching cost, in truth, represents an assumption that all parameters have a normal distribution with the same co-variance matrices and thus the probability of the observation and the model being equal is solely dependent on the mean of the attributes (whose distances we minimize, maximizing the probability). One of the aims of the project is to both provide a method for parameter learning and to test this assumption.

The calibration phantom may be improved by increasing its simulation capability, through the representation of other body structures – further increasing its capabilities of simulating real MRI data in a controlled manner – or, alternatively, through the development and usage of novel noise profiles, capable of representing problems commonly found in real data, such as acquisition noise caused by weaker magnetic fields, the "spotlight" effect caused by field inhomogeneity, and the very common Gibbs ringing artifacts.

Finally, the parameters which compose the segmentation pipeline may be explored with greater depth; future works may explore a larger set of both matching and improvement algorithms, as well as different cost measures, in order to improve the quality of the final segmentation; and explore other super-segmentation algorithms in order to improve the potential quality of the segmentation.

---

[1]https://github.com/MarEe0/STAP-liver

Bibliography

[AKJ16]   MR Avendi, Arash Kheradvar e Hamid Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. *Medical image analysis*, 30:108–119, 2016. 11

[ASS+12]   Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, Sabine Süsstrunk et al. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 26, 31

[BCDK+16] Avi Ben-Cohen, Idit Diamant, Eyal Klang, Michal Amitai e Hayit Greenspan. Fully convolutional network for liver segmentation and lesions detection. Em *Deep Learning and Data Labeling for Medical Applications*, páginas 77–85. Springer, 2016. 10

[BEG+15]   Birol Baytan, Melike Sezgin Evim, Salih Güler, Adalet Meral Güneş e Mehmet Okan. Acute central nervous system complications in pediatric acute lymphoblastic leukemia. *Pediatric Neurology*, 53(4):312–318, 2015. 1

[BZ15]   David Bonekamp e Stefan L. Zimmerman. *Gibbs ringing artifact*, páginas 154–158. Cambridge University Press, 1 2015. 22

[BZS+15]   Spyridon Bakas, Ke Zeng, Aristeidis Sotiras, Saima Rathore, Hamed Akbari, Bilwaj Gaonkar, Martin Rozycki, Sarthak Pati e C Davazikos. Segmentation of gliomas in multimodal magnetic resonance imaging volumes based on a hybrid generative-discriminative framework. *Proceeding of the Multimodal Brain Tumor Image Segmentation Challenge*, páginas 5–12, 2015. 1

[CCB06]   Olivier Colliot, Oscar Camara e Isabelle Bloch. Integration of fuzzy spatial relations in deformable models—application to brain mri segmentation. *Pattern Recognition*, 39(8):1401–1414, 2006. 11, 12, 13, 23, 24

[CEE+16]   Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. Em *International Conference on Medical Image Computing and Computer-Assisted Intervention*, páginas 415–423. Springer, 2016. 10

[CJBBL05] Roberto M Cesar Jr, Endika Bengoetxea, Isabelle Bloch e Pedro Larrañaga. Inexact graph matching for model-based recognition: Evaluation and comparison of optimization algorithms. *Pattern Recognition*, 38(11):2099–2113, 2005. 17, 18

[CMK+13]   M Jorge Cardoso, Andrew Melbourne, Giles S Kendall, Marc Modat, Nicola J Robertson, Neil Marlow e Sebastien Ourselin. AdaPT: an adaptive preterm segmentation algorithm for neonatal brain mri. *NeuroImage*, 65:97–108, 2013. 10

[CR08]   Anthony G Cohn e Jochen Renz. Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence*, 3:551–596, 2008. 11

[DAGB15] Sonia Dahdouh, Elsa D Angelini, Gilles Grangé e Isabelle Bloch. Segmentation of embryonic and fetal 3d ultrasound images based on pixel intensity distributions and shape priors. *Medical Image Analysis*, 24(1):255–268, 2015. 10

[DB15]   Sonia Dahdouh e Isabelle Bloch. Shape-based multi-region segmentation framework: application to 3d infants mri data. Em *SPIE Medical Imaging*, páginas 941312–941312. International Society for Optics and Photonics, 2015. 10

[DCSA15] Chelli N Devi, Anupama Chandrasekharan, VK Sundararaman e Zachariah C Alex. Neonatal brain mri segmentation: A review. *Computers in biology and medicine*, 64:163–178, 2015. 1, 8, 9

[Dix84]   W Thomas Dixon. Simple proton spectroscopic imaging. *Radiology*, 153(1):189–194, 1984. 24

[DVO+16] S Dahdouh, N Varsier, MA Nunez Ochoa, J Wiart, A Peyman e I Bloch. Infants and young children modeling method for numerical dosimetry studies: application to plane wave exposure. *Physics in Medicine and Biology*, 61(4):1500, 2016. 10

[Ede99]   Shimon Edelman. *Representation and recognition in vision*. MIT press, 1999. 11

[FAB12]   Geoffroy Fouquier, Jamal Atif e Isabelle Bloch. Sequential model-based segmentation and recognition of image structures driven by visual features and spatial relations. *Computer Vision and Image Understanding*, 116(1):146–165, 2012. 12, 13

[FHHGM01] Lynn M Fletcher-Heath, Lawrence O Hall, Dmitry B Goldgof e F Reed Murtagh. Automatic segmentation of non-enhancing brain tumors in magnetic resonance images. *Artificial Intelligence in Medicine*, 21(1):43–63, 2001. 1

[FSB+02] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002. 1

[GCB14]   Ana B. V. Graciano, Roberto M. Cesar Jr. e Isabelle Bloch. Modeling and recognition of structured objects: a statistical-relational approach. Em *Stockholm. Proc. Workshop FEAST - ICPR 2014*. IEEE, 2014. 2, 23, 24

[GLFN+13] Daniel García-Lorenzo, Simon Francis, Sridar Narayanan, Douglas L Arnold e D Louis Collins. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical Image Analysis*, 17(1):1–18, 2013. 1

[GMJ+13] Daniel Rodriguez Gutierrez, Muftah Manita, Tim Jaspan, Robert A Dineen, Richard G Grundy e Dorothee P Auer. Serial mr diffusion to predict treatment response in high-grade pediatric brain tumors: a comparison of regional and voxel-based diffusion change metrics. *Neuro-oncology*, página not034, 2013. 1

[GPU+05] Jeroen JG Geurts, Petra JW Pouwels, Bernard MJ Uitdehaag, Chris H Polman, Frederik Barkhof e Jonas A Castelijns. Intracortical lesions in multiple sclerosis: improved detection with 3d double inversion-recovery mr imaging. *Radiology*, 236(1):254–260, 2005. 7

[GR17]   Hyun Woo Goo e Young-Shin Ra. Advanced mri for pediatric brain tumors with emphasis on clinical benefits. *Korean Journal of Radiology*, 18(1):194–207, 2017. 1

[Gra12]   Ana Beatriz Vicentim Graciano. *Modelagem e reconhecimento de objetos estruturados: uma abordagem estatístico-estrutural.* Tese de Doutorado, University of São Paulo, 2012. 2, 12, 13, 23, 29, 34

[GZM15]   Lars J Grimm, Jing Zhang e Maciej A Mazurowski. Computational approach to radiogenomics of breast cancer: luminal a and luminal b molecular subtypes are associated with imaging features on routine breast mri extracted using computer vision algorithms. *Journal of Magnetic Resonance Imaging*, 42(4):902–907, 2015. 1

[HAB08]   Céline Hudelot, Jamal Atif e Isabelle Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159(15):1929–1951, 2008. 11

[HAB17]   S Harish, GF Ali Ahammed e Reshma Banu. An extensive research survey on brain mri enhancement, segmentation and classification. Em *Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2017 International Conference on*, páginas 1–8. IEEE, 2017. 1

[HCH+05]   E Mark Haacke, Norman YC Cheng, Michael J House, Qiang Liu, Jaladhar Neelavalli, Robert J Ogg, Asadullah Khan, Muhammad Ayaz, Wolff Kirsch e Andre Obenaus. Imaging iron stores in the brain using magnetic resonance imaging. *Magnetic resonance imaging*, 23(1):1–25, 2005. 7

[HDWF+17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin e Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017. 10

[HLO+18]   Mathieu Hatt, Baptiste Laurent, Anouar Ouahabi, Hadi Fayad, Shan Tan, Laquan Li, Wei Lu, Vincent Jaouen, Clovis Tauber, Jakub Czakon et al. The first miccai challenge on pet tumor segmentation. *Medical image analysis*, 44:177–195, 2018. 5

[HVGS+09] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE transactions on medical imaging*, 28(8):1251–1265, 2009. 10

[HWK+98]   Petra S Hüppi, Simon Warfield, Ron Kikinis, Patrick D Barnes, Gary P Zientara, Ferenc A Jolesz, Miles K Tsuji e Joseph J Volpe. Quantitative magnetic resonance imaging of brain development in premature and mature newborns. *Annals of neurology*, 43(2):224–235, 1998. 1, 8

[HZM+16]   Abhilash Rakkunedeth Hareendranathan, Dornoosh Zonoobi, Myles Mabee, Chad Diederichs, Kumaradevan Punithakumar, Michelle Noga e Jacob L Jaremko. Hip segmentation from mri volumes in infants for ddh diagnosis and treatment planning. Em *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, páginas 1046–1049. IEEE, 2016. 10

[Jac01]   Clarence Martin Jackson. A method of teaching relational anatomy. *Journal of the American Medical Association*, 37(12):731–735, 1901. 11

[JKZ16]   Tushar H Jaware, KB Khanchandani e Anita Zurani. Multi-kernel support vector machine and levenberg-marquardt classification approach for neonatal brain mr images. Em *Power Electronics, Intelligent Control and Energy Systems (ICPEICES), IEEE International Conference on*, páginas 1–4. IEEE, 2016. 10

[KCAB09]   Hassan Khotanlou, Olivier Colliot, Jamal Atif e Isabelle Bloch. 3d brain tumor segmentation in mri using fuzzy classification, symmetry analysis and spatially constrained deformable models. *Fuzzy sets and systems*, 160(10):1457–1473, 2009. 12

[KFD+13]   Sun Hyung Kim, Vladimir S Fonov, Cheryl Dietrich, Clement Vachet, Heather C Haz-
           lett, Rachel G Smith, Michael M Graves, Joseph Piven, John H Gilmore, Stephen R
           Dager et al. Adaptive prior probability and spatial temporal intensity change esti-
           mation for segmentation of the one-year-old human brain. *Journal of Neuroscience
           Methods*, 212(1):43–55, 2013. 10

[KGSH90]   JP Kaltwasser, R Gottschalk, KP Schalk e W Hartl. Non-invasive quantitation of
           liver iron-overload by magnetic resonance imaging. *British Journal of Haematology*,
           74(3):360–363, 1990. 2

[KMAS+11] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Va-
           lentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Ruther-
           ford, A David Edwards et al. A dynamic 4d probabilistic atlas of the developing brain.
           NeuroImage, 54(4):2750–2763, 2011. 10

[LAFB16]   D. Lesage, E. D. Angelini, G. Funka-Lea e I. Bloch. Adaptive particle filtering for
           coronary artery segmentation from 3d ct angiograms. *Computer Vision and Image
           Understanding*, 151:29–46, 2016. 10

[LKB+17]   Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio,
           Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginne-
           ken e Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical
           image analysis*, 42:60–88, 2017. 5

[LWH+17]   Fang Lu, Fa Wu, Peijun Hu, Zhiyi Peng e Dexing Kong. Automatic 3d liver loca-
           tion and segmentation via convolutional neural network and graph cut. *International
           journal of computer assisted radiology and surgery*, 12(2):171–182, 2017. 10

[MAT+16]   B. Morel, G. Antoni, J.R. Teglas, I. Bloch e C. Adamsbaum. Neonatal brain MRI:
           How reliable is the radiologist's eye? *Neuroradiology*, 58:189–193, 2016. 10

[MBCJ16]   Henrique Morimitsu, Isabelle Bloch e Roberto M Cesar-Jr. Exploring structure for
           long-term tracking of multiple objects in sports videos. *Computer Vision and Image
           Understanding*, 2016. 24

[MEBT+16] Iska Moxon-Emre, Eric Bouffet, Michael D Taylor, Normand Laperriere, Michael B
           Sharpe, Suzanne Laughlin, Ute Bartels, Nadia Scantlebury, Nicole Law, David Malkin
           et al. Vulnerability of white matter to insult during childhood: evidence from patients
           treated for medulloblastoma. *Journal of Neurosurgery: Pediatrics*, 18(1):29–40, 2016.
           1

[MNA16]   Fausto Milletari, Nassir Navab e Seyed-Ahmad Ahmadi. V-net: Fully convolutional
           neural networks for volumetric medical image segmentation. Em *3D Vision (3DV),
           2016 Fourth International Conference on*, páginas 565–571. IEEE, 2016. 8, 10

[Moe16]   P Moeskops. *Automatic MRI-based quantication of brain characteristics in preterm
           newborns*. Tese de Doutorado, Utrecht University, 2016. 10

[MVM+16]   Pim Moeskops, Max A Viergever, Adriënne M Mendrik, Linda S de Vries, Manon JNL
           Benders e Ivana Išgum. Automatic segmentation of mr brain images with a convoluti-
           onal neural network. *IEEE Transactions on Medical Imaging*, 35(5):1252–1261, 2016.
           10

[MXV+16]   Baptiste Morel, Yongchao Xu, Alessio Virzi, Thierry Géraud, Catherine Adamsbaum
           e Isabelle Bloch. A challenging issue: Detection of white matter hyperintensities in
           neonatal brain mri. Em *Engineering in Medicine and Biology Society (EMBC), 2016
           IEEE 38th Annual International Conference of the*, páginas 93–96. IEEE, 2016. 12,
           13

[NAB13] Olivier Nempont, Jamal Atif e Isabelle Bloch. A constraint propagation approach to structural model based image segmentation and recognition. *Information Sciences*, 246:1–27, 2013. 13

[NP14]   Peer Neubert e Peter Protzel. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. Em *Pattern Recognition (ICPR), 2014 22nd International Conference on*, páginas 996–1001. IEEE, 2014. 26, 31

[NPD+13] Henning Neubauer, Thomas Pabst, Anke Dick, Wolfram Machann, Laura Evangelista, Clemens Wirth, Herbert Köstler, Dietbert Hahn e Meinrad Beer. Small-bowel mri in children and young adults with crohn disease: retrospective head-to-head comparison of contrast-enhanced and diffusion-weighted mri. *Pediatric radiology*, 43(1):103–114, 2013. 1, 8

[PKV+11] Tina Young Poussaint, Mehmet Kocak, Sridhar Vajapeyam, Roger I Packer, Richard L Robertson, Russell Geyer, Daphne Haas-Kogan, Ian F Pollack, Gilbert Vezina, Robert Zimmerman et al. Mri as a central component of clinical trials analysis in brainstem glioma: a report from the pediatric brain tumor consortium (pbtc). *Neuro-oncology*, 13(4):417–427, 2011. 1

[RGC+97] Wilburn E Reddick, John O Glass, Edwin N Cook, T David Elkin e Russell J Deaton. Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks. *IEEE Transactions on Medical Imaging*, 16(6):911–918, 1997. 1

[SAB+12] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J Counsell, James P Boardman, Mary A Rutherford, A David Edwards, Joseph V Hajnal e Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*, 59(3):2255–2265, 2012. 10

[SDM+04] Dieter Seghers, Emiliano D'Agostino, Frederik Maes, Dirk Vandermeulen e Paul Suetens. Construction of a brain template from mr images using state-of-the-art registration and segmentation techniques. Em *International Conference on Medical Image Computing and Computer-Assisted Intervention*, páginas 696–703. Springer, 2004. 10

[SFT+10] Feng Shi, Yong Fan, Songyuan Tang, John H Gilmore, Weili Lin e Dinggang Shen. Neonatal brain image segmentation in longitudinal mri studies. *NeuroImage*, 49(1):391–400, 2010. 10

[SLM+13] V Strenger, H Lackner, R Mayer, P Sminia, P Sovinz, M Mokry, A Pilhatsch, M Benesch, W Schwinger, M Seidel et al. Incidence and clinical course of radionecrosis in children with brain tumors. *Strahlentherapie und Onkologie*, 189(9):759–764, 2013. 1

[SRE12] Caroline A Schneider, Wayne S Rasband e Kevin W Eliceiri. Nih image to imagej: 25 years of image analysis. *Nature methods*, 9(7):671, 2012. 5

[SSLS+01] David W Shattuck, Stephanie R Sandor-Leahy, Kirt A Schaper, David A Rottenberg e Richard M Leahy. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, 13(5):856–876, 2001. 1

[Sue02] Paul Suetens. *Fundamentals of medical imaging.* Cambridge university press, 2002. 7

[USG+16] E Udayakumar, S Santhi, R Gowrishankar, C Ramesh e T Gowthaman. Region growing image segmentation for newborn brain mri. *BioTechnology: An Indian Journal*, 12(12), 2016. 10

[VGM+18] A Virzì, P Gori, CO Muller, E Mille, Q Peyrot, L Berteloot, N Boddaert, S Sarnacki e I Bloch. Segmentation of pelvic vessels in pediatric mri using a patch-based deep learning approach. Em *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*, páginas 97–106. Springer, 2018. 10

[Vie97] Laure Vieu. Spatial representation and reasoning in artificial intelligence. Em *Spatial and temporal reasoning*, páginas 5–41. Springer, 1997. 11

[WPC+15] Lijia Wang, Mengchao Pei, Noel CF Codella, Minisha Kochar, Jonathan W Weinsaft, Jianqi Li, Martin R Prince e Yi Wang. Left ventricle: fully automated segmentation based on spatiotemporal continuity and myocardium information in cine cardiac magnetic resonance imaging (lv-fast). *BioMed Research International*, 2015, 2015. 1

[WW09] Neil I Weisenfeld e Simon K Warfield. Automatic segmentation of newborn brain mri. *NeuroImage*, 47(2):564–572, 2009. 10

[WZL+16] Shuihua Wang, Yudong Zhang, Ge Liu, Preetha Phillips e Ti-Fei Yuan. Detection of alzheimer's disease by three-dimensional displacement field estimation in structural magnetic resonance imaging. *Journal of Alzheimer's Disease*, 50(1):233–248, 2016. 1

[XGB17] Yongchao XU, Thierry Geraud e Isabelle BLOCH. Segmentation d'irm de cerveaux de nouveau-nés en quelques secondesa l'aide d'un réseau de neurones convolutif pseudo-3d et de transfert d'apprentissage. Em *Actes du 26e Colloque GRETSI, Juan-les-Pins, France*, 2017. 10

[XSJ+07] Hui Xue, Latha Srinivasan, Shuzhou Jiang, Mary Rutherford, A David Edwards, Daniel Rueckert e Joseph V Hajnal. Automatic segmentation and reconstruction of the cortex from neonatal mri. *NeuroImage*, 38(3):461–477, 2007. 10

[ZBS01] Yongyue Zhang, Michael Brady e Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001. 1

[ZLD+15] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji e Dinggang Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214–224, 2015. 10

[ZSN+04] Reza A Zoroofi, Yoshinobu Sato, Takashi Nishii, Nobuhiko Sugano, Hideki Yoshikawa e Shinichi Tamura. Automated segmentation of necrotic femoral head from 3d mr data. *Computerized Medical Imaging and Graphics*, 28(5):267–278, 2004. 1