University of Alberta

Library Release Form

Name of Author: Mark Schmidt

Title of Thesis: Automatic Brain Tumor Segmentation

Degree: Master of Science

Year this Degree Granted: 2005

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author’s prior written permission.

____

Mark Schmidt
10623-108 St.
Westlock, AB
Canada, T7P 1E1

Date: ____________________

University of Alberta

AUTOMATIC BRAIN TUMOR SEGMENTATION

by

Mark Schmidt

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

in

Computing Science

Edmonton, Alberta
Fall 2005

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Automatic Brain Tumor Segmentation** submitted by Mark Schmidt in partial fulfillment of the requirements for the degree of **Master of Science**.

____

Russell Greiner

____

Peter Allen

____

Albert Murtha

____

Walter Bischof

Date: ____________________

Abstract

This thesis addresses the task of automatically segmenting brain tumors and edema in magnetic resonance images. This is motivated by potential applications in assessing tumor growth, assessing treatment responses, enhancing computer-assisted surgery, planning radiation therapy, and constructing tumor growth models. The presented framework forms an image processing *pipeline*, consisting of noise reduction, spatial registration, intensity standardization, feature extraction, pixel classification, and label relaxation. The key advantage of this framework is the simultaneous use of features computed from the image intensity properties, and the locations of pixels within an aligned template brain. Automatically learning to combine these features allows recognition of tumors and edema that have relatively normal intensity properties. Our results on 11 patients with brain tumors show that the system achieves nearly perfect performance given patient-specific training, but also achieves accurate results in segmenting patients not used in training.

Table of Contents

Chapter 1

Introduction

1.1 Chapter Summary

This introductory chapter will discuss the context for our task, the problem of automatic brain tumor segmentation, beginning with a definition of the problem in Section 1.2. This will be followed by a discussion of the practical applications that motivate the study of this problem in Section 1.3, while Section 1.4 will discuss the various challenging aspects of the problem that complicate the development of automatic methods. This chapter will conclude with a summary of the contribution of this dissertation, and will outline the remaining chapters in the work. Note that at the end of this document, there is a glossary of commonly used terms in this work.

1.2 Problem Definition

The problem addressed in this thesis is the automatic post-acquisition segmentation of brain tumors and associated edema in multi-spectral Magnetic Resonance (MR) images. Since there exist multiple interpretations of this problem, this section outlines the formal task definition that will be used.

Our input is a series of slices taken from different MR modalities of the same individual in the same session (Figure 1.1). The output will be a binary segmentation of the images, where each pixel in the input images is labeled as either *normal* or *abnormal*. As illustrated in Figure 1.2, there exist different possible definitions of abnormality. The segmentation of *enhancing pixels* and *enhancing tumor regions* have often been used in the literature as definitions of abnormality in order to simplify the development of automatic methods (in addition to the segmentation of homogeneous tumors). Although the approach presented in this work can be used to address these simplified cases, this work specifically addresses the more challenging tasks of segmenting the *Gross Tumor Volume* (GTV) and the full Tumor and Edema (swelling) area. Given these segmentations, the related tasks of computing the *GTV contour* or a *full brain segmentation* would be greatly simplified.

In each of the possible output cases, the desired output is defined manually by human experts based on the visible abnormality in the image data, which is limited by the imaging protocol used, and is subject to interpretation. As a result of this, the goal will not be defined as the determination of the absolute location of the abnormality, but to perform the segmentation as a human expert would.

In order for the algorithm to be practically useful for segmenting existing data, two major constraints are required. The first major constraint is that the processing will be post-acquisition. Specifically, only the image data will be used in order to produce the final result. The second major constraint is that the system must be able to utilize common MR modalities (such as T1-weighted and T2-weighted images [Brown and Semeka, 2003]). Having no control over the acquisition and employing only commonly used MR modalities complicates the task, since acquisition protocols vary considerably and may not necessarily facilitate a straightforward segmentation approach. However,

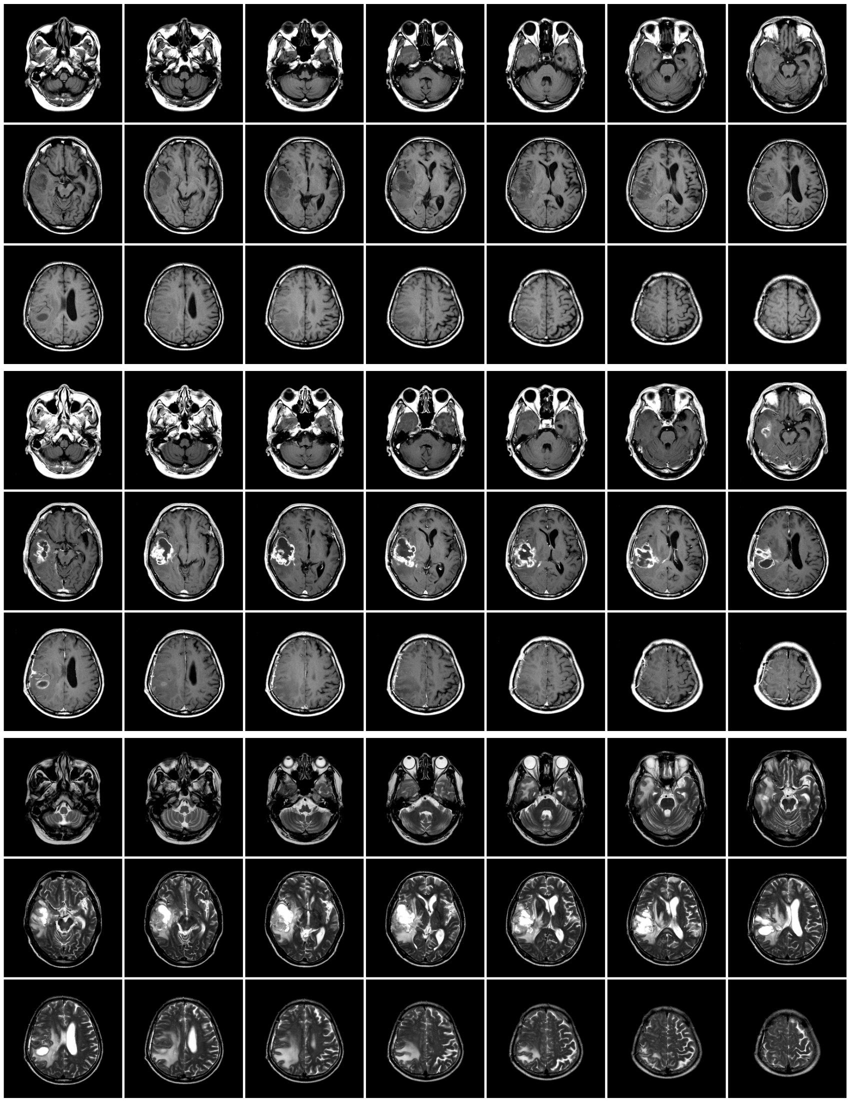


Figure 1.1: Multi-spectral MR Images (*axial* view). T1-weighted images (first 3 rows), T1-weighted images with contrast agent (middle 3 rows), T2-weighted images (bottom 3 rows). Slices within modalities are ordered from left to right, then top to bottom. This order corresponds to moving from the bottom to the top of the head.

these constraints are imposed in order for the technique to be practically applied to the segmentation of the large amounts of existing data, and in centers where more advanced protocols are not in place.

Since there may be various interpretations of what constitutes an ‘automatic’ method, we will now clarify this constraint. In the approach developed in this work, the algorithm receives as input the images and the names of the corresponding modalities (ie. axial T1-weighted image). An *automatic method* is defined as an approach that takes these modality-labeled images and produces a final segmentation without any human interaction. This excludes operations such as manual seed selection, manual contour initialization, manual prototype selection, manual contrast adjustment, manually restarting the algorithms if a human expert decides that a local minimum is found based on visual analysis, manual cluster selection, or other forms of manual input or adjustment. Although training data is employed in the presented system (and many of the experiments presented in Chapter 5 use patient-specific training), upon completion of the training phase, the system represents a fully automatic method that can segment images not incorporated within the training phase.

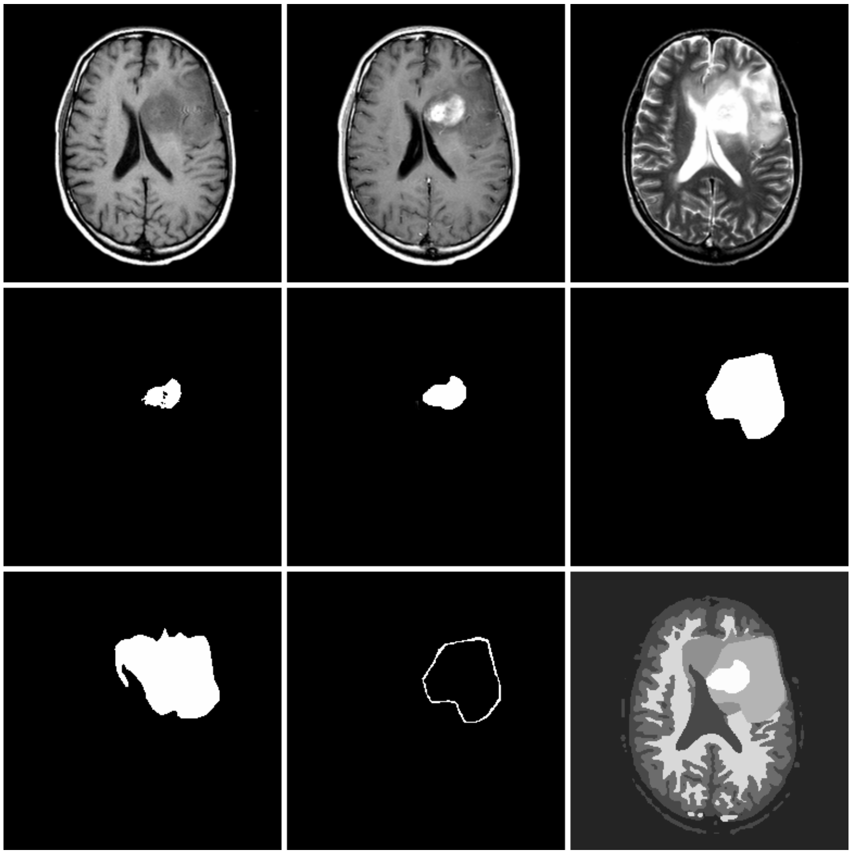


Figure 1.2: Different Tasks with Respect to Automatic Brain Tumor Segmentation. Top, input images, left to right: T1-weighted image, T1-weighted image with contrast agent, T2-weighted image (all of the same patient). Middle, tumor segmentation tasks, left to right: Enhancing pixel segmentation, Enhancing Area segmentation, Gross Tumor Volume (GTV) segmentation. Bottom, related tasks, left to right: Tumor and Edema area segmentation, GTV contour, full brain segmentation into 3 normal and 3 abnormal tissue classes.

1.3 Magnetic Resonance Imaging and Brain Tumors

Magnetic Resonance Imaging (MRI) is a powerful visualization technique that allows images of internal anatomy to be acquired in a safe and non-invasive way. It is based on the principles of Nuclear Magnetic Resonance (NMR), and allows a vast array of different types of visualizations to be performed. This imaging medium has been of particular relevance for producing images of the brain, due to the ability of MRI to record signals that can distinguish between different 'soft' tissues (such as gray matter and white matter) [Brown and Semeka, 2003]. In imaging the brain, two of the most commonly used MRI visualizations are T1-weighted and T2-weighted images. These weightings refer to the dominant signal (whether it be the T1 time or the T2 time) measured to produce the contrast observed in the image [Brown and Semeka, 2003]. Since areas with high fat content have a short T1 time relative to water, T1-weighted images can be thought of as visualizing locations of fat. In contrast, since areas with high water content have a short T2 time relative to areas of high fat content, T2-weighted images can be thought of as visualizing locations of water. Figure 1.3 demonstrates an example T1- and T2-weighted image, and the locations of two normal tissue types in these modalities.

In visualizing brain tumors, a second T1-weighted image is often acquired after the injection of a 'contrast agent'. These 'contrast agent' compounds usually contain an element whose composition causes a decrease in the T1 time of nearby tissue (gadolinium is one example)

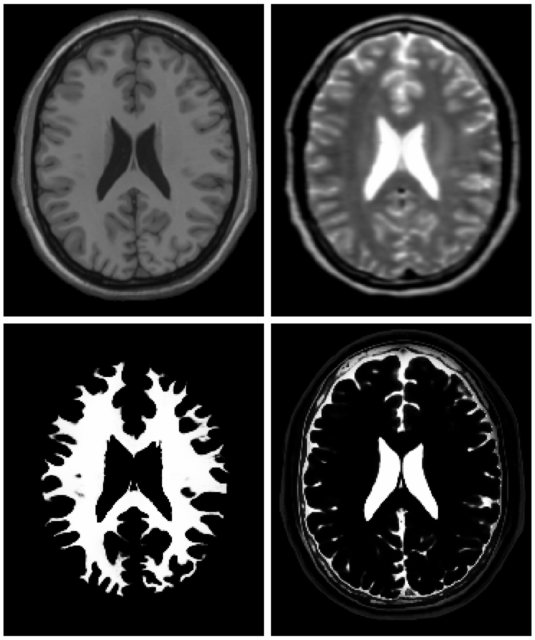


Figure 1.3: T1-weighted and T2-weighted signal properties. Top left: T1-weighted image (light regions visualize locations of fat). Top right: T2-weighted image (light regions visualize locations of water). Bottom left: White matter (high fat) locations. Bottom right: Cerebrospinal fluid (high water) locations. Images generated using the ICBM View software [ICBM View, Online], segmentations generated using Statistical Parametric Mapping [SPM, Online].

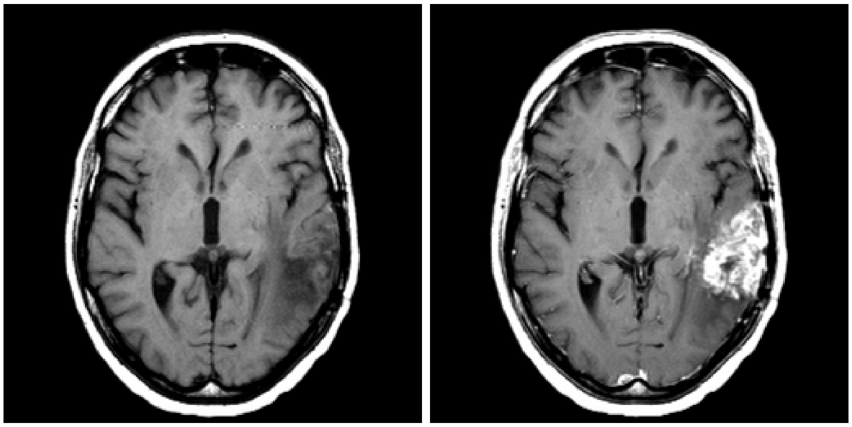


Figure 1.4: Effects of contrast agent on T1-weighted image data. Left: T1-weighted image prior to the injection of a contrast agent. Right: T1-weighted image after the injection of a contrast agent

[Brown and Semeka, 2003]. This results in bright regions observed at image locations that contain 'leaky' blood cells (where blood moves through the brain-blood barrier). The presence of this type of 'enhancing' area can indicate the presence of a tumor. Figure 1.4 illustrates a T1-weighted image before and after the injection of a contrast agent. Although the presence of this 'enhancement' can be a strong indicator of tumor location, there exist a large variety of types of brain tumors, and their appearance in MR images can vary considerably. Although some may be fully 'enhancing' (ie. appear hyper-intense after the injection of a contrast agent) or may have an 'enhancing' boundary, many types of tumors display only partial enhancement or no enhancement at all (such as those examined in [Fletcher-Heath et al., 2001]). Edema (swelling) can also be observed in many types of primary tumors, and appears as hyper-intense in T2 images. Treatment of primary brain tumors often involves a combination of surgical resection, radiation therapy, and chemotherapy [Murray, 2003]. MRI is used in tumor diagnosis, monitoring tumor progression, planning treatments, and monitoring responses to treatment.

1.4 Motivations for Automatic Segmentation

There are diverse motivations for the development of methods for automatic medical image segmentation. Accurate segmentations are needed or would be useful in clinical and scientific applications, but the need for *manual* intervention is both time consuming and subject to manual variation. This section will first examine applications of segmentation, and proceed to discuss the two drawbacks of manual segmentation. This section will conclude by exploring the properties of this problem that make it an excellent research challenge in the fields of Machine Learning and Pattern Recognition.

Many of the current and potential applications of segmentation are discussed in detail in Chapter 1 of [O'Donnell, 2001]. These include enhanced visualizations, high-throughput and consistent volume measurements, research into structural shape and variations, image-guided surgery, and change detection in images acquired at different times. With respect to brain tumors, change detection and volume measurements are often used to evaluate tumor growth or treatment response, but this problematic since current standard methods of tumor volume measurement consist of simple heuristics [Miller et al., 1981, Therasse et al., 2000], that are inaccurate compared to manual segmentations, and where only large changes can be deemed statistically significant. Change detection is also important with respect to evaluating the effectiveness of treatments, since tumors will have varied responses to different types of treatment. Change detection can be relevant over long periods or time, or can be used to detect small changes over short periods of time to assess the immediate patient- and tumor-specific effectiveness of different treatment methods.

Another motivation for pursuing automatic tumor segmentation methods is alleviating the manual work and reducing the variability associated with defining radiation therapy target areas. This is especially important with respect to new technologies such as radiosurgery and intensity-modulated radiation therapy that allow more precise treatment options than traditional technology [Pirzkall et al., 2001]. Accurate automatic segmentation methods could also lead to new applications, including effective content based image retrieval in large medical databases. This could allow clinicians to find similar images in historical data based on tumor location, grade, size, enhancement, extent of edema, similar patterns of growth, or a variety of other factors. This information could help clinicians in making decisions, in addition to being a useful research tool for exploring patterns in the historical data. In a similar vein, accurate high-throughput segmentations could be used in combination with relevant features and Machine Learning methods to improve tumor grading in cases where grading is ambivalent (or to discover potentially useful distinctions within grades), and to provide a more accurate and patient-specific prognosis.

Although manual segmentation by qualified professionals remains superior in quality to automatic methods, it has two drawbacks. The first drawback is that producing manual segmentations or semi-automatic segmentations is extremely time consuming, with higher accuracies on more finely detailed volumes demanding increased time from medical experts. It was estimated that the expected number of people performing manual segmentations at any time during an average day at the Surgical Planning Laboratory at Brigham and Women's Hospital is ten [O'Donnell, 2001]. Although this statistic indicates that segmentations are important in clinical settings, it also demonstrates that automatic methods that could achieve a sufficient level of accuracy would be highly desirable for their ability to perform high-throughput segmentation. The second problem with manual and semi-automatic segmentations is that the segmentation is subject to variations both between observers and within the same observer. For example, a recent study quantified an average of $28\% \pm 12\%$ variation in quantified volume between individuals performing the same brain tumor segmentation task (the variation ranged from 11% to 69%), and quantified a $20\% \pm 15\%$ variation within individuals repeating the task three times at 1 month intervals [Mazzara et al., 2004]. Accurate automatic methods would be advantageous since they are not subject to this variation and thus the significance of changes in volumes could be more easily assessed. In addition to tumor volume calculation, accurate automatic segmentation methods additionally have the potential to reduce the variability and increase the standardization of other measurements and protocols, including the quantification of edema or necrosis.

The study of automatic brain tumor segmentation represents an interesting research problem in Machine Learning and Pattern Recognition, since it represents a problem that humans can learn to do effectively, but developing highly accurate automatic methods remains a challenging problem. This is easily explained by the fact that humans must use high-level visual processing, and must incorporate specialized domain knowledge to perform this task [Prastawa et al., 2003], which makes developing fully automatic methods extremely challenging. Although this is true of many pattern recognition and vision problems, this problem has several properties that diminish the advantage that humans have over machines. This includes:

- The size of pixels is known, thus the ability to compensate for scale (using scene context) has no advantage.
- There is no temporal component and the brain remains stationary, therefore being able to visually track objects over time has no advantage.
- Humans view the data as a series of two-dimensional slices, therefore the ability of humans to use three-dimensional information in segmentation is also diminished in this task since there is no three-dimensional modeling of structures based on a large range of views of the object.
- There is no occlusion
- The viewpoint is known
- The behavior of different tissue types in different MR channels is well characterized [Just and Thelen, 1988]
- There are robust algorithms for correcting intensity inhomogeneity [Sled et al., 1999] (making the ability to compensate for differences in illumination less of an advantage).
- Finally, the head's appearance in MR images is relatively predictable, and the brain is well quantified structurally.

The latter information takes the form of atlases, templates, spatial prior probabilities for tissues, spatial prior probabilities for intensities, spatial prior probabilities for structures, and anatomic variability maps, that can be used, in theory, to offset the advantage humans have in incorporating domain knowledge to aid in this task (examples of these are shown in Figure 1.5).

1.5 Challenges in Automatic Segmentation

Despite the appealing properties listed above and the large amount of research focusing on brain tumor segmentation in MR images, robust and automatic methods that achieve an accuracy comparable to human experts have remained out of reach. This section will highlight many of the challenges associated with this problem that contribute to this disparity. The factors that need to be considered in performing quantitative analysis of MR images will first be discussed, distinguishing those that will and those that will not be addressed in this work. This will be followed by a discussion of the factors that further complicate the segmentation of brain tumors compared to normal tissues within the brain.

With respect to the MR imaging modality, this thesis will focus on five problems that can complicate the segmentation task:

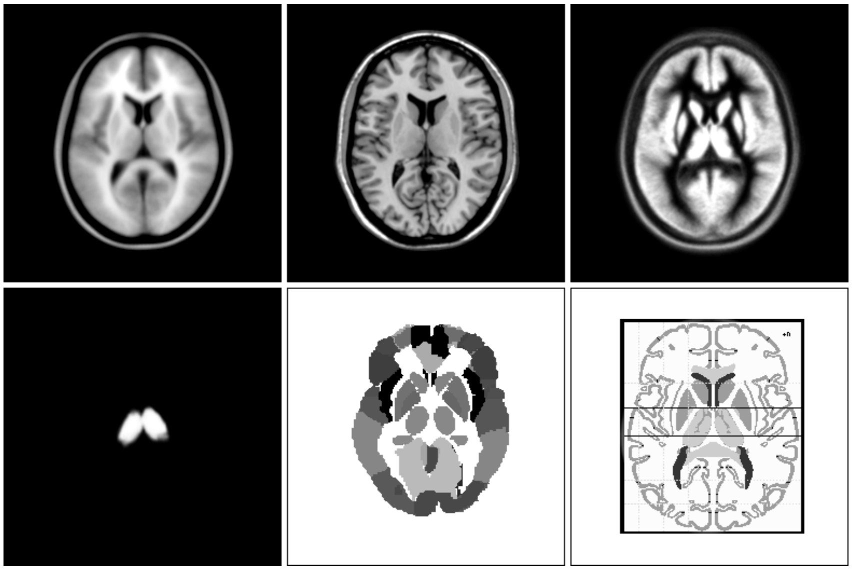


Figure 1.5: Examples of structural quantification of the brain: Top, left to right: Average T1-weighted image from 152 aligned normal adult brains, T1-weighted single subject template, spatial tissue probability map for gray matter [ICBM View, Online]. Bottom, left to right: spatial prior probability for thalamus [Mazziotta et al., 2001], anatomic atlas labels [Tzourio-Mazoyer et al., 2002], Talairach Daemon [TD, Online].

1. Local Noise
2. Partial Volume Averaging
3. Intensity Inhomogeneity
4. Inter-slice Intensity Variatoins
5. Intensity Non-Standardization

Local noise corrupts the signal measured for each pixel. A simulation of this effect is illustrated in Figure 1.6. The effect of this noise is often modeled as a Gaussian that is independent of the underlying tissue type [Sled et al., 1999] (although this is not strictly true as discussed in [Gering, 2003b]). *Partial volume averaging* is the result of the finite resolution represented by acquired pixels. Since the pixels have a finite size, an individual pixel can represent more than one type of tissue, resulting in *partial volume artifacts*. The intensity recorded for these partial volume artifacts will be a combination of the intensities of the structures that intersect at the pixel location. Figure 1.7 illustrates a simulation of the partial volume averaging effect using pixels of different sizes along the dimension orthogonal to the slice. *Intensity inhomogeneity* refers to variations in the recorded intensity observed within a set of slices, that can lead to a 10% to 20% variation in the intensity values recorded for homogeneous tissues [Sled et al., 1999]. This effect is illustrated in Figure 1.8, and is the result of a variety of factors related to the acquisition environment (such as the strength of the magnet and the type of receiver coil used), and to patient specific effects (such as attenuation of the radiofrequency signal as it passes through different tissue types). Discussions of the various causes of this inhomogeneity are presented in [Brown and Semeka, 2003, Sled, 1997], while a study on the effects of many of these factors on images generated with a 1.5 Tesla magnet can be found in [Simmons et al., 1994]. *Inter-slice intensity variations* are a specific type of intensity inhomogeneity that refers to rapid changes in the intensities of adjacent slices caused by gradient eddy currents and cross-talk between the slices in multi-slice acquisition protocols [Leemput et al., 1999a]. As quantified in [Simmons et al., 1994], this can result in even-numbered slices being noticeably darker than odd-numbered slices or vice versa (Figure 1.9).

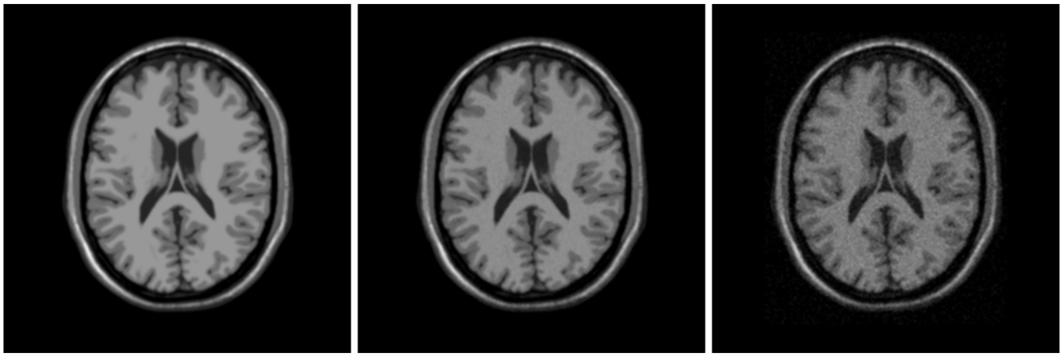


Figure 1.6: Local noise simulated using BrainWeb [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998]. Left to right: Noise free image, image with 3% noise, image with 9% noise.

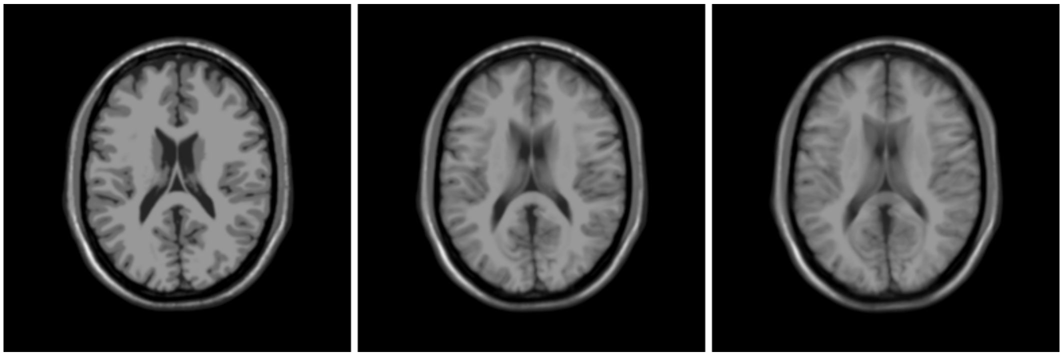


Figure 1.7: Partial volume averaging simulated using BrainWeb [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998]. The 'in-plane' size of pixels is kept constant while the size orthogonal to this plane is increased. Left to right: Image with 1mm pixels, image with 5mm pixel thickness, image with 9mm pixel thickness

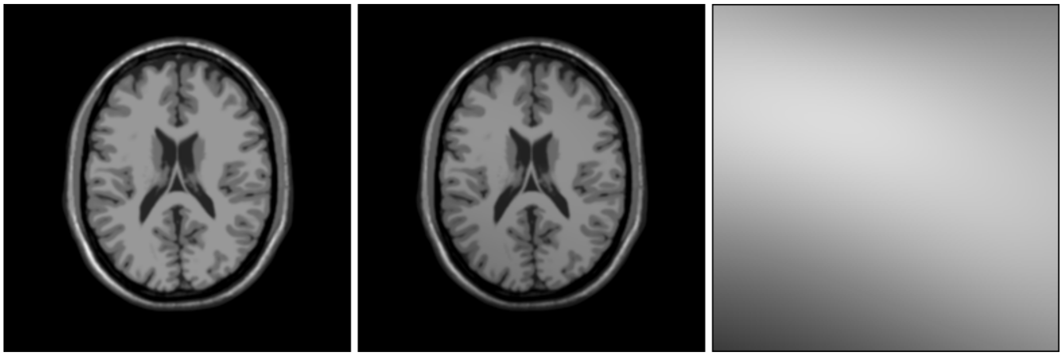


Figure 1.8: Intensity inhomogeneity simulated using BrainWeb [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998]. Left to right: Original image, image with 40% intensity inhomogeneity, applied inhomogeneity field

The final challenge in segmenting MR images that will be addressed in this thesis is the issue of *intensity non-standardization*. The versatility of MR imaging has led to the existence of a large variety of protocols for generating images with similar visual properties. The acquisition of MR images is therefore not a calibrated measure, and the intensities represented in the image do not have an exact meaning with respect to the underlying tissue [Clatz et al., 2004]. This variation can

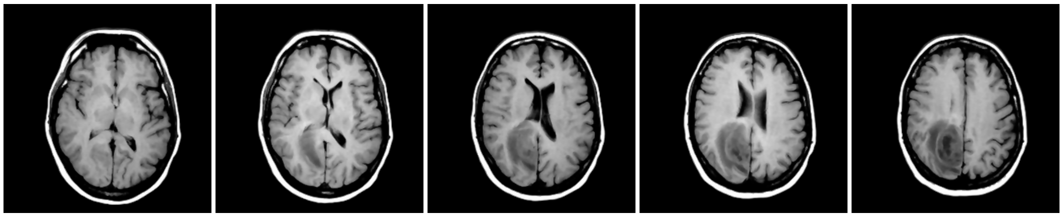


Figure 1.9: Inter-slice intensity variations following an 'even-odd' pattern. Five consecutive slices from a T1-weighted series are shown. Observe that the even numbered slices are brighter than the odd numbered slices.

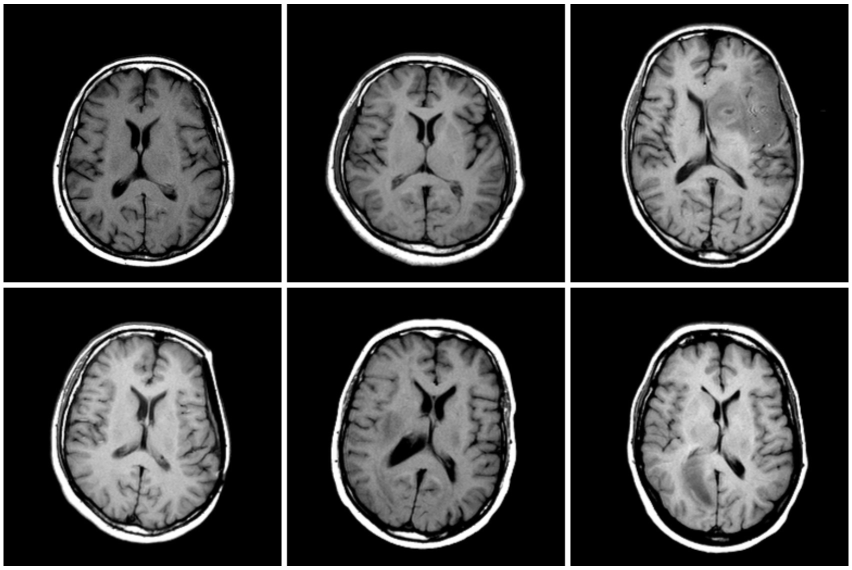


Figure 1.10: Intensity Non-standardization in a controlled situation. 6 slices from different patients at approximately the same area of the different heads, acquired using the same scanner and protocol, show large intensity variations.

cause major problems in intensity based segmentation methods, since differences in a wide variety of factors can lead to different observed intensity distributions. These intensity differences are present even in very controlled settings:

> "Unlike in other modalities, such as X-ray computerized tomography, MR images taken for the same patient on the same scanner at different times may appear different from each other due to a variety of scanner-dependent variations and, therefore, the absolute intensity values do not have a fixed meaning." [Nyul et al., 2000]

We can generally compensate for local noise, partial volume averaging, intensity inhomogeneity, and inter-slice intensity variations by preprocessing or postprocessing steps. However, intensity non-standardization represents a major problem in the quantitative analysis of MR images. Figure 1.10 illustrates the intensity differences between several T1-weighted images acquired using the same scanner and protocol (note that the differences observed between images acquired with different scanners and protocols will be much more significant).

There are several other factors that can make segmentation and quantitative analysis of MR images challenging. These include geometric distortions, inter-slice gaps, anisotropic pixels, the 'Gibbs-ringing' effect [Gering, 2003b], and finally misalignment within image series and motion artifacts due to patient movement. This dissertation will not specifically focus on these issues, as most of these issues do not interfere significantly with the segmentation task (with the exception of

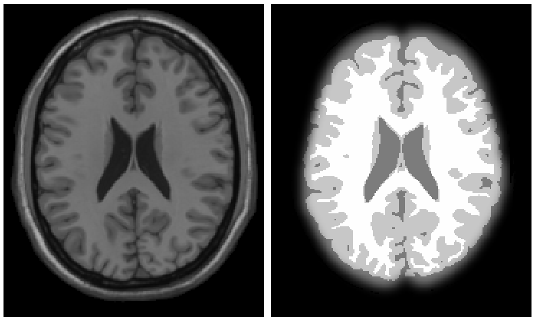


Figure 1.11: Normal brain segmentation. Left: T1-weighted image. Right: Automatic segmentation into the three normal tissue classes.

motion artifacts), and methods to reduce many factors such as these are often incorporated during acquisition and thus are not typically considered in performing post-acquisition image analysis.

Despite the presence of the above five challenges for automatic segmentation of MR images, effective methods for the automatic segmentation of normal brains into different normal brain tissue classes exist (see Figure 1.11 for an example input and output). This will be discussed further in the next chapter, but the main insight underlying these methods is that the locations of different tissues within the brain are relatively predictable (at the level of regions, not necessarily voxels), and thus spatial information can be used to improve an intensity-based model, resulting in an accurate and automatic segmentation. There are several factors that complicate the direct application of these types of algorithms to the segmentation of brain tumors, but the two major factors are the difficulty in predicting the tumor's spatial location *a priori*, and the potentially complex tumor intensity distributions that often violate simple assumptions. With respect to tumor intensity distributions, the most severe complicating factor is that tumor pixels can have similar or identical signals to normal pixels [Just and Thelen, 1988], even within the same image. This complicates intensity-based methods since different outputs will be desired for the same input intensities. Another challenge associated with using intensities for brain tumor segmentation is that tumor areas often have heterogeneous intensities (and do not follow the simple parametric distributions that can model normal tissues), and even homogeneous tumors can appear in different areas of the intensity spectrum. Another property of brain tumors that complicates their segmentation is that they can displace, infiltrate, and destroy nearby normal tissue [Price et al., 2004]. One result of this is that regions of mis-registration after the alignment of a normal brain are not necessarily pathological, since they may appear abnormal on the image due to pressure from the displacement. Other challenges in brain tumor segmentation from conventional MR images include the presence of ambiguous boundaries due to the lack of a clear contrast at the boundaries, and that regions of abnormality must be distinguished from regions of normal variation (or variations in normal areas observed due to the presence of the abnormality). These various additional challenges with respect to MR imaging and brain tumor segmentation make simple intensity models inadequate for accurate automatic brain tumor segmentation.

1.6 Thesis Contribution and Outline

This section has introduced the problem of automatic brain tumor segmentation in Magnetic Resonance images. Magnetic Resonance is an excellent modality for visualizing brain tumors, and there exist a large variety of current and potential applications for brain tumor segmentation in this modality. Intensity-based segmentation in MR images is complicated by local noise, partial volume averaging, intensity inhomogeneity, inter-slice intensity variations, and intensity non-standardization. Brain tumor segmentation in MR images is further complicated by the lack of *a priori* knowledge about tumor location and the tumor's intensity distribution, in addition to the intensity overlap observed between normal and tumor tissues, the potential intensity heterogeneity of the tumor region,

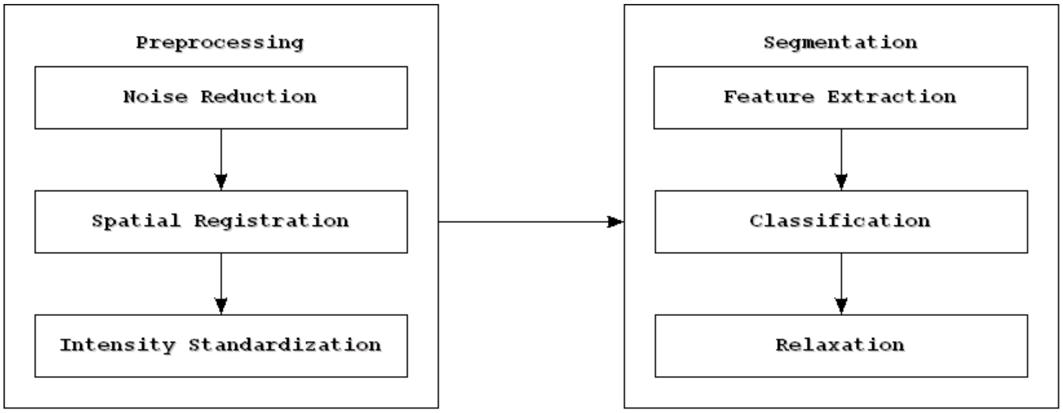


Figure 1.12: High Level Outline of the Presented Framework

the potential lack of contrast between tumors and adjacent normal tissue, and finally the challenge of distinguishing tumor or edema regions from displaced normal tissue and normal (or tumor induced) anatomic variations.

The many challenges associated with this problem complicate the development of automatic methods to perform this task. However, despite these challenges, humans experts can learn to perform this pattern recognition task effectively. This is primarily due to the fact that humans are able to fuse information from a variety of sources including multiple imaging modalities (T1 and T2 images), anatomic knowledge about shape and relative positions of structures, bi-lateral symmetry, the relative image intensities of different tissue types within an image, patient-specific diagnostic information, and the characterization of image regions and shapes (including analysis of textures and the extrapolation of boundaries between different structures where boundaries may be ambiguous).

As will be detailed in the next chapter, there have been various approaches proposed for automatic brain tumor segmentation in MR images, many of which rely on an intensity-based classification scheme. As will be discussed, although additional image-based features (such as textures, for example) can significantly improve results relative to intensity-based classification, large training sets are often required in order to achieve accurate results, since the training set must be large enough to sufficiently characterize the different tissue classes based on the larger feature sets. The next chapter will also outline several recent alternate approaches to improving intensity based classification methods, that use the spatial alignment of a template with known properties in order to improve an intensity-based classification.

The recent methods that encode spatial information to improve intensity-based classifications have demonstrated impressive results, even in circumstances that use small training sets or are fully unsupervised. However, the ability of these systems to address the most challenging cases remains bottlenecked by a reliance on a primarily intensity-based classification as the major component of the system that performs the *recognition* task (locating the general abnormality). This problem exists even in systems that use advanced anatomy based pre- or post-processing steps, and in systems that use neighborhood or shape based post-processing steps. This bottleneck remains due to the fact that it is not obvious how to optimally and simultaneously incorporate diverse forms of prior knowledge (such as bi-lateral symmetry or similarity to a normal brain) into a system for tumor segmentation, even though it is obvious that these forms of prior knowledge are important.

The contribution of this dissertation is a fully automatic framework for brain tumor segmentation in MR images that alleviates this bottleneck in the segmentation process, by augmenting an intensity-based classification model with features that encode diverse forms of prior knowledge, obtained after the spatial registration of the image with a template in a standard coordinate system. The combination of image-based and prior knowledge-based features allow the classification phase to overcome many image-based problems such as intensity overlap, intensity heterogeneity, and the lack of contrast at structure boundaries. The use of knowledge-based features also allows accurate results to be obtained with a relatively small number of training images. These features are incorporated into a fully automatic segmentation framework, that notably also contains preprocessing steps that reduce the problems associated with using the MR image intensities directly. Thus, the thesis underlying this dissertation is that accurate automatic brain tumor segmentation can be performed by learning to combine different sources of information, including the (standardized) regional intensity information and features derived from the spatial alignment of a template image in a standard coordinate system.

The implementation of the segmentation framework was tested for the segmentation of primary brain tumors and associated edema in T1-weighted and T2-weighted images. However, the framework was designed to be applicable to other segmentation tasks or with additional MR (or non-MR) modalities, or could potentially use entirely different sets of modalities (although this will not be demonstrated). Another advantage of this framework is its modular design, illustrated in Figure 1.12. Dividing the task into a series of steps that do not produce intermediate segmentations will allow new algorithms to replace existing elements of the presented implementation of the framework to improve the quality of the system's output.

Chapter 1 has introduced the problem of automatic brain tumor segmentation in MR images, along with applications, challenges, and a high-level description of the contribution of this thesis. The remaining chapters are organized as follows:

- Chapter 2 will extensively survey previous approaches proposed for this problem.
- Chapter 3 will present the automatic segmentation framework, and highlight the purpose of each step in the process.
- Chapter 4 will present in detail the implemented instantiation of this framework, including motivations for design decisions and potential improvements for each step.
- Chapter 5 will presents experimental results evaluating the implemented instantiation.
- Chapter 6 will present a summary of this work, and discuss potential future directions of research.

Chapter 2

Existing Approaches to Automatic Brain Tumor Segmentation

There is an immense array of scientific literature focusing on the task of image segmentation. Medical image segmentation has also received significant attention, due to the many practical applications of segmentation results. An impressively large amount of research effort has even focused on specific areas of the body or specific modalities, such as the segmentation of images of the brain in MR images. Although this section will not cover in detail all of the approaches proposed for the segmentation of MR images of the brain, it will provide a survey of many of the proposed approaches for automatic brain tumor segmentation in MR images. The focus of this section may therefore seem limited in scope; however there has been a large amount of research effort directed towards this problem and many of the approaches that will be discussed here represent prototypical examples of state of the art methods in the general area of medical image segmentation.

The remainder of this chapter will be divided into sections based on general properties of the systems. The first two types of methods we examine are *unsupervised* and *supervised* methods that do not incorporate spatial registration. The difference between these two is that *supervised* methods make use of *training data* that has been manually labeled, while *unsupervised* methods do not. The third set of methods that we discuss are the recent methods that incorporate *spatial registration*, while the final class of methods covered will be the recent methods that additionally incorporate *spatial prior probabilities*.

2.1 Unsupervised Segmentation

Image segmentation is the task of dividing an image into homogeneous regions. This requires an *objective* measure that is used to define homogeneity. The image segmentation task addressed in this thesis uses an *anatomic* objective measure to assess segmentation quality, in contrast to methods that use an *image-based* objective measure. This refers to the fact that the goal is to segment the image into regions that have homogeneous (and known) anatomic properties, rather than regions that have similar intensities or textures. Section 2.1.2 will discuss several methods proposed for brain tumor segmentation that use an *image-based* objective measure, but Section 2.1 will primarily focus on unsupervised approaches that aim to segment the image into at least two anatomically meaningful regions, at least one of which is *tumor* or *edema*.

2.1.1 Unsupervised Segmentation with an Anatomic Objective Measure

[Gibbs et al., 1996] presented an unsupervised approach for the segmentation of enhancing tumor pixels from T1-weighted post-contrast images. This system first applied an intensity threshold to a manually selected region of interest, then used a region growing algorithm to expand the thresholded

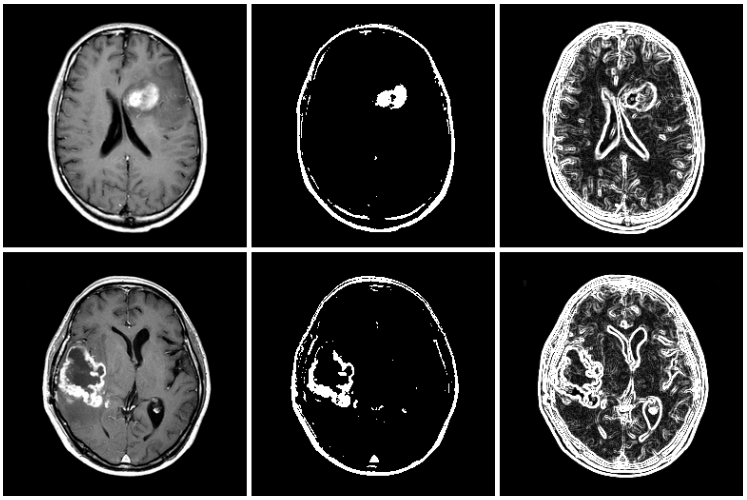


Figure 2.1: Examples of low-level image processing in segmentation of enhancing tumors. Top, left to right: T1-weighted post-contrast image, image after intensity thresholding, edge probabilities resulting from a Sobel filter. Bottom: the same image and operations applied to a different patient. A large amount of false positives are associated with normal structures in the thresholded image, and false negatives are associated with regions that are not sufficiently hyper-intense.

regions up to the edges defined by a Sobel edge detection filter. Figure 2.1 demonstrates intensity thresholding and Sobel edge detection results. The region growing result was refined through iterations of *dilation* (causing the defined tumor region to grow), and *erosion* (conversely causing the defined tumor region to shrink). These two operations change the labels assigned to individual pixels by examining the labels of neighboring pixels, and are commonly referred to as *morphological operations*. A similar approach was proposed in [Zhu and Yan, 1997] for the segmentation element of their enhancing tumor boundary detection approach.

These methods represent a clearly justified approach for segmenting image objects that are different in intensity than their surroundings. Although the requirement of manual slice or region of interest selection is one disadvantage of these methods, a more severe disadvantage is that these methods do not effectively take into account the presence of hyper-intense pixels representing normal structures in T1 post-contrast images. These false positives include non-tumor structures that have short T1 times (locations of fat), in addition to normal structures that may also uptake the contrast agent. Another major disadvantage of these methods is the assumption that the entire boundary will have a large intensity difference between its surrounding tissues, which is not always the case.

[Ho et al., 2002] presented a more recent fully unsupervised approach for tumor segmentation. This approach also focused on segmenting tumors with an enhancing border, but was not subject to many of the disadvantages of the approaches discussed above. This system used both the T1-weighted pre-contrast and the T1-weighted post-contrast images as input, and the first step in this system was the *coregistration* of these two volumes. Coregistration refers to the spatial alignment of two volumes that may not be of the same modality, but that represent a (potentially unaligned) measurement of the same underlying object. After this alignment step, an image was computed that represented the difference between the T1-weighted images before and after the injection of the contrast agent (Figure 2.2). A *Mixture Model* was then applied to the histogram of this difference image. The parametric distribution fit to the pixels that had a large difference was used to initialize a *Level Set* active contour, that 'evolves' to find a final, smooth 'blob-like' three-dimensional segmentation (that was post-processed to remove connected regions below a size threshold).

The [Ho et al., 2002] method has clear advantages over the methods discussed earlier. The use of a Mixture Model allows the technique to *adaptively* find the enhancing area, and is thus more robust to differences in intensity between images due to intensity non-standardization. Another advantage

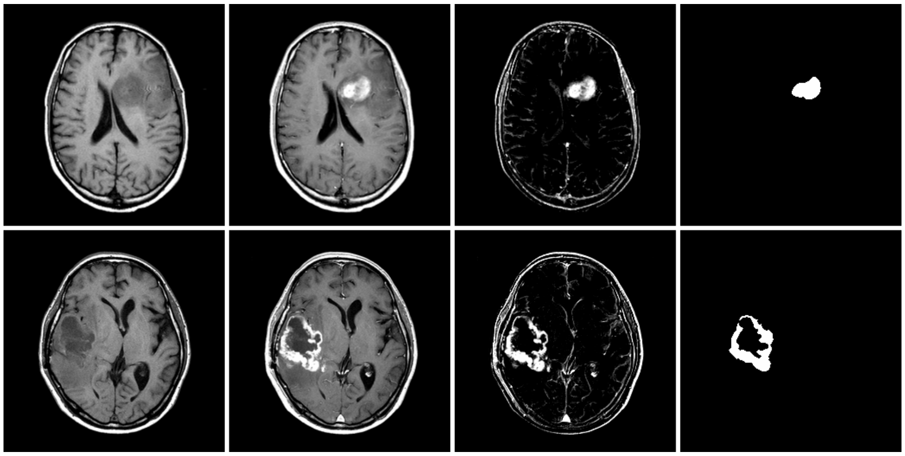


Figure 2.2: Example of the contrast agent difference image and a Level Set active contour applied to this image. Left to right: Original T1-weighted image, coregistered T1-weighted image after contrast injection, pixel-wise difference between the pre-injection and post-injection images, segmentation results with a Level Set active contour with manual initialization and manual parameter tuning. Bottom: the same images and operation applied to a different patient. The difference image reduces false positives compared thresholding, but false positives remain due to the presence of the normal contrast enhancing structures. The contour provides a very accurate delineation of the enhancing area in the top image, while not producing as accurate of a segmentation in the bottom image. In addition to a large hole in the middle of the segmentation of the lower image, there is a gap in the tumor boundary in the upper left area of the segmentation.

of this system is that non-enhancing areas surrounded by enhancing areas will be included in the segmentation through the use of the active contour. The use of the difference information rather than the post-contrast image directly is also advantageous, since it allows false positives associated with the many structures that are normally hyper-intense in T1 images to be removed (important if the enhancing region is adjacent to a hyper-intense normal structure). Although this has the potential to remove a significant amount of false positives, there may still be systematic false positives associated with this method, since it does not account for normal structures that are also affected by the contrast agent. Another disadvantage of this system is the large number of sensitive parameters that must be set for the Level Set method to converge to, and terminate at, an appropriate solution.

[Clark et al., 1998] presented a different type of approach for unsupervised automatic tumor segmentation, and is one of the most extensively validated system to date. This work focused on the segmentation of enhancing pixels from T1-weighted (post-contrast), T2-weighted, and ρ-weighted images (an additional MR modality that is often acquired simultaneously with T2-weighted images). The two main components of this system are *Fuzzy C-Means* (FCM) clustering (Figure 2.3), and a linear sequence of human-enginnered *knowledge-based* rules and operations. The clustering algorithm divides the pixels into groups with similar multi-spectral intensities (an unsupervised image segmentation with an image-based objective function), while the knowledge-based rules are a set of (intensity and anatomy based) rules and low-level image processing operations designed to select and process the results of the clustering algorithm in order to achieve a final segmentation. This system proceeds by clustering the data, applying rules to remove certain clusters or process others, then re-clustering and applying more rules on the reduced and refined segmentation. The system proceeds from clustering the entire image to clustering very specific areas, since the rules allow the identification of clusters that do not have tumor properties. Examples of these rules include that cerebrospinal fluid (CSF) will be the cluster within the brain that has the lowest T1 value, that pathological pixels will be assigned to the 3 highest intensity ρ-weighted clusters, and that clusters with tumor pixels will be closer to the highest T1 cluster than the lowest. The image processing 'rules' include morphological operations such as erosion and closing, in addition to cluster evaluation tech-

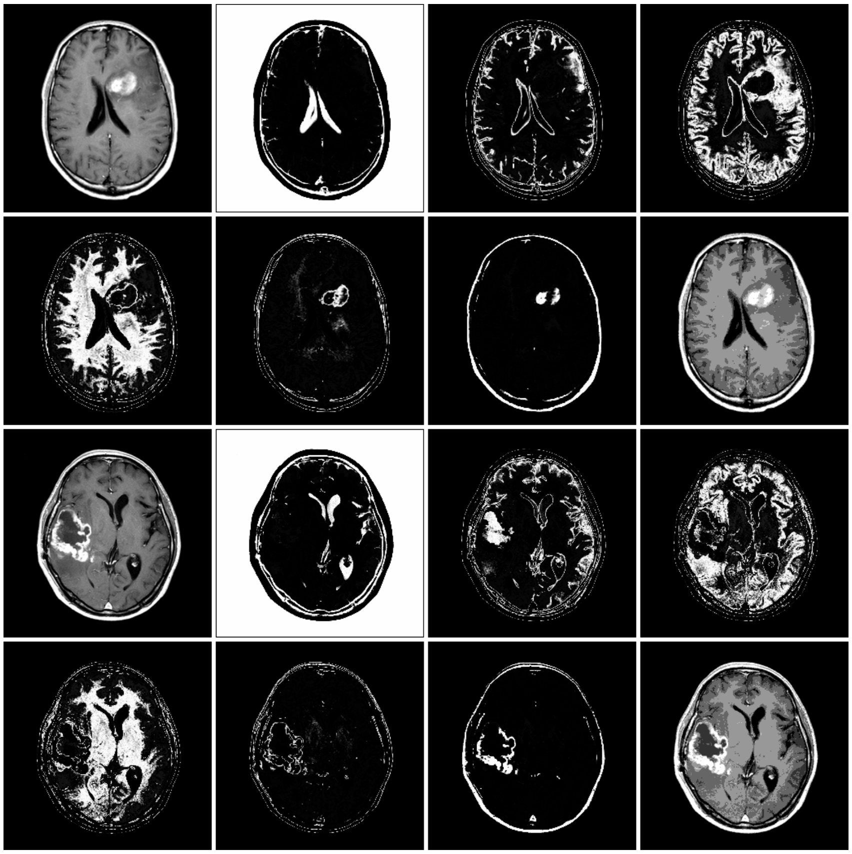


Figure 2.3: Example Fuzzy C-Means Clustering into 6 clusters. Top, left to right: T1-weighted image after contrast injection, first three clusters. Second row: left to right: last three clusters, image visualizing all 6 clusters. Bottom rows: The same image and operations applied to a different patient. This represents an unsupervised segmentation with an image-based objective measure; note that the tumor has been divided among multiple classes and utilizing the intensity data has produced noisy results, motivating the need for significant post-processing in order to reach a final segmentation.

niques such as cluster density thresholding. Note that these rules are not learned automatically from the data, but rather are manually engineered by the system's designer.

This *knowledge-based* approach to brain tumor segmentation results in a final system that is fairly intuitive. Iteratively proceeding from examining the entire image down to specific regions of interest is a logical approach, and most of the rules are well-motivated and based on tissue properties or anatomic structures (though some of the rules are based on ideas from image processing or clever observations). An obvious advantage of this system is that it contains rules that account for normal structures that also appear hyper-intense due to the injection of the contrast agent.

Criticisms of this type of approach include that the rules may not be robust to intensity non-standardization and that errors can propagate if the assumptions of early rules in the sequence are violated. However, the main criticism of this type of *knowledge-based* approach is that it requires considerable manual engineering. This is primarily due to the difficulty involved in translating complex anatomic knowledge and visual analysis into sequential low-level operations and rules. Even for the simplest definition of tumor segmentation (labeling enhancing pixels), the final system requires a large amount of rules and manual data analysis. The required manual data engineering makes this type of approach difficult to apply in cases where tumor tissue closely resembles normal tissue, does not have a clearly defined boundary, or is heterogeneous. Additionally, a system following this approach would need to be completely re-engineered in order to use a different set of modalities. Nevertheless, this type of approach has been employed in various works, including

[Yoon et al., 1999] and [Gosche et al., 1999]. More recent systems based on the use of FCM and knowledge-based rules include [Fletcher-Heath et al., 2001], which focused on the segmentation of non-enhancing tumors, and [Shen et al., 2003], which incorporated *intensity standardization* as a preprocessing step, and utilized a modified FCM algorithm that incorporated dependencies between neighboring pixels.

2.1.2 Unsupervised Segmentation with an Image-Based Objective Measure

There has been substantial research effort directed towards techniques for unsupervised brain tumor segmentation in MR images that do not use an anatomic objective measure. Rather than dividing the image along anatomically meaningful distinctions, these methods divide images into homogeneous regions using image-based features such as intensities and/or textures (clustering is one method to do this). These methods will not be covered in great detail, since there are major disadvantages to this type of approach. These include the facts that (1) the number of regions often needs to be pre-specified, (2) tumors can be divided into multiple regions, and (3) tumors may not have clearly defined intensity or textural boundaries. These factors are especially evident when considering heterogeneous tumors, since these factors necessitate manual intervention (or rule-based systems as before) in order to identify (and possibly split, merge, or process) the tumor regions that are to be used for quantitative analysis. Although not directly applicable for quantitative analysis, these techniques are often appropriate for enhanced visualizations. This includes, for example, producing a visualization that highlights the different regions present in a heterogeneous tumor.

[Sammouda et al., 1996] examined three methods to perform unsupervised brain tumor segmentation with an image-based objective measure: Hopfield Neural Networks, Boltzmann Machines, and the ISODATA algorithm. [Capelle et al., 2000] presented a more recent version of this type of approach. This method has advantages over similar methods due to the use of an automatic 'brain masking' preprocessing operation (removing pixels from the analysis that are not part of the brain area, alternately referred to as 'skull stripping' in the literature), and the use of a Markov Random Field model that statistically uses influences that neighboring pixels should have on each other's labels, removing the need for morphological operations. This work assumed that the tissue classes (gray matter, white matter, CSF, tumor, and edema) could be modeled by a Mixture Model (of Gaussians), and trained the Markov Random Field with the Iterated Condition Modes (ICM) algorithm. More recently, [Capelle et al., 2004] presented another approach of this nature, that also used 'brain masking' and a Gaussian Mixture Model (learned using an Expectation Maximization approach), but used an Evidence Theory formulation rather than a Markov Random Field to take into account neighboring pixel dependencies.

2.1.3 Summary of Unsupervised Segmentation

Although unsupervised segmentation methods that use an anatomic objective measure would be preferred over supervised methods since they avoid the human variability associated with manual training data is avoided, they have thus been of limited applicability. Most proposed methods of this type have focused solely on the segmentation of enhancing tumor areas, a greatly simplified task compared to the segmentation of edema or non-enhancing tumors. This is primarily due to the difficulty in translating the visual processing and anatomic knowledge used by human experts into operations that yield the desired results. Unsupervised segmentation methods that use an intensity or texture based objective measure can handle more complicated cases and are useful in enhancing visualizations, but the results are often not appropriate for automatic quantitative analysis since intensity and texture distinctions often do not correspond to the *appropriate* anatomic distinctions.

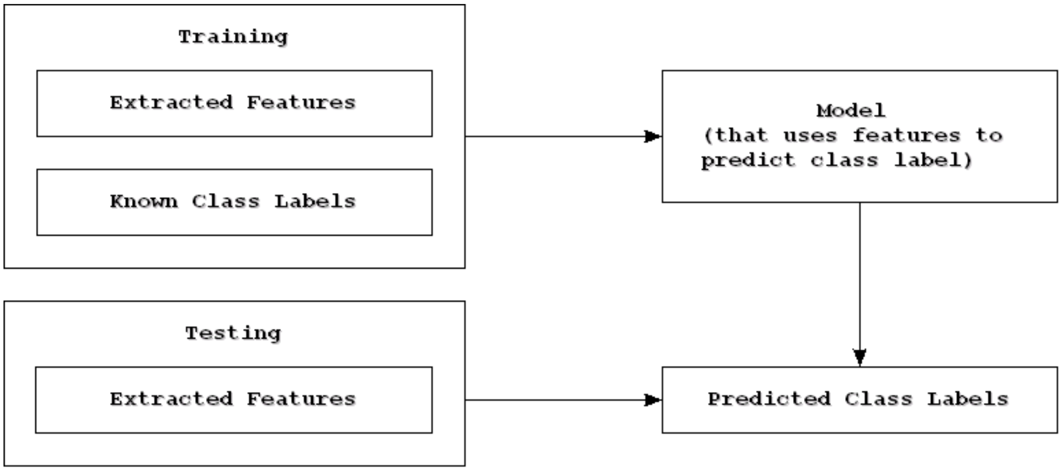


Figure 2.4: Overview of supervised learning framework. The *training* phase uses labeled data and extracted features to generate a model mapping from the values of the features to the labels. The *testing* phase uses this model to predict labels from extracted features where the label is not known.

2.2 Supervised Segmentation

Supervised methods for image segmentation differ from unsupervised methods through the use of labeled training data, used to automatically *learn* a model for segmentation. The advantage that data-driven approaches such as supervised methods offer is that relevant patterns in the data are discovered automatically, rather than through manual experimentation and intuition. The *classification* problem formulation is a popular method to perform image segmentation using a supervised approach. The task in classification is to assign a class, from a finite set of classes, to an entity based on a set of features. Supervised classification involves both a *training* phase that uses labeled data to learn a model that maps from features to labels, and a *testing* phase that is used to assign labels to unlabeled data based on the measured features (Figure 2.4). While many unsupervised approaches also use these 2 phases, the use of labeled data in the training phase of supervised approaches forces the model to focus on making discriminations in the feature space that correspond to the desired semantic discriminations.

One straightforward method of formulating the brain tumor segmentation task as a supervised classification problem is to use the labels *normal* and *tumor* as classes, and to use the intensities in the different MR images as features. The training phase under this formulation would consist of learning a model that uses the MR image intensities to discriminate between *normal* and *tumor* pixels. The testing phase would consist of the use of this model to classify unlabeled pixels into one of the two classes based on their intensities (later, we will consider other features, beyond just the intensities).

A major advantage of using a supervised formulation is that supervised methods can perform different tasks simply by changing the training set. This was exemplified by a recent study that assessed the unsupervised *knowledge-based* technique discussed in the previous section and a supervised segmentation algorithm based on the simple supervised classification formulation above (utilizing the *k-nearest neighbors* classifier and patient-specific training pixels) [Mazzara et al., 2004]. In this study, the *knowledge-based* (referred to as KG) and the classification-based (referred to as kNN) methods performed similarly quantitatively, but the

> "kNN method was able to segment all cases, whereas the KG method was limited to enhancing tumors and gliomas with clear enhancing edges and no cystic formation." [Mazzara et al., 2004]

This statement made by the authors supports the argument that supervised methods have the potential to be much more versatile than manually-engineered unsupervised methods. This is due to the fact

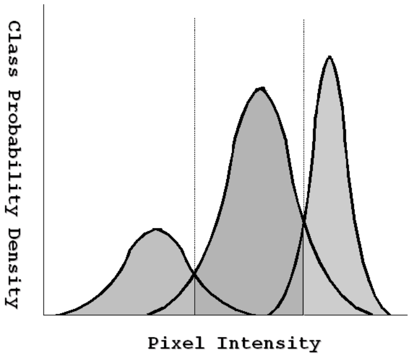


Figure 2.5: Illustration of a simple maximum likelihood classification model (three classes). Three Gaussian distributions are used to model the observed data, and classifications are made by assigning pixels to the class with the highest probability density based on its intensity (represented by the three different gray levels under the curves).

that they use a generic approach to learning to perform a task based on patterns present in the data, that is independent of the actual data used. This reduces the manual engineering task to providing labeled data, appropriate features, and appropriate parameters for the learning algorithm. Section 2.2.1 will discuss many of the systems proposed for brain tumor segmentation in conventional MR imaging modalities (T1-weighted, T2-weighted, and ρ-weighted images) that utilize a supervised classification approach. Section 2.2.2 will then briefly look at recent methods that use supervised approaches with more discriminating imaging protocols.

2.2.1 Supervised Segmentation with Conventional Image Modalities

[Clarke, 1991] was one of the first studies that examined a supervised classification approach for brain tumor segmentation in MR images. This short article compared a *Maximum Likelihood* (ML) classifier to an *Artificial Neural Network* (ANN), finding that the ANN performed better than the ML approach. Training ML classifiers consists of optimizing the parameters of an assumed model of the features (often assuming a parametric model such a univariate or multivariate Gaussian), and assigning pixels to the class that they are statistically most likely to belong to, based on these models (Figure 2.5). In contrast, ANN approaches ‘feed’ the features through a series of nodes, where mathematical operations are applied to the input values at each node and a classification is made at the final output nodes (Figure 2.6). Training for these models consists of determining the values of the parameters for the mathematical operations such that the error in the predictions made by the output nodes is minimized. Since no parametric distribution (such as a Gaussian distribution) is assumed of the data, ANN approaches are non-parametric techniques and, with the use of ‘hidden’ layers of nodes, allow the modeling of non-linear dependencies in the features. Although training of ANN models is more complex than simpler ML models, the ability to model non-trivial distributions offers clear practical advantages. This is noteworthy in the case of tumor segmentation since assuming the data follows a simple Gaussian distribution may not be appropriate for the segmentation of heterogeneous tumors.

[Ozkan et al., 1993] also examined ANN and ML classification methods. This system used the pixel intensities in the different channels and used patient-specific training (meaning that the training data was obtained from the volume to be segmented). This work also showed that Neural Networks outperformed Maximum Likelihood methods, and presented a simple method to account for inter-slice intensity variations. A ML method was used in the ECHO system of [McClain et al., 1995]. This work used patient-specific training and examined the effects of utilizing different combinations of modalities (among T1 pre-contrast, T1 post-contrast, T2, and ρ images), finding that classification

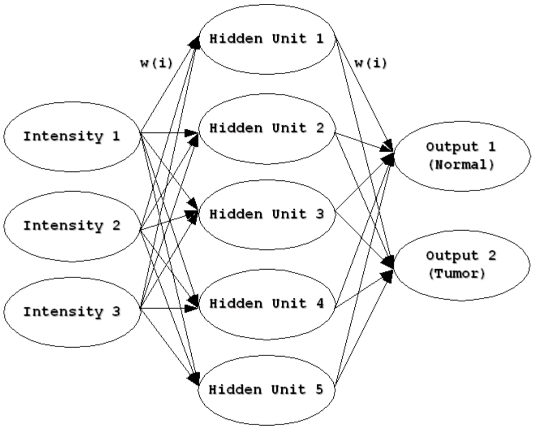


Figure 2.6: Example artificial neural network architecture. Multi-spectral intensities represent the input to this network, linear combinations of the intensities (weighted along the edges with values $w(i)$) are input to the hidden layer of nodes (that may 'squash' the values using a sigmoid or other function), while the output node values are formed from linear combinations of the results of the hidden layer transformations. Pixels are assigned to the class whose output node has the highest value.

improved with additional modalities. This work illustrates the fact that the use of a classification model allows the use of different modalities without any re-engineering of the system.

A very different early system that employed a supervised approach for brain tumor segmentation in MR images was presented in [Schad et al., 1993]. This group used a *Decision Tree* classifier (although this terminology is not used) based on first-order and second-order texture features (a.k.a. *statistical moments* and *spatial cooccurrence* features, respectively). Figure 2.7 illustrates several second-order features computed from a T1-weighted image. Decision Trees are a popular classification technique due to their ability to model non-linear dependencies in the features, and their intuitive graphical representation of the learned model (as opposed to, for example, ANNs). Decision Trees perform classification by making a set of decisions based on the features, beginning from a root 'node' and following decision made to other nodes in the tree where new decisions are made, leading finally to a 'leaf' where a classification is made. Figure 2.8 illustrates our interpretation of the Decision Tree developed in the work of [Schad et al., 1993], that illustrates this concept. Although this group chose their decision rules manually as a set of linear discriminants, there exist a large variety of methods for automatic Decision Tree learning, including the popular C4.5 classifier [Quinlan, 1993]. Although the manual decision selection technically makes the method of [Schad et al., 1993] appear to be closer to the knowledge-based unsupervised approaches (since the authors would need to be contacted in order to apply this learning approach to a different task), the use of the common Decision Tree framework would allow other Decision Tree learners (or other classification methods) to be substituted into this method, making it a supervised classification approach that could be used for different tasks. Schad et al. incorporated their Decision Tree into an expert system written in a variant of PROLOG (a logic programming language) that took into account additional information such as neighborhood analysis. This system segmented images into the 3 normal tissue classes and 2 abnormal tissue classes (tumor and edema) based on T1 pre-contrast and T2 images.

[Vinitski et al., 1997] presented a supervised method that addressed several issues previously ignored in most automatic systems for tumor segmentation. This method used several preprocessing steps before the classification in order to improve results. The first preprocessing step was the coregistration of the different modalities to improve their alignment (in comparison, the method of [McClain et al., 1995] discarded cases where the misalignment was large). The second preprocessing step used by this system was an *anisotropic diffusion filter*, which is a method for edge-preserving non-linear smoothing. This filtering reduces the detrimental effects of local noise (and

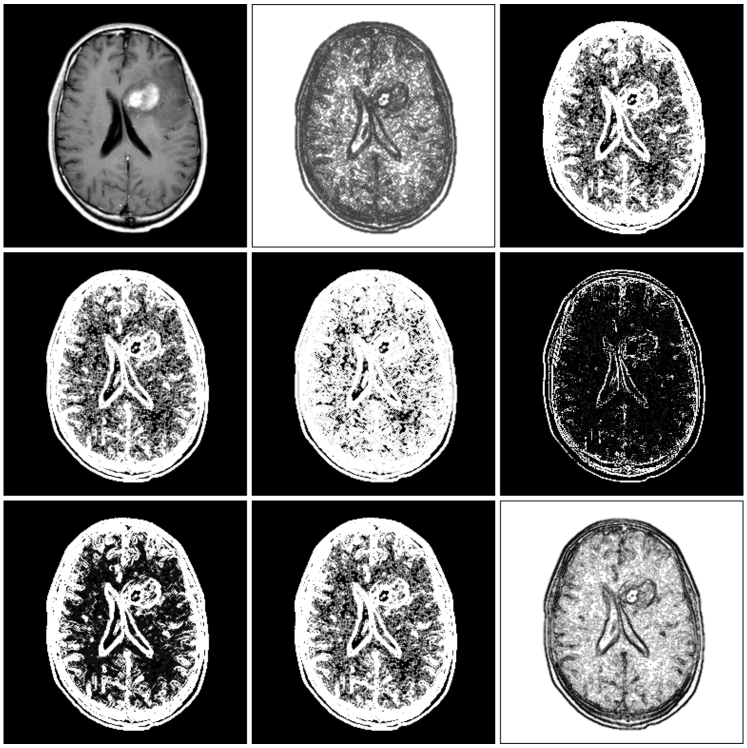


Figure 2.7: Examples of second-order textures computed from a spatial coocurrence matrix. Top, left to right: Original T1-weighted image after the injection of a contrast agent, angular second momentum texture features, contrast texture features. Middle, left to right: absolute value, entropy, cluster shade. Bottom, left to right: Cluster prominence, inertia, local homogeneity

potentially partial volume averaging) on the classification. The third preprocessing step was an *intensity inhomogeneity correction* algorithm. This preprocessing step aimed to reduce errors associated with the intensity inhomogeneity present in the images (discussed in Chapter 1). The method in [Vinitski et al., 1997] used patient-specific training and classified the T1-, T2-, and ρ-weighted images into 10 tissue classes. A method to remove outliers in the training data, in order to account for human errors, was also presented. The classifier used was a *k-Nearest Neighbors* classifier, that assigns labels to pixels based on the most frequent label among the k closest training points under a distance metric applied to the features (referred to as 'lazy' learning, since no explicit model is learned). The kNN algorithm is a simple and effective method for multi-class classification, that is able to model non-linear distributions. Disadvantages of the kNN algorithm include the dependence on the parameter k, large storage requirements (the model consists of all training points), sensitivity to noise in the training data, and the undesirable behavior that can occur in cases where a class is underrepresented in the training data.

A major limitation of most supervised methods for brain tumor segmentation is that patient-specific training is required. [Dickson and Thomas, 1997] presented one of the rare supervised methods that does not require patient-specific. This system used a set of 50 hand-labeled MR slices from the same area of the head of different patients with acoustic neuromas, and learned to automatically label this type of tumor without patient specific training. The features used in this system included not only the pixel intensities, but the intensities of neighboring pixels and the pixel's location within the image. This work compared the use of a kNN classifier, a Learning Vector Quantization (LVQ) classifier, and an ANN. The comparative studies done in this work have provided valuable insights into the problem. These results indicated that the ANN outperformed the other two methods, that pixel neighborhood intensities increase classification performance, that the combination of intensity and texture information performed better than either individually, and that 1 hidden layer

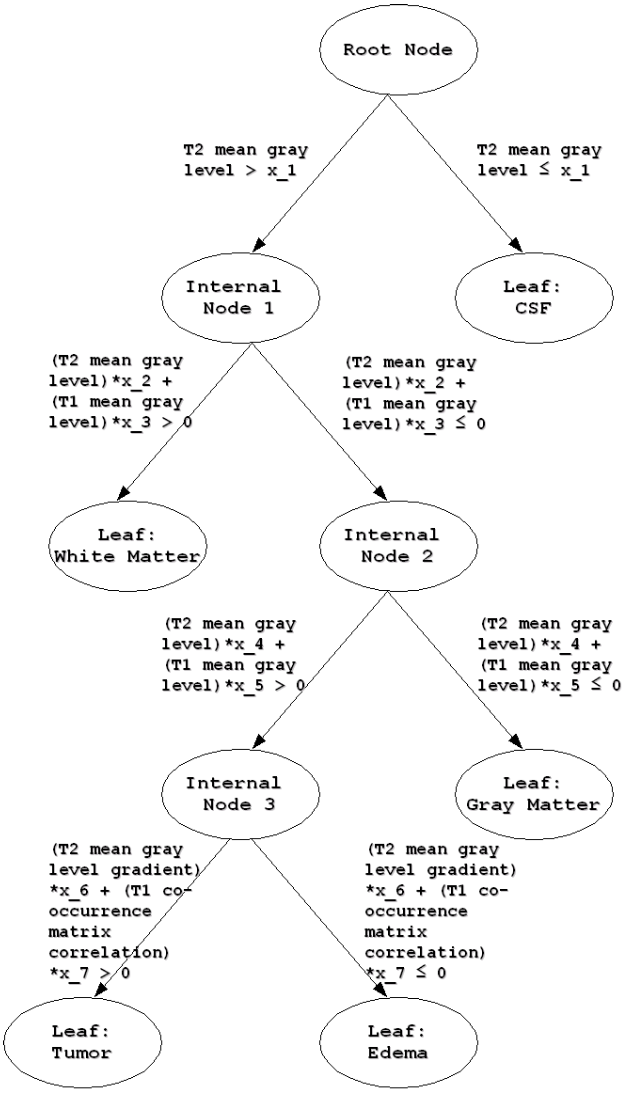

Root Node
T2 mean gray level > x_1
T2 mean gray level ≤ x_1
Internal Node 1
Leaf: CSF
(T2 mean gray level)*x_2 + (T1 mean gray level)*x_3 > 0
(T2 mean gray level)*x_2 + (T1 mean gray level)*x_3 ≤ 0
Leaf: White Matter
Internal Node 2
(T2 mean gray level)*x_4 + (T1 mean gray level)*x_5 > 0
(T2 mean gray level)*x_4 + (T1 mean gray level)*x_5 ≤ 0
Internal Node 3
Leaf: Gray Matter
(T2 mean gray level gradient) *x_6 + (T1 co-occurrence matrix correlation) *x_7 > 0
(T2 mean gray level gradient) *x_6 + (T1 co-occurrence matrix correlation) *x_7 ≤ 0
Leaf: Tumor
Leaf: Edema

Figure 2.8: Our interpretation of the brain MRI tissue classification decision tree from [Schad et al., 1993]. The values x_i are learned parameters of the model (typically the tree topology is also learned). An initial decision is made at the root node to determine if regions represent CSF based on their T2 mean gray level. Those that are not classified as CSF proceed to the first internal node where subsequent decisions are made until the pixel is classified.

in the network topology outperformed 0 or 2 hidden layers. After pixel classification with the ANN, this system performed an unsupervised segmentation to divide the image into homogeneous regions. These regions were assigned a label based on the results of the classifier, and were processed with morphological operations. A second ANN was used to determine whether the resulting regions represented tumors based on a feature set that included shape characterization, the presence of a symmetric region, the structure location, and an approximation of circularity. The accuracy of the results of this system are impressive, and this remains one of the only supervised approaches that does not require patient-specific training to account for intensity non-standardization. In addition to a large training set (50 manually labeled images to learn a segmentation model for one slice of the three-dimensional volume) and a relative degree of intensity standardization across the images (all from the same scanner and protocol), the task in this case was also simplified by the highly localized nature of acoustic neuromas. Since this type of tumor only occurs in a specific location, coordinates were used to enhance the classification accuracy of the first ANN, while the global angle was able to enhance the classification of the second ANN. These features would not represent effective features for tumors that have less predictable locations, unless much larger training sets were available, that allowed the characterization of the entire brain area. [Alirezaie et al., 1997] also examined Neural networks and LVQ classification methods based on pixel intensities and the intensities of neighboring pixels, although this system required patient-specific training.

[Busch, 1997] presented another supervised method that did not require patient-specific training. This work focused on the segmentation of a specific type of non-enhancing homogeneous tumor (low-grade astrocytomas) from ρ-weighted, T2-weighted, and coregistered CT (X-ray) images. This method utilized five texture extraction methods to compute features, and trained 5 LVQ classifiers based on these five feature sets (that were preprocessed with a Kohonen Feature Map). The results of the 5 classifiers were weighted and combined, and the results were post-processed with morphological operations. The use of multiple classifiers (an ‘ensemble’ method) allowed a more robust classification than the individual classifiers. Second-order (spatial cooccurrence) textures provided the worst classification performance among the five texture extraction methods, while first-order (statistical moment) textures performed significantly worse than the other three methods. The Wavelet-based texture features gave a small improvement over the remaining two methods examined. This system achieved highly accurate results, although a large number of training samples were required to achieve this.

[Zhang et al., 2004] presented one of the most recent approach to automatic tumor segmentation in MR images. This approach used Support Vector Machines (SVMs), which are currently an extremely popular method for performing binary classification (in addition to a large variety of other tasks). SVMs will be covered in greater detail in Chapter 4, since the approach presented in this work also takes advantage of their appealing classification properties and their often impressive empirical results. Zhang et al. proposed a simple system for the segmentation of nasopharyngeal carcinomas (another highly localized type of tumor), that used an SVM to perform a binary classification into either the tumor or non-tumor class based on the T1 pre-contrast and post-contrast intensities, and morphological post-processing. This system used patient-specific training and compared two different types of Support Vector Machines, the standard ‘2-class’ method and the more recent ‘1-class’ method. Both methods performed at a similar level of accuracy for this task. However, the advantage of using a ‘1-class’ method was a reduction in the manual time needed to perform patient specific training, since only training examples for the tumor class were needed.

[Garcia and Moreno, 2004] proposed another recent approach for automatic brain tumor segmentation with Support Vector Machines. This work also used patient specific training, and used the intensities of a neighborhood of pixels to make pixel classifications. A 2-class Support Vector Machine (trained by the Adatron algorithm) was used to perform an initial pixel classification, followed by a 1-class Support Vector method that constructed a three-dimensional tumor model from the pixel classifications.

2.2.2 Supervised Segmentation with Advanced Image Modalities

To complete this survey of supervised approaches to brain tumor segmentation, there have been several supervised (and some unsupervised) approaches to tumor segmentation that use more discriminative MR imaging protocols. Although these will not be covered in detail, several recent methods will be briefly discussed. The advantages of these methods are that they may facilitate an easier automatic segmentation task, and that they may more appropriately characterize the extent of the tumor infiltration. The disadvantages of these approaches are that they require additional acquisition time, and that the additional modalities are not available for historical data, nor are these acquisition protocols commonly used. [Soltanian-Zadeh et al., 1998] evaluated a method to segment tumors from the combination of 4 T2-weighted images with different parameters and 2 T1-weighted images of different parameters, or the combination of 2 T2-weighted images, 2 T1-weighted images, and a Fluid-Attenuated Inversion Recovery (FLAIR) image, that is similar to a T2 image but

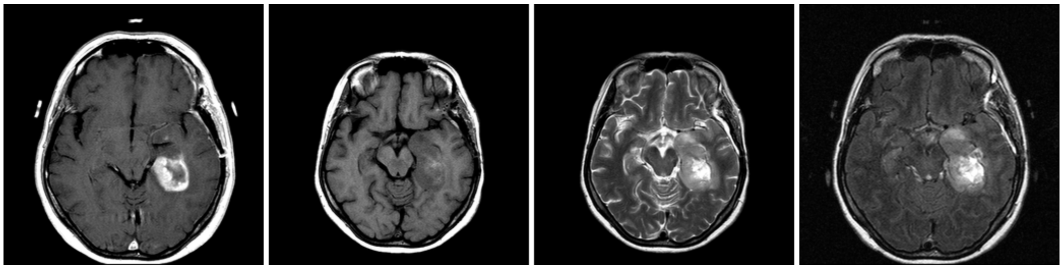


Figure 2.9: Example of a set of imaging modalities that would more easily facilitate segmentation. Left to right: T1-weighted post-contrast injection image, T1-weighted pre-contrast injection image, T2-weigthed image, FLAIR image. The FLAIR image measures a signal similar to the T2 image, but suppresses free water. Thus the abnormal area appears hyper-intense (as in the T2), but many normal areas that also appear hyper-intense in the T2 image remain hypo-intense in the FLAIR.

allows better tumor and edema visualization since free water is suppressed (Figure 2.9). The system presented to segment this large combination of images used patient specific training, and consisted of coregistration, brain masking, anisotropic filtering, intensity non-uniformity correction, and finally an *eigenimage* (or principal component) analysis that detected similarity to the lesion class. The results of this system indicated tumor infiltration extended beyond the regions visible in normal images, which was confirmed by biopsy data. [Peck et al., 2001] also proposed a technique for supervised segmentation of brain tumors in more advanced modalities, performing segmentation based on MR Spectroscopy data that measured the presence of Choline, Creatine, NAA, and Lactate. This work used patient-specific training based on normal areas, and the segmentation was performed based on a measure of abnormality. Two other works that use a simpler approach and a simpler measure of abnormality using patient-specific training on normal areas in MR Spectroscopy images were [Pirzkall et al., 2001] and [Stadlbauer et al., 2004].

2.2.3 Summary of Supervised Segmentation

This section has surveyed a variety of techniques proposed to perform tumor segmentation that incorporate training data. Several general observations that can be made from examining this work include that intensity and texture information may be complimentary to each other, that additional modalities can simplify the task, that coregistration is important in using multi-modality data, that intensity preprocessing methods (such as anisotropic filtering, inter-slice intensity variation correction, and intensity inhomogeneity correction) can improve results, and that classifiers that do not assume a simple distribution and take into account non-linear dependencies in the features (such as ANN models) have tended to outperform other methods (although the SVMs and model averaging techniques used by several recent works have not yet been compared directly to an ANN model).

Although highly effective and versatile, supervised methods of brain tumor segmentation in MR images often suffer from the disadvantage of requiring patient-specific training, with only a few exceptions. The exceptions that were able to perform inter-patient classification focused on relatively simplified tasks, and required a large amount of training data. A final noteworthy point is that the approaches that utilize additional or more advanced MR modalities can facilitate a simpler segmentation task and may provide a more appropriate definition of the actual tumor location, and thus the method presented in this dissertation has been designed with the goal that additional modalities can be easily incorporated (or replace existing modalities) to improve results.

2.3 Registration-Based Segmentation

Classification-based segmentation is a popular and appealing technique that takes advantage of labeled training data. An alternate approach to using labeled data is through the use of *spatial reg-*

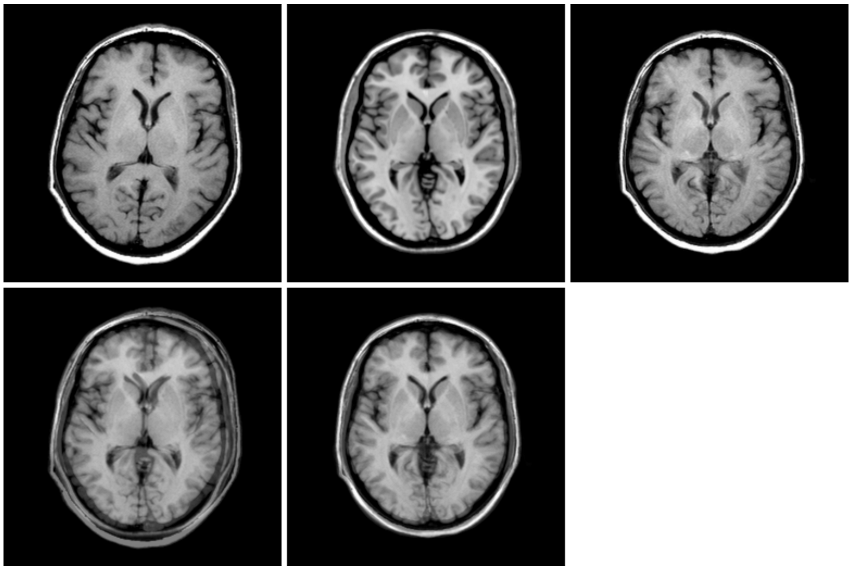


Figure 2.10: Example of non-linear spatial registration. Top, left to right: T1-weighted image to be spatially aligned, T1-weighted template used in alignment (from [ICBM View, Online]), T1-weighted image after non-linear registration to the template. Bottom left: average intensity of T1-weighted original image and registration template before spatial registration. Bottom middle: average intensity of T1-weighted original image and registration template after spatial registration. Non-linear spatial registration not only aligns the images and corresponding structures, but also takes into account differences in overall head shape and anatomic variability.

istration (Figure 2.10). Segmentation approaches based on this idea typically first align a labeled template (or atlas) image with the image to segmented, and infer the labels for the new image by assuming that they correspond to the labels of the aligned template. The advantage of this type of method is that spatial information is encoded through the use of the template, as opposed to pixel classification based methods that encode limited spatial information. The major disadvantages of this type of method are that the registration may not be perfect, and that there may be anatomical differences between the template and the image to be segmented. These disadvantages make template registration methods inappropriate to apply directly for tumor segmentation, since the template does not have a tumor, nor is its anatomy affected by the presence of a tumor. However, the ability to use spatial information derived from the spatial alignment of a template is appealing, and there has been considerable recent effort focusing on the incorporation of template registration into methods for tumor segmentation.

The influential work in [Kaus et al., 2001] presented a method that incorporated both supervised classification and template registration for the segmentation of homogeneous brain tumors from T1-weighted images. This method employed a kNN classification algorithm with patient-specific training, used a label-based registration algorithm based on principles of optical flow, and preprocessed images with an anisotropic diffusion filter before analysis. After preprocessing, the segmentation consisted of performing kNN classifications (refined through the use of morphological operations), followed by the use of non-linear registration with a labeled template. A ‘distance transform’ was computed based on each pixel’s distance to the template labels and was used to refine the kNN classifications. As in the knowledge-based systems, this system proceeds from operating on the entire image to focusing on specific areas of interest. Tumor tissue was initially included as part of the ‘intra-cranial cavity’ (brain) class, but this class is eventually classified into tumor, ventricles, and an ‘other’ class. This idea of using registration to remove normal structures that may have similar intensities to tumor pixels greatly simplifies this task, by taking advantage of the predictable spatial location of the brain and ventricles within the template. The validation work presented for this sys-

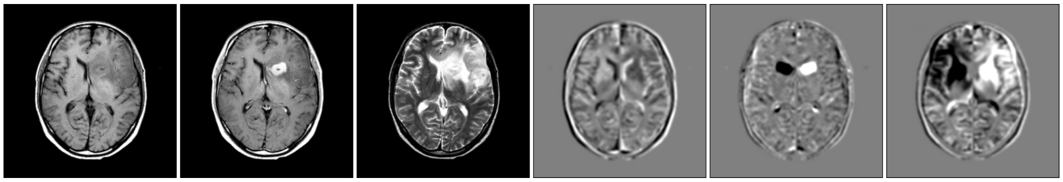


Figure 2.11: Example of bi-lateral symmetry. Left to right: Original T1-weighted image, T1-weighted image after contrast injection, T2-weighted image, pixel-level bi-lateral symmetry of the T1 image, pixel level bi-lateral symmetry of the contrast difference image, pixel level symmetry of the T2 image (symmetry images are smoothed by a Gaussian filter).

tem indicates near perfect segmentations for the two types of homogeneous tumors it was applied to.

A more recent system that proposed to take advantage of template registration in segmentation was presented in [Gering, 2003a]. This method proposed to use a database of normal brains as training data. Each normal brain would be registered with the image to be segmented, and a simple multi-resolution statistic would be computed for each pixel to determine how significantly it differed from the most similar normal brain at that location. Although the statistic computed makes false independence assumptions (a density estimation method may be more appropriate) and the use of equally weighted square neighborhoods resulted in a ‘blocky’ effect, this represents an interesting and very different approach to the problem. One of its weaknesses is that it does not account for the intensity non-standardization effects that would be present in a large database of normal brains, while another weakness is the lack of availability of a database of completely normal brains. The author addresses this second weakness by proposing that bi-lateral symmetry could be used as a patient-specific template representing an average normal brain (Figure 2.11), and presents impressive results using this idea, given the simplicity of the method. Although this indicates that symmetry is clearly an important feature, the sole use of symmetry is obviously insufficient since the statistic will be equal for both sides of the brain (including the ‘normal’ side), and tumors that cross the mid-saggital plane may not be asymmetric at the pixel level. Another complication with using symmetry is locating the axis of symmetry, which is not discussed in this work, but is a non-trivial task in the case of large tumors that deform normal anatomy. Another consideration when examining this method is that an ‘abnormal’ area with respect to normal brains does not necessarily mean that the area represents tumor. For example, tumors can have associated edema, can deform nearby normal structures, and can result in other physiological effects (such as enlargement of the ventricles).

2.4 Expectation Maximization Segmentation with Spatial Prior Probabilities

For the task of segmenting head MR images into the three normal brain classes (grey matter, white matter, and cerebrospinal fulid), *Expectation Maximization* approaches have become a popular framework, since they have shown to be robust to both intensity inhomogeneity and intensity non-standardization. [Wells et al., 1996] was the first group to formulate the task of normal brain segmentation as an Expectation Maximization problem. The main observation underlying this work was that segmentation could be simplified if the inhomogeneity field was given, while estimating the inhomogeneity field would be simplified if a segmentation was given. Expectation Maximization (EM) schemes are a natural approach to perform classification in this situation. This algorithm consists of an Expectation step where the tissue class parameters are computed given the current estimation of the inhomogeneity field, and a Maximization step where the inhomogeneity field is computed given the current estimation of the tissue class parameters. This algorithm converges to a local optimum, and the use of an (adaptive) Gaussian Mixture Model makes the method robust

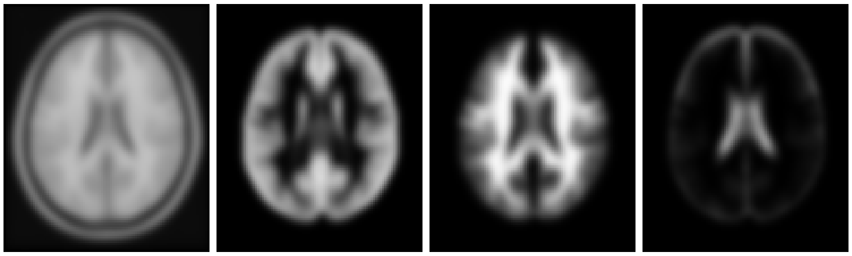


Figure 2.12: SPM priors [SPM, Online]. Left to right: T1 registration template, gray matter spatial prior probability, white matter spatial prior probability, CSF spatial prior probability.

to intensity non-standardization. They also present an alternate formulation using non-parametric tissue class estimation with Parzen Windowing.

Many extensions have been proposed to the original method of Wells et al. One of the most notable is the method in [Leemput et al., 1999a, Leemput et al., 1999b]. This work implemented a method proposed by Wells et al. to automate the initialization of the tissue parameters through the use of template registration and empirical *spatial prior probabilities*. The spatial prior probabilities used were those provided with SPM96 (derived from work in [Evans et al., 1992a, Evans et al., 1992b, Evans and Collins, 1993, Collins et al., 1994] and shown in Figure 2.12), which includes prior probabilities for the three normal tissue classes. These priors represent the empirical likelihood, at each pixel in an image registered into the template's standard coordinate system, that the pixel belongs to each of the three classes. An intensity-based registration method was utilized (as opposed to a label-based registration method as in [Kaus et al., 2001]) to register new images with the template, and these prior probabilities allowed a robust initialization of the tissue classes, negating the need for patient-specific training, since the algorithm adaptively adjusts for intensity non-standardization effects. An appealing aspect of this algorithm is that the use of spatial information makes the algorithm robust to the type of MR image used, and thus can be applied to a set of coregistered images of different modalities or to a single image modality. Another major contribution in the work of [Leemput et al., 1999a, Leemput et al., 1999b] was the incorporation of a Markov Random Field to probabilistically smooth the labels of adjacent pixels, making the technique more robust to pixel-level noise.

Several factors complicate the application of Expectation Maximization approaches to the segmentation of brain tumors. The first main factor is that there is no prior available for the tumor class, removing the ability to use the straightforward and robust initialization of the tissue class parameters using the priors. Tumor heterogeneity is a second factor that complicates the direct application of this algorithm, since heterogeneous tumors are not modeled effectively by Gaussians. A third complicating factor is that tumors interfere with both the inhomogeneity estimation and the tissue class estimation, potentially leading to poor local optimums. The final complication of applying this algorithm to tumor segmentation is that the tumor class can be very similar or can overlap in intensity with the normal tissue classes, and this approach will have significant difficulty discriminating between different areas that may have similar or the same intensities.

[Moon et al., 2002] was the first approach that adapted an Expectation Maximization approach with spatial prior probabilities to the task of tumor segmentation. This work utilized a simple method to approximate a prior for enhancing tumor pixels and another for edema. The enhancing tumor prior was constructed from the contrast agent difference image, while edema was assumed to be co-located with white matter. The preprocessing phase consisted of coregistration of the modalities and template registration, both done as a linear affine transformation maximizing Mutual Information. A version of the Expectation Maximization segmentation method of Van Leemput et al. was used, and was applied to T1 pre-contrast and T2 images for the segmentation of tumors with an enhancing boundary and the segmentation of edema. Unfortunately, this approach is only applicable to the segmentation of enhancing tumors, and it is not obvious how to obtain a meaningful approximate prior for more difficult cases.

This group later presented a more general approach in [Prastawa et al., 2004]. This approach addressed additional challenges associated with applying Expectation Maximization approaches to tumor segmentation. This work used a single abnormal class to identify both tumor and edema, but detected abnormal pixels as outliers from the three normal tissue classes (where were estimated using a non-parametric model). This allowed the identification of tumors that were both enhancing and non-enhancing, and also allowed the identification of heterogeneous tumors. This work also presented a method to make the initialization of the normal tissue class parameters robust to the presence of tumors, by detecting and removing outliers during the tissue class estimation. This system post-processed the Expectation Maximization results to divide the abnormal class into tumor and edema (if necessary), and a Level Set method was applied to the tumor class while morphological operations were applied to the edema class.

[Gering, 2003b] outlined a third recent approach using a variation of the Expectation Maximization approach, which also detected tumors as intensity outliers from the normal classes. This work addressed the problem of tumor interference with the inhomogeneity field estimation by reducing the weight, during inhomogeneity field estimation, given to pixels that have a greater degree of abnormality. This work refined the Expectation Maximization results with a Markov Random Field (using a mean field approximation), incorporated a 'structure to boundary' constraint using a multi-level Markov Random Field, and presented a method to discriminate partial volume pixels from tumor pixels by creating an adaptive spatial prior for pixels that are at the boundaries of normal structures. These three impressive additions to the Expectation Maximization algorithm are combined into a structured 'Contextual-Dependency Network' for the segmentation of brain tumors from T1-weighted images. The multi-level Markov Random Field in particular addressed a major weakness of the Expectation Maximization methods since it allows the identification of tumor structures that have normal intensities but are too 'thick' to be normal. Unfortunately, this is only applicable to tumors that are homogeneous enough to be segmented into a single normal tissue class, and therefore is not generally applicable to heterogeneous tumors.

2.5 Summary of Existing Approaches

The previous section has introduced the popular Expectation Maximization approach for the segmentation of normal brains, and discussed several efforts to employ this methodology in methods for the segmentation of brain tumors. Several of the problems with this transfer have been addressed. This includes detecting tumors as outliers or utilizing the contrast agent difference information to account for the lack of a prior, using robust estimators in tissue class initialization, assigning less weight to increasingly abnormal pixels in the inhomogeneity field estimation, and the detection of large homogeneous tumor areas that have relatively normal intensities. In our opinion, the current state of the art methods for automatic brain tumor segmentation are the two Expectation Maximization approaches that use outlier detection [Prastawa et al., 2004, Gering, 2003b] (due to the robustness to intensity non-standardization), along with the registration-based method of [Kaus et al., 2001] (due to the use of non-linear registration and the more extensive use of spatial information to enhance discrimination) and the neural network based methods of [Dickson and Thomas, 1997, Busch, 1997] (due to the use of textural information and more powerful classification techniques). In this work, we extend these methods by presenting a system that integrates the use of a powerful classification technique, the use of textural information, the use of a large amount of spatial information to enhance discrimination through spatial registration, and finally the use of a template-based intensity standardization step that makes the technique robust to intensity non-standardization.

Chapter 3

Overview of Automatic Segmentation Framework

The previous chapter discussed a variety of approaches proposed in the literature for the task of automatic brain tumor segmentation. This chapter will first motivate our proposed approach for automatic brain tumor and edema segmentation, and then present the different elements of the framework. We will begin by briefly reviewing the important components that a system should have, based on the previous work. This will be followed by a comparison of the advantages and disadvantages of the four types of approaches discussed in the previous chapter, and a discussion of how human experts perform this task. This leads naturally into the motivation for developing our approach, which incorporates many of the ideas presented in the previous works, but further explores the utility of spatial registration in order to additionally take advantage of the types of information used by human experts to perform this task. The sections following this motivation will present the individual components in the segmentation framework.

Based on the existing literature, several general conclusions can be drawn with respect to elements of a segmentation system that can be used to improve performance:

- Edge-preserving low-pass filtering can reduce the effects of local noise and partial volume averaging.
- Intensity inhomogeneity correction can reduce the effects of intra-volume intensity inhomogeneity.
- Additional modalities can enhance discrimination between the classes.
- Coregistration can allow the use of modalities that otherwise may not be aligned.
- Intensity and texture information can be combined to improve discrimination between the classes.
- Spatial registration with a template can allow template information to be used to enhance discrimination.
- Spatial registration with a template can allow the template to be used in intensity standardization.
- Label relaxation operations can improve pixel-level classification results.
- Information about shape, size, symmetry, and normal anatomic variability can improve segmentation results.

In order to decide which type of approach to use, we will first briefly review the advantages and disadvantages of each the four general types of approach discussed in the previous Chapter. These can be summarized as follows:

- Unsupervised Methods: These approaches offer the advantage that they do not rely on training data, and therefore are not subject to any degree of variation due to human interpretation. These methods have the disadvantage that they have been limited to simple tasks, where there is an obvious indicator of abnormality such as the presence of a contrast agent. Another disadvantage of these methods is that significant re-engineering is required in order to apply these methods to new tasks or to use different modalities than those the system was designed for.

- Supervised Methods: These methods have the appealing advantage that they can be applied to new tasks or can use different modalities without the need for redesign. Another advantage of supervised methods is that learning to meaningfully combine different potential sources of evidence for the presence of tumor can be done automatically. These methods are much more effective than unsupervised methods at tasks where the different tissue classes may be very similar, since the training aspect focuses on learning a model that will be effective at exactly this task. The major disadvantage of this type of approach is that methods have typically required patient-specific training (ie. training pixels from the volume to be segmented as opposed to from other volumes), primarily due to the problem of intensity non-standardization. The requirement of patient-specific training means that these methods are not fully automatic, and that they are also subject to manual variability. A small number of systems have presented inter-patient training methods, but these have focused on relatively simplified cases and required large training sets.

- Registration-Based Methods: Registration provides the ability to use spatial patterns and constraints within a system. Although it is clear that this could be used to significantly improve results, it is not necessarily obvious how the information provided by registration should be used. Existing methods use registration to focus the segmentation on the brain area (and to differentiate tumors from the ventricles), or to assess how much pixels deviate from corresponding locations in normal brains. Although these can demonstrably help in segmentation, it is obvious that there remains the potential to incorporate additional registration-based information.

- Methods initializing tissue models with spatial priors: These recent methods are appealing since they provide a structured and statistically sound method to confer a large degree of robustness to problems related to intensity non-standardization, through the initialization of tissue parameter models with spatial information. One disadvantage of these methods is that abnormal tissues may not adhere to the assumed model for normal tissue, or may interfere in its estimation. However, recent methods that model tumor tissues as intensity outliers have largely addressed this issue and provide the advantage that they can be used to segment different types of tumors (that are intensity outliers) without any change to the algorithm. The problem remaining with the outlier detection approaches is that fine discrimination between normal pixels and tumor pixels that have similar intensities is not possible. Some preliminary work has been done to address these issues, but these have only addressed simple cases (areas visible due to the presence of a contrast agent, areas adjacent to tumor pixels that are intensity outliers, and sufficiently thick homogeneous regions), while real data is often much more complex.

In designing a new system to perform this task, the existing work is a valuable source of insight into this problem, since it is clear that there are a variety of different properties of the problem that can be exploited. However, since the goal of our work is to automate a task performed by human experts, the methods used by human experts also provide insights into the problem. It is clear that human experts do not build an internal intensity-based pixel classifier, and are able to incorporate much more complex information. This includes knowledge of the expected appearance, location, and variability of normal anatomy, but also includes patient-specific bi-lateral symmetry, knowledge about the expected intensities of different tissues relative to each other in a modality, and the evaluation of the appearance of regions of pixels and/or shapes present within the image. Furthermore, humans are able to simultaneously consider and combine these diverse sources of evidence, and can consider previous experience in related tasks.

It is obvious that in cases where the discrimination between normal and abnormal areas is not trivial, that a variety of sources of evidence are used in making a decision. These various sources of evidence could include the intensities in different modalities (relative to each other), the textures observed, bi-lateral symmetry, similarity to a normal brain, expected tissues or structures at spatial positions, expected and observed shapes of different structures, assessing normal anatomic variations, assessing variations due to the presence of a tumor, and evaluating pixel regions. As outlined in Chapter 2, there has been an attempt to incorporate each of these items (to at least some extent), into systems for automatic brain tumor segmentation. However, usually only a few of these items are considered, and the different sources of evidence are often not considered simultaneously. This is understandable due to the fact that it is not obvious how to use each of these items meaningfully towards the goal of achieving an accurate segmentation. It is clear, however, that improved results could be achieved if a system could consider a variety of sources of evidence simultaneously in performing a segmentation.

Before examining how these different sources of evidence could be incorporated into an automatic system, we will first examine an important question. Given a variety of sources of evidence that can help to indicate the presence or absence of tumor, how can these sources be combined meaningfully to achieve an accurate result? Although the goal of the *knowledge-based* systems discussed earlier was to find ways to meaningfully incorporate different sources of information as rules, these systems fall short of this goal in ambiguous cases due to the fact that the patterns are complex and involve interactions between the different sources of evidence. These complex interactions are difficult to represent with a set of 'hard' manually determined rules. A supervised approach seems ideal, since supervised learning focuses specifically on finding patterns in (potentially large and complex) sets of (potentially interacting) features in order to optimize a performance measure, such as the number of misclassified pixels. In fact, the literature already supports the idea that intensity and texture information or different textural measures can be combined to increase performance. Unfortunately, it has also been mentioned that supervised learning methods have typically required patient-specific training in order to achieve accurate results, due to intensity non-standardization.

The INSECT system (Intensity Normalized Stereotaxic Environment for Classification of Tissues) is a popular image processing 'pipeline' for the automatic segmentation of Multiple Sclerosis lesions in MRI [Zijdenbos et al., 1998]. This system uses a large number of preprocessing steps that have previously been discussed such as intensity inhomogeneity correction, noise reduction, interslice intensity variation correction, template registration, coregistration, and brain masking. What makes this especially noteworthy is that this system uses a supervised learning (ANN) approach, but does not require patient-specific training, even though it has been validated at seven different locations and thus was subject to data with variations in intensities. A key additional operation that allowed this was a simple preprocessing step prior to classification, an intensity standardization step. Although an obvious solution to intensity non-standardization, standardizing intensities is not necessarily a trivial operation, since pathology can interfere with standard histogram-based methods. The INSECT system uses the spatially aligned template image in performing this standardization step, allowing spatial information in addition to intensity information to be used in the intensity normalization step.

In addition to providing a method to escape the dependency of patient-specific training in a supervised learning framework, the INSECT system also provides a simple method to incorporate an additional source of evidence into the classification. The classifier in this system uses 6 pixel-level features as inputs, consisting of the intensities in the three channels (T1, T2, and ρ), and the values of the three spatial prior probabilities (gray matter, white matter, and CSF). The incorporation of the spatial prior probabilities within the classification step improves accuracy compared to using the three intensities alone [Zijdenbos et al., 2002]. This is due to the fact that the classifier can use these probabilities to incorporate additional 'spatial context' into the classification. The increased discriminative power allowed through the addition of the priors is similar to the addition of textural information discussed earlier, but these additional features have a more intuitive anatomical meaning. This can be illustrated by an example. Consider a lesion pixel that is similar to CSF based on its multi-spectral intensities, but occurs in a location that has a very low probability of being CSF based on its prior. An intensity based classifier may mis-classify this pixel as CSF due to its intensity properties, but a classifier that simultaneously incorporates the priors as features could use the pixel's low probability of being CSF to disambiguate the intensity information and classify the pixel as a lesion.

At first, the use of the priors in addition to intensities in a classifier appears to be similar to the Expectation Maximization approaches that make use of the same information. However, the use of a supervised classification method in the INSECT system allows the priors to be combined with the intensity information in a way that optimizes a classification performance measure. In addition, since the method of combining the priors and the intensity information is learned instead of being explicitly defined in the model, other features (where the interactions are not as obvious) could be added in the same way.

It is noteworthy that features like spatial prior probabilities are obtained in a way different than traditional features such as textural features would be obtained. While textural features are obtained directly from the image data (an image-based feature), the spatial priors are obtained through the use of registration to a standard coordinate system (a coordinate-based feature). The INSECT system has thus provided a means by which the requirement of patient-specific training can be removed (intensity standardization), and also presented a simple method to incorporate a form of coordinate-based pixel feature in the form of spatial prior probabilities. This leads naturally to two important observations that underlie the work in this dissertation:

- Spatial prior probabilities are not the only coordinate-based (or registration-based) features that can be used.
- It may be possible to combine more advanced image-based features with these coordinate-based features to achieve more accurate results than either could individually.

This chapter will present our automatic segmentation framework, which was designed to explore and take advantage of these ideas. This framework performs automatic brain tumor segmentation in MR images using the available modalities, in a supervised learning framework that incorporates preprocessing of the intensity data, template registration, and supervised classification based on a set of features that are derived from both the image and the registration of a template in a standard coordinate system. These features will be discussed further in Sections 3.4 and 4.4, but the main motivation behind them is that the classification of pixels as being tumor or normal will not be based solely on intensity and textural information nor solely on intensity and spatial priors, but will additionally and simultaneously consider information such as characterizations of anatomic variability, patient-specific bi-lateral symmetry, and the image properties at the corresponding location in the template image (all measured at multiple scales).

The remaining sections of this chapter will outline the automatic segmentation framework, while the next chapter will present an implementation of this framework. The motivation for this segregation is that, although our implementation contains well-justified and state-of-the-art methods for each of the steps, most of the steps in the framework represent open research problems. Thus the system has been purposely divided into a series of steps with the intention that improved methods can be incorporated into the framework at any stage in order to improve the final results. Another reason that the framework is distinguished from the instantiation is that the framework has been designed such that it can be easily adapted to different tasks (such as the segmentation of Multiple

Sclerosis lesions or normal structures in the brain), or to use different modalities (such as other or more advanced MR modalities, or other imaging modalities such as CT or PET), although this will not be explored in this work. This chapter will therefore concentrate on outlining the purpose of each step, and describing the general types of methods that are suitable for these steps. The steps in the framework are as follows (and are illustrated in Figure 3.1):

1. Noise Reduction: Initial image enhancement to reduce the effects of noise, inter-slice intensity variations, and intensity inhomogeneity.

2. Registration: Spatial alignment of the different modalities and alignment of the images with a template in a standard coordinate system.

3. Intensity Standardization: Transformation of the image intensities to approximately 'calibrate' them to the template intensities.

4. Feature Extraction: The calculation of pixel-level features that represent information from the image, the coordinate system, and the template registration.

5. Classification: The assignment of a class label to each pixel in the image based on its features and a learned supervised classification model.

6. Relaxation: Refinement of the pixel classifications by taking into account dependencies in the class labels of neighboring pixels.

7. Post-Processing: Additional registration, segmentation, or other processing steps

3.1 Noise Reduction

The first stage in the processing pipeline is noise reduction. This stage aims to reduce the effects of local noise, inter-slice intensity variations, and intensity inhomogeneity, before further processing of the images takes place. This step is not vital, since the image acquisition protocols are often specifically designed to reduce these effects, seeking to have high signal to noise ratios, to reduce the intensity inhomogeneity within slices, and to have consistent intensities between slices. Furthermore, the features that will be discussed in Sections 3.4 and 4.4 will make the classifier relatively robust to these minor effects. Since these effects will typically not be severe, an important property of the algorithms used for this step is that they do not introduce additional noise, and only seek to make minor corrections. Another important point to note is that the methods used for Noise Reduction should not depend on having a segmented image, and should not rely on a normal brain tissue model. To summarize, the Noise Reduction should:

1. not depend on a prior segmentation (there will be no segmentation available).

2. not depend on a prior tissue model (large abnormalities can interfere with this model).

3. introduce minimal additional noise (under-compensating is more desirable than over-compensating).

4. avoid making major corrections (these effects are considered minor).

The input to the Noise Reduction stage will be the raw images in the different modalities. The output of the Noise Reduction phase will be the same images, but with the same or reduced levels of local noise, inter-slice intensity variations, and intensity inhomogeneity.

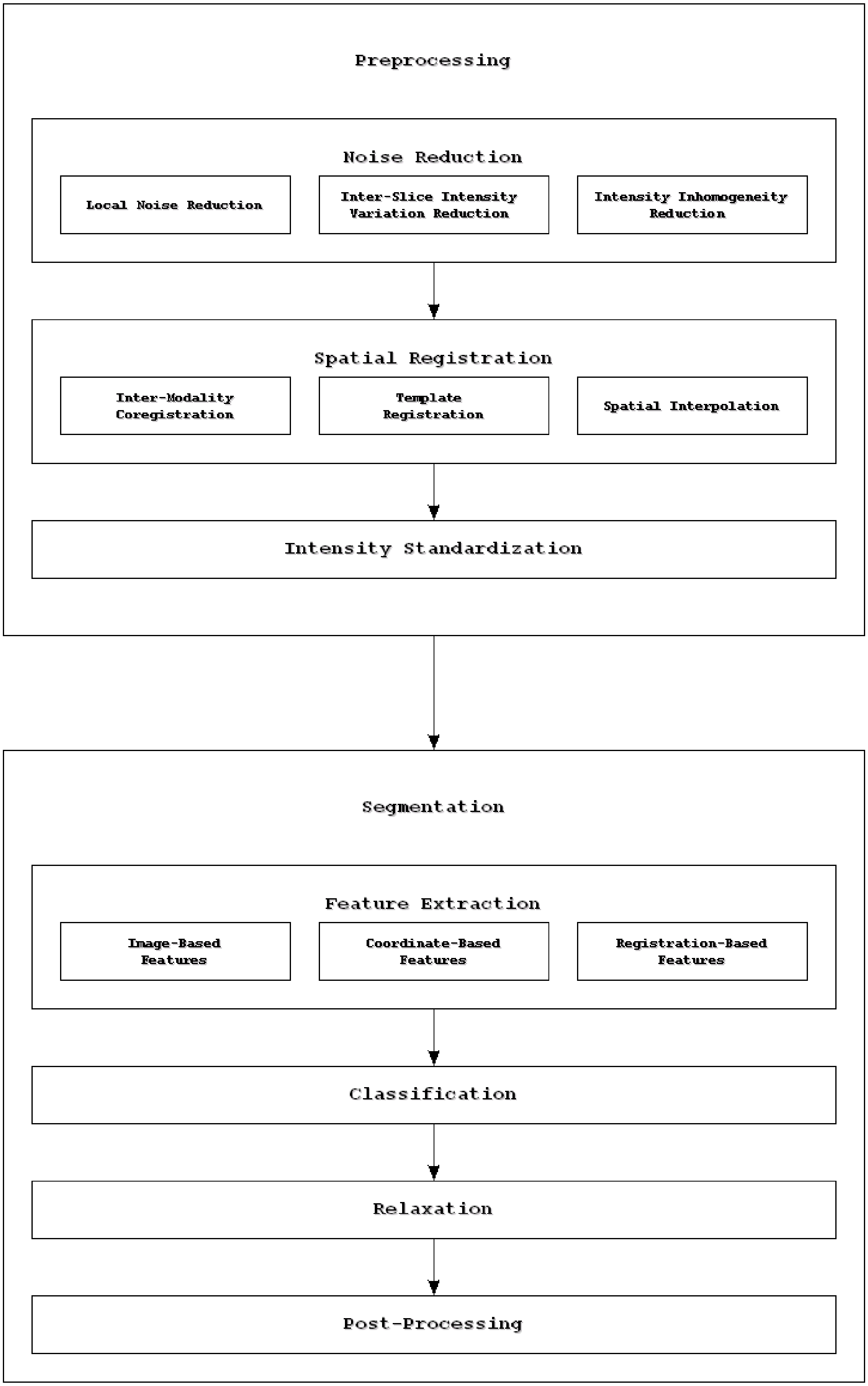


Figure 3.1: Overview of presented framework.

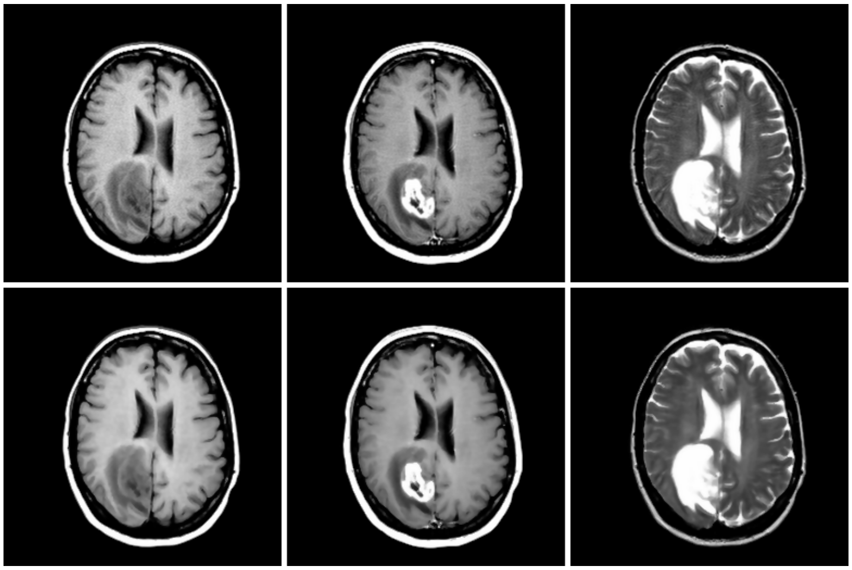


Figure 3.2: Example of Local Noise Reduction. Top: Original image modalities. Bottom: images after edge-preserving smoothing.

3.1.1 Local Noise Reduction

The phrase ‘Local Noise’ refers to noise that corrupts the signal recorded at each pixel. This noise is often considered to be additive and independent of pixel location, but dependent on the tissue measured at the location. This type of noise would be simple to correct given a complete segmentation into the different anatomic tissue classes, since this noise follows predictable distributions. Specifically, this noise has been characterized as following a Gaussian distribution near tissues with a strong signal and a Rayleigh distribution near tissues with a weak signal [Gering, 2003b]. A Rayleigh distribution is defined over the interval $[0, \infty]$ with parameter σ as:

$$P(r) = \frac{re^{-r^2/2\sigma^2}}{\sigma^2} \tag{3.1}$$

However, modeling this noise without a segmentation (or tissue model) is a more challenging task. As discussed in Chapter 2, *Anisotropic Diffusion Filtering* is a technique commonly used to reduce the effects of local noise. This technique was first adapted to MR images in [Gerig et al., 1992], and represents a simple method to reduce the effects of local noise without requiring a tissue model. Anisotropic Diffusion Filtering and other edge-preserving smoothing methods are examples of techniques that satisfy the desired properties for a Noise Reduction algorithm in this framework. They do not depend on a segmentation, do not require a tissue model, do not introduce additional noise (if the parameters are set appropriately), and only make small local correction (if the parameters are set appropriately). The effects of an edge-preserving smoothing operation are demonstrated in Figure 3.2.

3.1.2 Inter-Slice Intensity Variation Reduction

Inter-slice intensity variations refer to the sudden changes in intensity that can be observed between adjacent slices produced with some imaging techniques. Since the presence of this type of inhomogeneity is dependent completely on the acquisition protocol used, this step is only needed if this effect is present in the data. This effect can be corrected by modeling it as part of a three-dimensional inhomogeneity field as in [Leemput et al., 1999a, Leemput et al., 1999b], or it can be corrected by using techniques that standardize the intensities of adjacent slices as in the INSECT system (that uses the method from [Zijdenbos et al., 1995]). If inter-slice intensity variation reduction is included, it

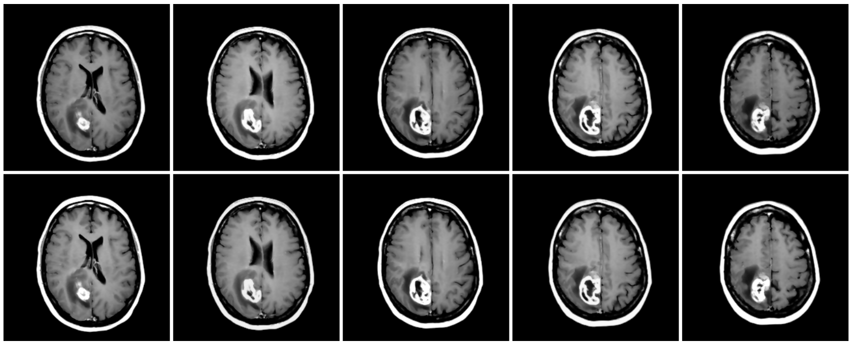


Figure 3.3: Inter-Slice Intensity Variation Reduction. Top: Original set of five adjacent slices after edge-preserving smoothing (note the increased brightness of the second and fourth slice). Bottom: Slices after reduction of inter-slice intensity variations.

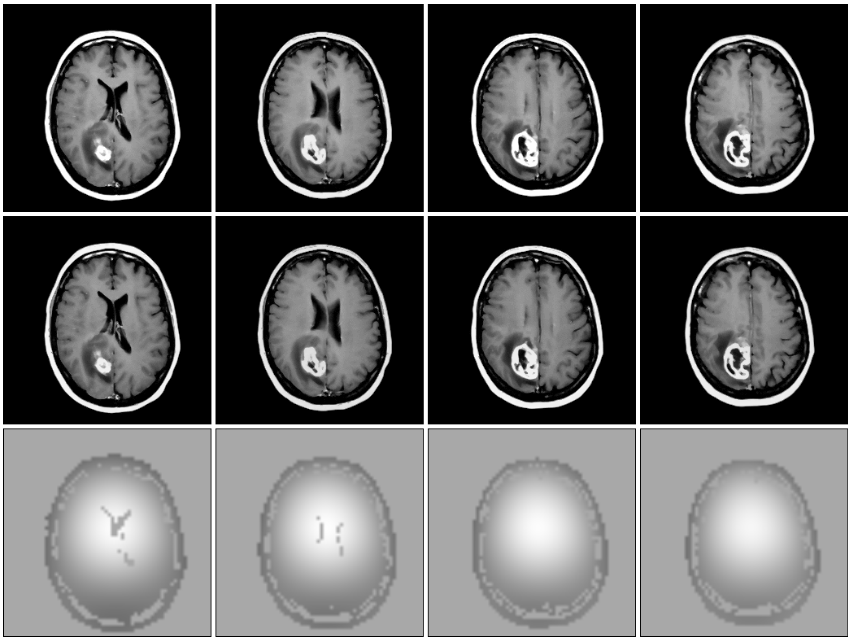


Figure 3.4: Example of Intensity Inhomogeneity Correction. Top: Set of adjacent slices after edge-preserving smoothing and reduction of inter-slice intensity variations. Middle: Slices after correction of intensity inhomogeneity by the N3 algorithm [Sled et al., 1999]. Bottom: Computed inhomogeneity fields (note that pixels below an intensity threshold are not used in estimating the field).

should follow the guidelines that apply to the other correction steps in this section. Since most approaches for this correction rely on a tissue model, a new method to perform this task that does not rely on a tissue model will be presented in section 4.1. An example of the input and output of this step is demonstrated in Figure 3.3.

3.1.3 Intensity Inhomogeneity Reduction

Intensity inhomogeneity within image volumes is due to a variety of factors. Although significant measures are typically taken to reduce inhomogeneity in the field produced, some residual inhomogeneity will remain due to effects (such as radiofrequency attenuation) that are dependent on the measured object [Vovk et al., 2004], and at higher magnetic field strengths the inhomogeneity becomes more pronounced [Gispert et al., 2004]. In performing image-based correction of the intensity inhomogeneity, the inhomogeneity is often modeled as a slowly varying multiplicative spatial field. After the field is estimated, it can be used to produce an image where the inhomogeneity is reduced. Similar to the other two noise reduction steps, estimating this field would be simple given an accurate segmentation, which is one of the main motivations for the development of the Expectation Maximization methods discussed earlier. However, developing methods to estimate this field is a challenging task in the presence of large abnormalities, since they interfere significantly with the tissue models assumed by many of the most widely-used methods, in addition to the estimation of the inhomogeneity field itself. Fortunately, there exist methods that do not rely on an explicit anatomic tissue model to estimate this field (such as the method of [Sled et al., 1999] used in the INSECT system), and are thus resistant to the presence of large abnormalities. The method chosen to perform this step should be one of these methods that is free of a tissue model, and it is important that parameters are set such that the method not add additional noise by *inducing* a bias field that is not actually present. Figure 3.4 illustrates the results of an intensity inhomogeneity correction algorithm.

3.2 Registration

Registration is the process of spatially aligning two images or volumes. This is done by computing a transformation that maps each location in an input volume onto a template volume, then 're-slices' the input image such that pixels align spatially and are the same size as the corresponding pixels in the template image. This process is often computed as a linear 9-parameter affine transformation (considering translation, rotation, and scaling in three dimensions) or as a non-linear warping field. Methods exist to register images of the same modality, such as aligning T1-weighted images with other T1-weighted images. But methods also exist to align images of different modalities, such as T1-weighted images with T2-weighted images or Positron Emission Tomography (PET) images. Registration methods typically aim to compute a transformation that minimizes a measure of dissimilarity between the images, or maximizes a measure of similarity. To prevent spurious and unrealistic transformations, some registration methods use 'regularization', penalizing less likely transformations under a set of assumptions.

The input to the Registration phase will be the images produced by the Noise Reduction phase. The purpose of registration in this framework is to allow the use of multiple imaging modalities (coregistration), and the use of information derived from the alignment of a template in a standard coordinate system (template registration). Therefore, the output of the registration phase will be images in the different modalities that have been aligned with each other and have been additionally aligned with a template in a standard coordinate system.

The methods used for registration within this framework need to have several properties. First of all, they need to be fully automatic, and therefore should not rely on manually selected landmarks. The methods should also not rely on *extrinsic* markers. Although the incorporation of these markers greatly simplifies registration, it cannot not be assumed that these will be present in all images to be segmented. The registration methods should also not rely on a segmentation, or even the automatic recognition of specific landmarks, since the presence of large tumors can interfere significantly with these operations. These restrictions indicate that intensity- or information-based registration algorithms are most the appropriate, and fortunately there is an abundance of research into these methods for medical image registration. Each of the registration steps should be performed with three-dimensional volumes, even if only a subset of the slices is to be segmented. This is due to the fact that utilizing three-dimensional information allows a more accurate registration than the alignment of individual images, since the registration of three-dimensional volumes (of known scales) will be more constrained than the registration of two-dimensional images.

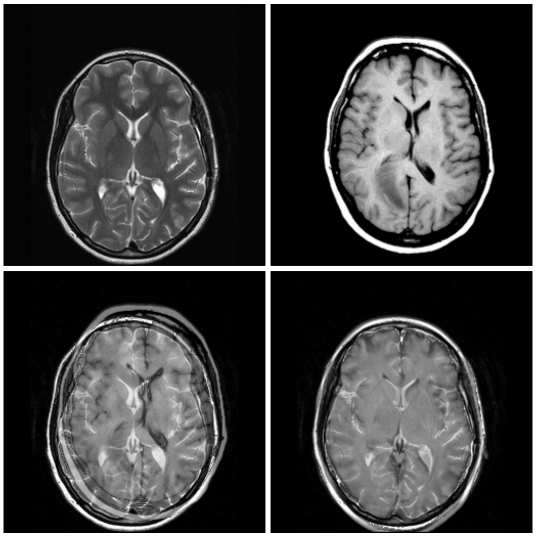


Figure 3.5: Inter-Modality Registration by Maximization of Mutual Information. Top left: T2-weighted image from individual A. Top right: T1-weighted image individual B. Bottom left: T1-weighted image from individual B overlayed on T2-weighted image from individual A before registration. Bottom right: T1-weighted image from individual B overlayed on T2-weighted image from individual A after registration by maximization of Mutual Information.

3.2.1 Coregistration

Coregistration is the task of aligning different modalities of the same patient (ie. aligning the T2-weighted image with the T1-weighted image). This step is essential if modalities are used that may not be in perfect alignment, which is often be the case with real data. However, including this step not only allows the use of modalities that are not in perfect alignment, but more generally allows the use of modalities that were not necessarily taken at the same time. This includes the registration of T1-weighted images before and after contrast injection, registration of T1-weighted images with T2-weighted images, or registration between any other two modalities of the same underlying object.

For coregistration, minimizing a dissimilarity metric based on intensities is obviously not appropriate, since the intensities of corresponding regions in different modalities may be different. Methods based on aligning labeled regions are also not appropriate, since this preceeds segmentation. Methods that seek to maximize a measure of correlation or information that is invariant to intensity differences are more appropriate, and methods based on the maximization of different forms of Mutual Information are currently popular for the task of coregistration. Mutual Information based registration has been used in several of the works discussed in Chapter 2 [Moon et al., 2002, Prastawa et al., 2004, Gering, 2003b], and is demonstrated in Figure 3.5.

We use a linear transformation for coregistration, as opposed to a non-linear warping. This is due to the fact that the same object is present in the template and input image, and thus an exact correspondence should exist without warping. The only case where non-linear registration for coregistration would be useful is if modalities were used that have different geometric distortions. In this case, it may make sense to use the modality with the least geometric distortion as the template, and perform a non-linear coregistration of the other modalities to this template. Although a high degree of regularization would be needed if non-linear coregistration is performed, this is not typically required for linear coregistration.

3.2.2 Template Registration

After aligning the images in the different modalities, the second registration step aligns the modalities with a template image in a standard coordinate system. The main purpose of this step is to

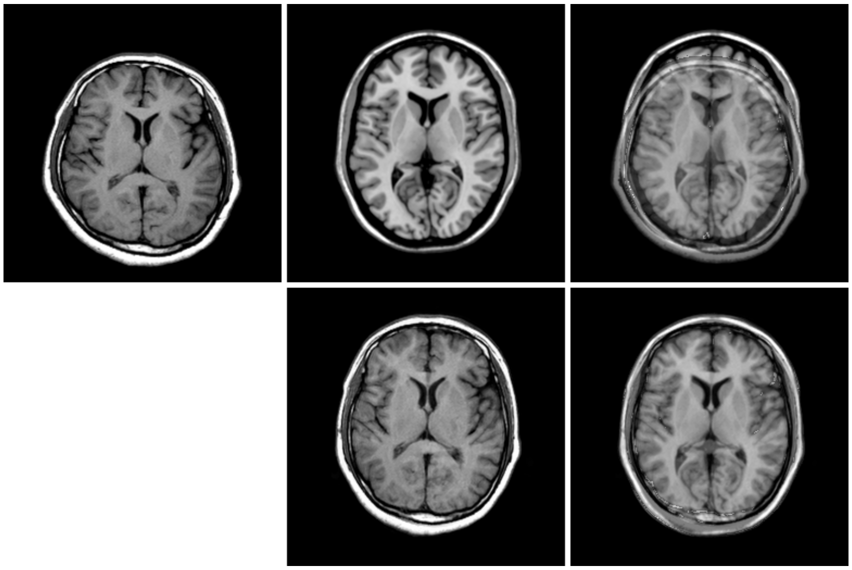


Figure 3.6: Template Registration. Top, left to right: T1-weighted image. T1-weighted template [Holmes et al., 1998], T1-weighted image overlayed on T1-weighted template. Bottom left to right: T1-weighted image after spatial registration with T1-weighted template, registered T1-weighted image overlayed on T1-weighted template.

allow the subsequent use of coordinate-based and registration-based features. However, a second purpose of this step is to standardizes the size of pixels, such that the pixel-level classifier acts only on pixels that are of the same size, even though pixels from the original images may be of different sizes. Finally, template registration will also be essential for performing inter-volume intensity normalization.

The Talairach coordinate system is the most widely recognized brain coordinate system [Talairach and Tourneaux, 1988]. However, a coordinate system such as the 'MNI' coordinate system (originally defined in [Evans et al., 1992a, Evans et al., 1992b]) that is defined based on a population rather than an individual is more appealing, since these coordinate systems are more representative of the shape of average brains, have available intensity templates in different modalities, and have available spatial prior probabilities.

The template used should have the same modality as one of the images to be segmented. Templates in the MNI coordinate system include the 'MNI305' T1-weighted average template from 305 normal individuals [Evans and Collins, 1993, Collins et al., 1994], and the more recent and higher quality 'ICBM152' data set that contains T1-weighted, T2-weighted, and ρ-weighted average images from a set of 152 'normal' individuals [Mazziotta et al., 2001]. A T1-weighted template of the average of a single individual that was imaged 27 times and registered with the MNI coordinate system is also available [Holmes et al., 1998]. Average images of a single subject or multiple subjects are both suitable for this stage, if the registration algorithm used is capable of estimating the appropriate transformation from the image to the template

Since the modalities are already aligned with each other, only a single modality needs to be registered with the template. The transformation computed to register this modality with the template can then be used to transform the other modalities. However, in cases where different modalities of the template are available, using additional modalities to estimate the transformation may confer an additional degree of robustness.

We initially perform template registration as linear affine transformation, to initialize the non-linear registration performed in the next step. This registration stage can seek to maximize measures such as Mutual Information, but can also employ strategies that minimize a measure of intensity dissimilarity. Methods based on aligning labeled regions should not be used, since recognition of

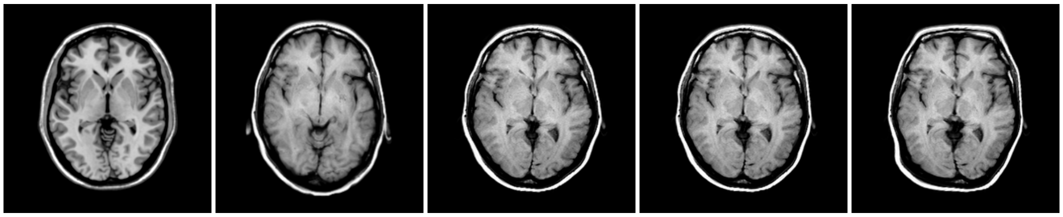


Figure 3.7: Comparison of registration with different degrees of freedom. Left to right: T1-weighted template [Holmes et al., 1998], , T1-weighted image after linear 6-parameter rigid-body registration to T1-weighted template, T1-weighted image after linear 12-parameter affine registration to T1-weighted template, T1-weighted image after further heavily regularized non-linear registration to the template image, T1-weighted image after lightly regularized non-linear registration to template image. The affine transformation provides a higher correspondence to the template than the rigid-body transformation, and although the difference is subtle, the overall correspondence with the template has been increased in the heavily regularized non-linearly transformed image compared to the affine transformation due to small corrections for overall head and brain shape, without introducing excessive and unrealistic deformations as in the lightly regularized non-linearly transformed image. It also noteworthy that the heavily regularized non-linearly registered image is the most symmetric among the transformed images.

the pathology has not yet been performed, and the template will likely not have a pathology label at the correct location (as the template is not of the same patient from approximately the same time). Since the template will not have a corresponding tumor region, regularization could be beneficial in determining the transformation. An example of template registration is shown is Figure 3.6.

3.2.3 Non-Linear Warping

After linear registration with the template, it is then possible to perform non-linear registration with the template. This step consists of finding a set of non-linear (global or local) deformations that can be applied to the images to produce ones that are more appropriately aligned. This has the effect of correcting for overall differences in head shape and reducing the anatomic variability between the images and a template, potentially increasing the degree of correspondence between image regions. This step can also be used to correct gross deformations caused by the presence of a large space-occupying abnormality, which is particularly helpful if symmetry is to be used as a feature. The main motivation for performing this step is to allow more meaningful direct comparisons to image regions in the template, and to take advantage of known locations in the template, including the bi-lateral line of symmetry.

The template used in this step should be an MRI of a single individual, rather than the average intensity over several individuals. Non-linear registration to an average template may lead to nonsensical warping as regions of higher anatomic variability in averaged images may not be representative of the expected appearance of individual images. This registration step can be used to warp the image data to match a template, or alternately to warp a template to match the image data. The primary advantage of warping the data to match a template is that it is much simpler to take advantage of symmetry (assuming a symmetric template). However, if a database of normal brains is available for comparison, it may be more desirable to warp each of the normal brains to the new image data. Fortunately, both the simple method of using symmetry and a database of normal brains for comparison can be used, if the image data is first registered non-linearly with a template, and then each of the normal brains is registered to the input data or to the template.

Non-linear registration has typically not been used for the segmentation of abnormalities. This is primarily due to the fact that non-linear registration with large abnormalities is a much more difficult problem than linear registration. Linear registration is relatively robust to the presence of abnormalities, since the goal is to maximize global correspondence with simple geometric transformations,

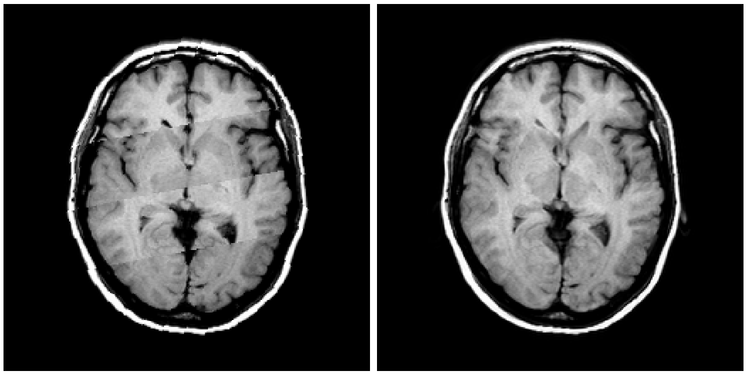


Figure 3.8: Comparison of a naive and an effective interpolation method. Left: Nearest Neighbor spatial interpolation after template registration. Right: High degree polynomial β-spline interpolation from the same original data and transformation. It is noteworthy that this volume was not corrected for inter-slice intensity variations, clearly visible in the left image (although they can be seen to a lesser extent in the right image).

and the image will be primarily comprised of normal regions that will have a high correspondence with the template. However, non-linear registration allows transformation that can improve local correspondence without global constraints, and thus local regions that have low correspondence with template regions can cause problems, since it is possible to grow or shrink abnormal regions to achieve a higher correspondence. For this reason, non-linear registration is left as an optional step in this framework, since it can be argued that it may cause undesirable effects in the quantitative results. If non-linear registration is used, a high degree of regularization is required, such that significant non-linear transformations are penalized heavily. Under a high degree of regularization, non-linear warping is advantageous since it can allow a more effective use of many of the features that use template information. A comparison of linear registration and a highly regularized non-linear registration is shown in Figure 3.7.

3.2.4 Spatial Interpolation

After computing the spatial transformation mapping one image to another, the final stage of spatial registration is Spatial Interpolation or ‘re-slicing’, since the spatial transformation computed can change the size and location of the pixels in the transformed image such that they do not correspond to the sizes and locations of pixels present in the original image. Spatial Interpolation uses the locations and intensities of the ‘old’ pixels after spatial transformation, to compute the intensity values at the ‘new’ pixel locations. In general, methods that enforce smoothly varying image derivatives (such as splines) should be preferred over methods that guarantee only local smoothness (or do not enforce smoothness at all). A visual comparison of a method that guarantees smoothly varying derivatives compared to one that does not guarantee smoothness at all is seen in Figure 3.8. Spatial Interpolation should only be performed after the transformation parameters for each of the registration steps have been computed, since performing intermediate interpolations may introduce unnecessary errors.

3.3 Intensity Standardization

Intensity Standardization is the vital step that allows the intensities to be used in a supervised classification framework, without requiring patient-specific training to account for differences in the meaning of the intensities acquired. The goal of this step is to convert the intensities of the input data to an intensity distribution where the values of the intensities have an approximate anatomical meaning. We use a template-based approach, since histogram-based and model-based approaches may not be appropriate if large abnormalities are present (due to interference with the histogram dis-

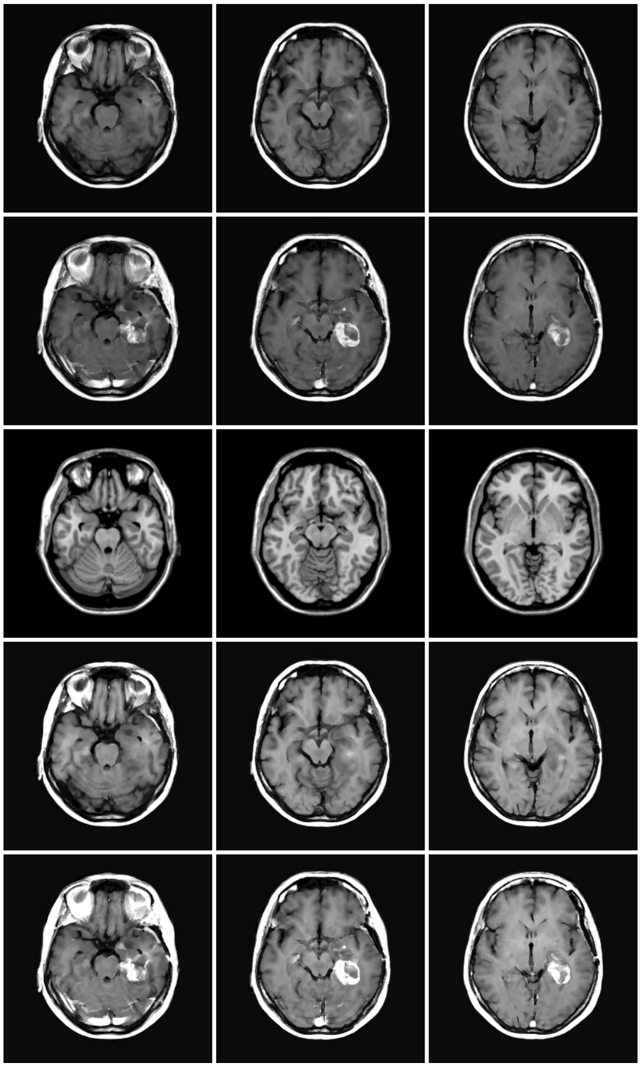


Figure 3.9: Template-based intensity Standardization. First row: T1-weighted images after noise reduction and spatial registration. Second row: T1-weighted post-contrast injection images after noise reduction and spatial registration. Third row: T1-weighted template used for standardization. Fourth row: T1-weighted images after intensity standardization. Fifth row: T1-weighted post-contrast injection images after intensity standardization. Although clearly not perfect, the intensity differences between similar tissue types have been decreased significantly.

tribution and model estimation). Although template-based approaches can also be affected by large abnormalities, we use a template-based measure of symmetry to make the intensity standardization robust to the presence of large (asymmetric) abnormalities.

Template-based methods perform registration as a preprocessing step, and estimate a transformation between the intensities in the image and the intensities in the template, based on pixels at corresponding locations. The advantage to this type of approach is that spatial information is used, in addition to the intensity distribution used by the other two approaches. The INSECT system utilizes a template-based approach for its Intensity Standardization step [Zijdenbos et al., 1998], estimating a linear intensity transformation between the input image and the template image based on regions of similarity. The single global scale factor computed in this system is applied to the original images to generate intensity standardized images. There are simple modifications that can be made to make the method more robust to large abnormalities (such as regularization or outlier modeling), and the use of spatial information offers a major advantage over methods that rely only on the intensity distribution. Thus, template-based methods represent the most appropriate type of Intensity Standardization technique for this framework. The inputs and outputs of a template-based Intensity Standardization technique are shown in Figure 3.9.

3.4 Feature Extraction

After Noise Reduction, Registration, and Intensity Standardization, the fourth stage in the framework is the calculation of the features that will be used in pixel classification. Features that have been used in previous works include intensities and textures [Dickson and Thomas, 1997], intensities and 'distances to labels' [Kaus et al., 2001], intensities and spatial tissue prior probabilities [Prastawa et al., 2004], and multi-resolution symmetry of the intensities [Gering, 2003a]. These feature combinations comprise a very limited characterization of a pixel within an image that has a relatively known structure. The main advantage of our framework is that it can learn to simultaneously use intensities, textures, 'distances to labels', spatial prior probabilities, and symmetry. Furthermore, we can use a variety of other features in this framework, such as features based on histogram analysis, anatomic variability maps, template comparisons, and voxel morphometry. This section will give an overview of the different types of features that can be used, while Section 4.4 will discuss specific implementation details for incorporating many of these features.

The main consideration when selecting features is that the features used should reflect properties measured at the pixel-level that can aid in discriminating between normal pixels and tumor pixels. However, there does not necessarily need to be an obvious (or even linear) relationship between the features used and the likelihood of a pixel representing a tumor, since the classification stage will learn an appropriate means of combining the features to perform the task. For example, the spatial prior probability for CSF is almost meaningless in discriminating normal and tumor pixels, since tumors can occur in regions of both high and low probability. However, combining this feature with the T2 image intensity can be used for much more effective discrimination, since the classifier could learn, for example, that a high T2 intensity and a low CSF probability are indications of tumor. There are four main sources for the generation of useful features:

1. Image-based Features: Features that can be calculated directly from the image data.

2. Coordinate-based Features: Features that take advantage of a standard coordinate system.

3. Registration-based Features: Features that use one or more (non-linearly) registered templates (including taking advantage of known properties of these templates such as labels or the location of the line of symmetry).

4. Feature-based Features: Features formed from subsets or combinations of other features.

3.4.1 Image-Based Features

Image-based features can be used to represent various properties of pixels and there neighborhoods. This is the type of feature most commonly used in brain tumor segmentation, and only recently have other types of features begun to be explored. The most obvious pixel-level feature of this type is the (standardized) pixel intensities from each modality.

Another set of features in this category are texture features. Texture features consist of a set of calculations that can characterize patterns in the intensities of the region containing a pixel, and have been used previously for tumor segmentation [Schad et al., 1993, Busch, 1997]. Another method that has also been discussed to characterize pixel regions is the inclusion of the intensities of neighboring pixels as additional features (as in [Dickson and Thomas, 1997, Garcia and Moreno, 2004]). In this case texture would not be modeled explicitly by computed texture features, but would be

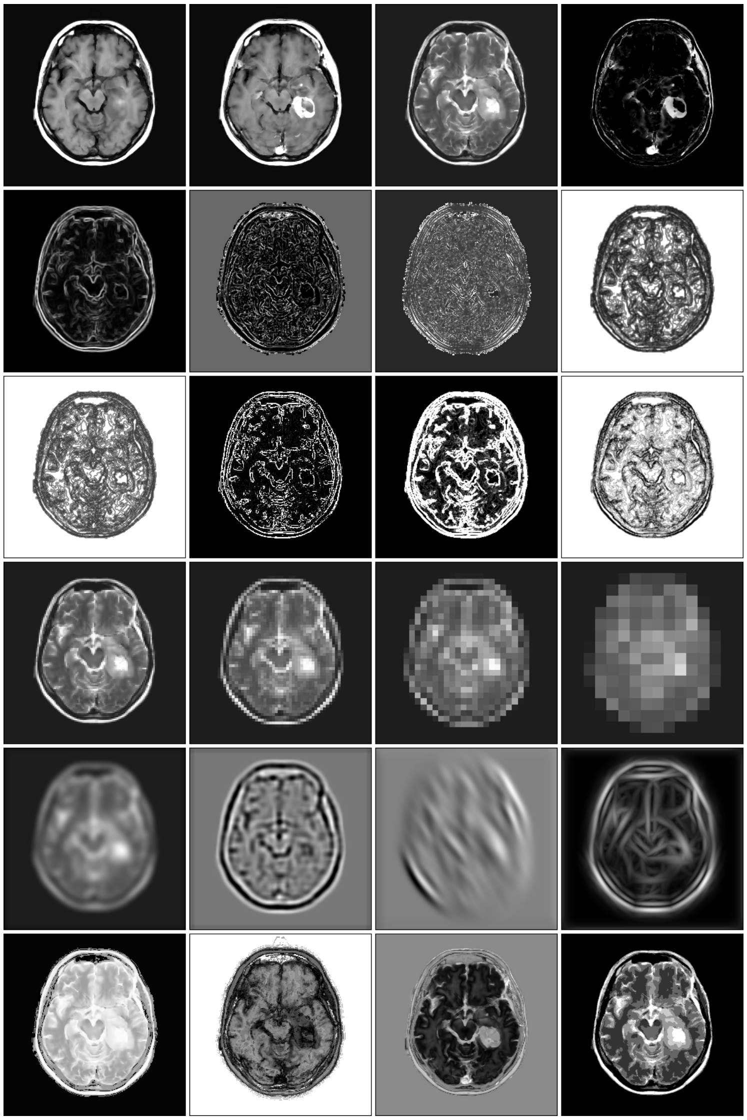


Figure 3.10: Examples of Image-Based Features: First row: Intensity Standardized intensities. Left to right: T1-weighted, T1-weighted image with a contrast agent, T2-weighted, contrast agent difference image. Second row: First order textures of T2 image. Left to right: Variance, skewness, kurtosis, energy. Third row: Second order textures of T2 image. Left to right: angular second momentum, cluster shade, inertia, local homogeneity. Fourth row: Four levels of a multi-scale Gaussian pyramid of the T2 image. Fifth row: Linear filtering outputs from the T2 image. Left to right: Gaussian filter output, Laplacian of Guassian filter output, Gabor filter output, Maximum Response Gabor Filter output. Sixth row, left to right: T2 Intensity percentile, multi-spectral (log) histogram density, multi-spectral distance to the template's average white matter intensities, unsupervised segmentation of the T2 image.

modeled implicitly by the classifier that will learn a method to combine the intensities in the neighborhood. An alternate method to characterize pixel neighborhoods is to measure features at multiple scales. The advantage of using multi-scale features is that fine details are represented in the fine scale features, while the coarse scale features represent gross neighborhood properties that are not as susceptible to noise.

Several other potentially useful image-based features can be inferred by examining previous systems for tumor segmentation that did not employ a supervised approach. The rule-based system in

[Clark et al., 1998] makes significant use of relative intensities. Relative intensities could be incorporated through simple histogram-based calculations. The rule-based system of [Clark et al., 1998] also uses a method called 'density screening', that assessed how many pixels in the image have similar multi-spectral intensity values, with the intuition that tumor pixels will be present in less dense areas of the multi-spectral intensity space than normal pixels. Although this makes an assumption that tumors represent intensity outliers and thus is not generally applicable, it could be used in this framework due to the fact that it represents a potentially useful feature for detecting tumor pixels that are intensity outliers, but its use does not necessitate that tumor pixels have to be intensity outliers, since other features can also be used to detect these cases. Another image-based feature, related to the outlier measure in [Gering, 2003b], could be to measure the distances from a pixel's multi-spectral intensities to the expected multi-spectral intensities of different normal tissue types in the template intensity distribution.

Another source of image-based features is computations applied to an unsupervised segmentation that divides the image into homogeneous regions. The idea in this case would be to calculate an unsupervised (hierarchical) image segmentation, and use properties of the regions that include the pixels as additional features of the pixels. This could include a structure 'thickness' evaluation similar to the one used in [Gering, 2003b], but could also include the volume of the resulting region or the distance that it extends from the pixel. Other features that could be calculated based on an unsupervised segmentation include the features used in the second Neural Network in [Dickson and Thomas, 1997], including simple characterizations of shape.

The different image-based features discussed in this section can be summarized, with several example features, in four broad categories as follows:

1. Intensity-based: A pixel's intensity in each channel, the intensities of the pixel's neighbors, and aggregations over the intensities in the pixel's neighborhood.

2. Texture-based: Explicit calculations of texture features based on a pixel's neighborhood.

3. Histogram-based: The intensity percentile of the pixel within the histogram, the multi-spectral distances to the intensities of normal tissues, and the number of pixels close to the pixel's intensities in the multi-channel intensity space.

4. Structure-based: After performing a hierarchical unsupervised segmentation, computing size and shape characteristics of the structures that the pixel is assigned to.

3.4.2 Coordinate-Based Features

We can also exploit features in this framework are features that are derived from the use of a standard coordinate system. Discussed extensively already are the spatial prior probabilities for gray matter, white matter, and CSF used in the Expectation Maximization approaches to tumor segmentation and the INSECT system. These features represent, at each pixel in the coordinate system, an empirical measure of the distribution of these tissue types in normal brains that are registered into the coordinate system. The major advantage of using this type of feature is that it encodes spatial information into the classification, which is not represented well by image-based features alone.

An obvious set of coordinate-based feature, other than the three spatial priors already seen, are priors for other structures. One useful prior for segmenting structures in the brain is a 'brain mask' prior. This obviously represents a useful feature for brain tumor segmentation, since it can be used to distinguish tumor tissue from tissue outside the brain that may have similar characteristics. The 'brain mask' prior is thus an excellent example of how a coordinate-based feature can enhance classification through a simple encoding of spatial information. The use of a probabilistic brain mask has the additional advantage that an explicit brain masking operation, that is subject to its own segmentation error, does not need to be applied prior to classification (as is done in the INSECT system). In some coordinate systems, spatial priors are available for a large variety of structures

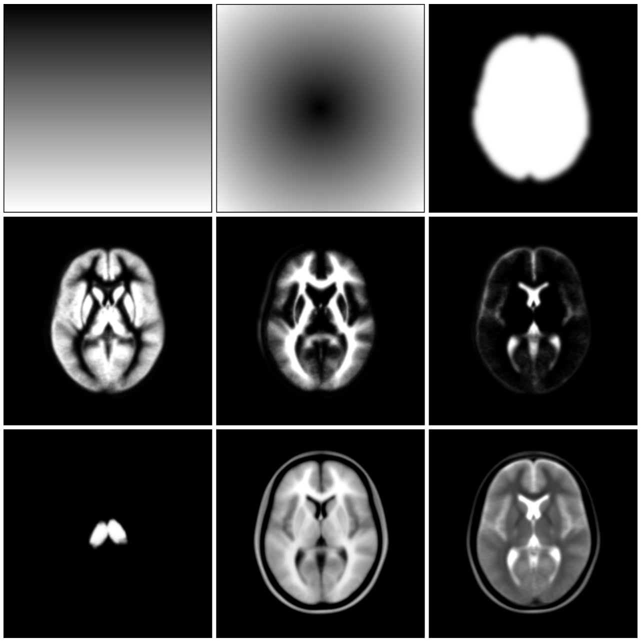


Figure 3.11: Examples of Coordinate-Based Features: First row, left to right: y-coordinate, distance to image center, brain mask prior probability [SPM, Online]. Second row, left to right: gray matter prior probability, white matter prior probability, CSF prior probability [ICBM View, Online]. Bottom row, left to right: thalamus prior probability [Mazziotta et al., 2001], average T1 intensity from a population [ICBM View, Online], average T2 intensity from a population [ICBM View, Online].

and tissue types. It is not recommended that all available priors be used, since there may not be sufficient training data available to provide examples of normal and abnormal structures at each specific spatial location. However, an additional prior that may be useful in addition to the three normal tissue priors and the brain mask prior would be a gray matter nuclei prior. This is due to the fact that gray matter nuclei can exhibit different intensity characteristics than cortical gray matter [Studholme et al., 2004]. Another important note on the use of spatial priors is that, if a prior for a relevant structure or tissue type is not available, spatial priors can be approximated from small amounts of empirical data using methods such as the one in [Joshi et al., 2003]. Another use of spatial priors could be intensity-based priors, rather than tissue or structure based priors. These priors would consist of the average intensity in a set of aligned brains, as in the template described in [Evans and Collins, 1993], and could aid in distinguishing pixels that are outliers based on both their intensity and spatial position, but not necessarily their intensity alone.

Another possible set of coordinate-based features are the actual x, y, and z coordinates. In general, these should not be used since spatial priors can encode a similar type of information with better generalization properties when classifying unseen test data. However, as was shown in [Dickson and Thomas, 1997], coordinate features can aid in the segmentation of highly localized tumors, and may be more appropriate than spatial priors in these cases, if sufficient labeled training data is available.

A third set of coordinate-based features are empirical measurements of the range of anatomic variability at locations within the coordinate system. The intuition behind including a feature such as this is that it may aid in discriminating between abnormal tissue and normal anatomic variation, since small changes in conserved regions are more significant than small changes in highly variable regions. The simplest method to incorporate this type of information would be to include the variance measured in constructing a spatial intensity prior. However, there are also more sophisticated methods than this exist. [Toga et al., 2003] discuss an extensive variety of probabilistic brain atlases, many of which could be used for this purpose, or to add additional coordinate-based (or registration-based) features.

The different types of coordinate based features can be summarized as follows:

1. Spatial Priors: Probabilities at each pixel in the coordinate system for different tissue types, structures, or intensities.

2. Coordinates: Cartesian or Spherical pixel coordinates, useful for segmenting highly localized structures when large training sets are available.

3. Variability Maps: Measures of normal anatomic variation at each pixel in the coordinate system, or measures of the significance of the measured anatomic variation over an image region.

3.4.3 Registration-Based Features

A third set of features are those that can be computed based on spatial registration with a normal brain. These features take a variety of forms, since there are many ways in which the registration and alignment information can be used. Although these features do not strictly require non-linear registration, they rely on the local template correspondence to a greater extent than coordinate-based features, and thus will be enhanced if an effective non-linear registration method is used. It is possible to use more than one registered normal brain to compute these features. If more than a single normal scan is used, the information from each scan could be averaged, the information from each scan could be used as a feature, or only the closest (or furthest) scan could be used.

The only registration-based feature previously used has been the ‘distance transform’, that depends on using a template with labeled regions. This feature computes the distance from a pixel to the location of a specific labeled region in the template (ie. how far is this pixel from the location of the labeled ‘brain’ area in the template). As discussed in Chapter 2, this feature was used in [Kaus et al., 2001] to improve intensity based classifications, by adding this simple form of spatial context. Another strategy to include information based on template labels is utilizing the template labels directly as pixel features, by expanding them into a set of binary features. These binary features that represent the exact locations of template labels could be smoothed to reflect uncertainty in the exact locations of structures. As with spatial priors, template labels should be used sparingly unless a large training set is available or the tumor type being segmented is highly localized. Labels for tissue types that are present throughout the brain should be preferred over labels of individual structures. One interesting use of label-based features could be the incorporation of text-based diagnostic information on the approximate tumor location. If the approximate brain region of the tumor is known, a patient specific approximate spatial prior probability could be constructed using template labels. A final use for label-based features would be to incorporate segmentations from earlier timepoints of the same individual (in this case the level of smoothing could be based on the date of the earlier timepoint), or to include other segmentations of the same patient (for example, segmenting edema is simpler than segmenting the GTV, but the GTV is often a subset of the edema and thus an edema segmentation would form a useful feature for the segmentation of the GTV).

A more rich source of registration-based features than template labels are the actual image properties at corresponding locations in the template. The abnormality metric from [Gering, 2003a] could be used as a feature of this type. This metric computed, at each pixel, the average intensity difference between the input image and the closest template image over three square window sizes. These three average absolute values were multiplied for each pixel location, and a threshold was selected to measure whether this was sufficiently abnormal. There are a large amount of variations or extensions to this type of method for using the template image information. For example, the naive combination rule could be removed and the measure computed at the different window sizes (ie. at multiple scales) could each be used as features. Other variations include using the template intensity information directly as additional features, assessing texture parameter differences, or computing a correlation or information-based measure such as Mutual Information.

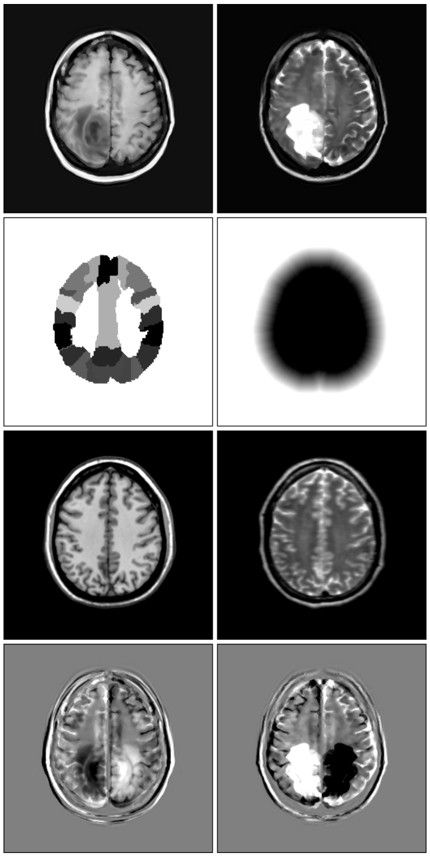


Figure 3.12: Examples of Registration-Based Features: First row, standardized and registered image data for visual comparison. Second row: Labels of normal structures in the template (left) [Tzourio-Mazoyer et al., 2002], distance transform to template brain area (right). Third row: Template image data at corresponding locations (note the much higher similarity between normal image regions than abnormal regions). Fourth row: Symmetry of the T1-weighted (left) and T2-weighted (right) image by using the template's known line of symmetry. The symmetry images were scaled, black areas indicate that the location is darker than the contra-lateral location, while white areas indicate that location is brighter than the contra-lateral location.

Registration-based features can also be derived from the registration process itself, in the case of non-linear registration. For many non-linear registration techniques, a 'warping' field can be produced that measures, at each pixel, the distance that the pixel was deformed from its original location (see [Ashburner and Friston, 2003a] for a discussion). The warping field could be a useful feature if a warping algorithm is used that tends to warp tumors in a relatively consistent way (for example, if tumor pixels often get displaced more than normal pixels). Alternately, if a non-linear algorithm is used that is constrained to force an exact match between the image and template, then the warping field could represent a useful feature for segmenting tumors since gross deformations would be required to grow or shrink the tumor into structures present in the normal template, compared to normal regions where gross deformations would not be needed. Unfortunately, this last feature is problematic since the registration task would be computationally complex, and a deformation model complex enough to transform tumors into normal tissue would be subject to many local optima. Another use for warping fields is the assessment of anatomic variability, since many available probabilistic atlases consider this type of information [Toga et al., 2003].

It is obvious that bi-lateral symmetry is an important feature, as normal brains have a large degree of symmetry [Joshi et al., 2003], while abnormalities are typically asymmetric. Unfortunately, assessing symmetry is not a trivial computation, and even automatically locating the mid-saggital axis of symmetry in MR images of the head has been the subject of recent research [Liu et al., 2001]. The use of symmetry with respect to brain tumors is further complicated by the fact that large tumors can deform the brain to such as extent that the 'line' of symmetry has become curved, making this axis more difficult to locate and symmetry more difficult to assess. Symmetry has been included as a registration-based feature rather than an image-based feature, since registration can be used to allow a simple and straightforward assessment of symmetry. This is due to the fact that the line of symmetry is known in the template, and therefore template registration can be used to approximately locate the line of symmetry. If the axis of symmetry in the template is parallel with one of the image axises, then a simple characterizations of symmetry can be made by mirroring the image around this axis. Pixel-level features can then be computed based on the corresponding mirrored region.

Even in cases where the line of symmetry has moved or changed shape, there exist methods to approximately recover symmetry. One way to do this is to use non-linear registration, since it can perform a small degree of 'unwarping' of tumor-induced deformations, in order to more appropriately align with the template. If residual differences need correcting after this step, then the template axis of symmetry could be used to initialize a search for the true axis. If the axis used is not a straight line, it is still possible to use symmetry based features, but their computation is slightly more complicated.

The different types of registration-based features can be summarized as follows:

1. Label-based Features: Features computed based on labels assigned to corresponding regions in a template (including available segmentations of the same patient).

2. Image-based Features: Features computed based on image properties at corresponding regions in the template.

3. Warping-based Features: Features computed based on the transformations used in non-linear warping.

4. Symmetry-based Features: Symmetry features computed by taking advantage of the the template's known axis of symmetry.

3.4.4 Feature-Based Features

After calculating the different features, an additional Feature Extraction stage can be used to construct a feature set that is more suitable for high classification performance. This feature processing stage consists of adding features that encode spatial context beyond pixel level measurements, performing feature selection, and/or performing dimensionality reduction.

The first and one of the most important of these feature processing steps is the addition of features that take into account a pixel's spatial context, through region-based features. Many of the features discussed in this section represent pixel-level measurements, ignoring the feature values of neighbors. The problem with simply using pixel-level coordinate- or registration-based features is that most classifiers make the assumption that the pixels to be classified are independent of their neighbors, and thus also ignore the values of the pixel's neighbors. Although computationally complex classifiers can be chosen that do not make this assumption, it is simple to construct features that will take additional regional context into account (these features will improve both classifiers that do and do not make the independence assumption). In the section on Image-Based features, several methods were discussed that can be used to represent additional regional context, including the use of texture parameters, neighboring pixel intensities, neighborhood aggregations, measuring features at multiple resolution, performing histogram analysis, or making calculations based on data clusterings. Since coordinate-based and registration-based features fundamentally encode a pixel-level measurement, these features measured at each pixel in the image can be constructed into an image, where each of the Image-Based methods to take into account regional context can be applied.

Another important feature processing stage is feature selection. This section has given an overview of many possible features that can be used, and we would like to use a feature set that encodes a diverse variety of types of information in order to enhance discrimination between normal and abnormal areas. However, when large and complicated feature sets are used to train on smaller training sets, classifiers can ‘over-fit’ the learned model, since it is likely that spurious patterns can be found that can accurately classify the training data, but are not relevant to unseen test data. Feature selection is partially up to the designer to select an appropriate feature set, but automatic methods can also be used. In selecting features, it is important to consider whether features will help in discriminating unseen data, and how complicated the interactions between the features are likely to be in order for them to be used in discrimination. An example of a feature subset that will not likely help in classifying unseen data is the inclusion of hundreds of template labels of anatomic and functional structures. If only a small number of training images are available, each of the structures will only have a few (or no) training instances where a tumor was actually present. This means that, for example, if a classifier never sees an example of a tumor present in pixels with the label of ‘temporal lobe’, it may learn a model that will never classify pixels with this label as tumor. It would be more logical to leave out labels such as this that do not aid in classifying unseen data, and use tissue types such as gray matter, since the model could use training data from gray matter pixels from other parts of the brain to decide if a pixel in the temporal lobe represents a tumor. Another example of a feature subset with interactions that may be too complicated to aid in discrimination with smaller training sets would be the extraction of a large amount of texture parameters and the inclusion of the pixel intensities in a large neighborhood around the pixel of interest. This feature subset is large and complex with a significant amount of non-informative information, since the exact intensities of distant pixels and many of the texture parameters may not aid in discrimination. It would be more appropriate to select a small set of good texture features, and use neighborhood aggregations of the neighborhood intensity data rather than including each individual intensity value. After an initial selection of potentially good features, an automated feature selection algorithm can be used to further narrow down the feature set to the most relevant features.

Another method of improving the performance of a feature set is dimensionality reduction. Dimensionality reduction methods construct a new feature set from the set of existing features, where the new feature set potentially has a lower total number of features, and a transformation can be applied to the new feature set to regain most of the information contained in the original features. The advantage of dimensionality reduction is that it can reduce redundancy and correlation in the features, resulting in a feature set that may be more likely to achieve high classification performance. Principle Component Analysis is one example of a linear dimensionality reduction method, a discussion of this method and a list of non-linear methods is contained in Chapter 3 of [Gering, 2003b].

3.5 Classification

After the feature set has been computed for each pixel, the feature set will then be used by a classifier to decide whether each pixel represents a tumor pixel or a normal pixel. The classification stage has two components, a training phase and a testing phase. In the training phase, pixel features and their corresponding manual labels represent the input, and the output is a model that uses the features (of a new voxel) to predict the corresponding label. This training phase needs to be done only once, since the model can then be used to classify new data. The input to the testing phase is a learned model and pixel features without corresponding classes, and the output of the testing phase is the predicted classes for the pixels based on their features. An example of the output of the testing phase is demonstrated in Figure 3.13.

Chapter 2 discussed a variety of classifiers including kNN, Decision Trees, Maximum Likelihood, Neural Networks, Ensemble Methods, Support Vector Machines, and Markov Random Fields. It was also discussed that methods that can learn non-linear dependencies in the features have experimentally outperformed those that do not. A noteworthy point is that any classifier can encode non-linear dependencies in the features through a change of basis (ie. using non-linear combinations of the features as additional features), but classifiers that can learn non-linear combinations without

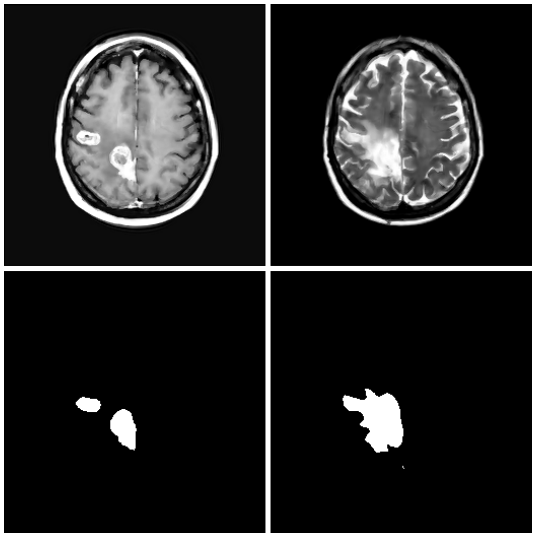


Figure 3.13: Classifier Output. Top: T1-weighted post-contrast injection (left) and T2-weighted image (right). Bottom: Classifier predictions for the 'Enhancing Tumor' class and the 'Tumor and Edema' class.

a change of basis are appealing since larger feature sets become computationally expensive to use. If multi-scale features are used, methods that embed the features in a (Euclidean) space and use distance metrics based on all of the features have the appealing property that neighboring pixels will tend to receive the same label, due to their high similarity in the coarse scale features. Although this property makes the classifier's predictions less noisy, it is not a necessary property for the classifier, since the next stage of the framework is a relaxation step that will remove noise in the segmentation. The time required for training is also an important factor in choosing a classifier, since each image has a large number of pixels and training times can grow large as feature sets and the number of training pixels increases. If the system is to be used in a semi-automatic way, where the user can input additional training pixels where the system has made mistakes, then a classifier that can be trained incrementally should be used to prevent the need to completely retrain.

For most classifiers, assigning classes based on a model is computationally efficient, while initially learning the model can be computationally intensive. Sub-sampling (using a subset of the full training data) is one method to ease the computational costs of the training phase, if the time needed to learn the model is prohibitively large. Although random sub-sampling can be used, spatial information can be used to produce a more strategic sub-sampling, that will not degrade the quality of the learned model to the same extent as random sub-sampling. An obvious non-random sub-sampling strategy that uses spatial information is to sub-sample proportionally to the pixel's prior probabilities of being part of the brain mask, since few pixels outside the brain will be needed in training (assuming that a brain mask prior probability is used). Non random sub-sampling using spatial information could also be used to sub-sample normal areas that have large distances from tumor pixels, since these should exhibit fairly typical behavior and will likely not help significantly in learning a model that appropriately classifies ambiguous instances.

3.6 Relaxation

Most classification techniques evaluate each pixel independently of the values of their neighbors, based on the pixel's features. In classifying a pixel as tumor or not, the classifier thus does not consider the classification of a the pixel's neighbors. Since the classifier may not be perfect at classifying unseen test pixels, the classifier may make errors. The purpose of the Relaxation stage is to smooth the classifier's output by considering dependencies in the labels of neighboring pixels,

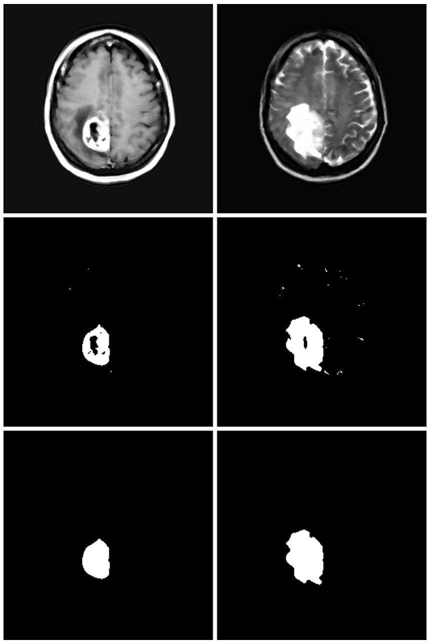


Figure 3.14: Relaxation of Classification Output. Top row: Image data. Middle: An example of predictions made by a noisy classifier. Bottom: The noisy classifier output relaxed using morphological operations that take into account the labels of neighboring and connected pixels.

which will correct pixel misclassifications by considering this additional spatial context. The algorithm needed for this stage will depend on how accurate and noisy the classifier is, since in some cases only minimal correction will be needed, while others may need more sophisticated methods. The strategy used for this step could thus range from simple methods including low-pass filtering or morphological operations applied to the classifier output, to more complicated methods including Markov Random Fields and Level Sets.

3.7 Post-Processing

Depending on the application that the segmentation is being used for, post-processing stages may be required after Relaxation. If the original (and not registered) images are to be labeled, this would include applying the reverse spatial transformations to the segmentation map to produce a segmentation map for the original image. Post-processing could alternately consist of simply unwarping the non-linear registration component, making the images suitable to be registered with a CT image or new MR images. Post-processing could also include steps such as segmenting the normal tissue classes, a task that would be greatly simplified since the abnormal pixels could be removed from the tissue class estimation.

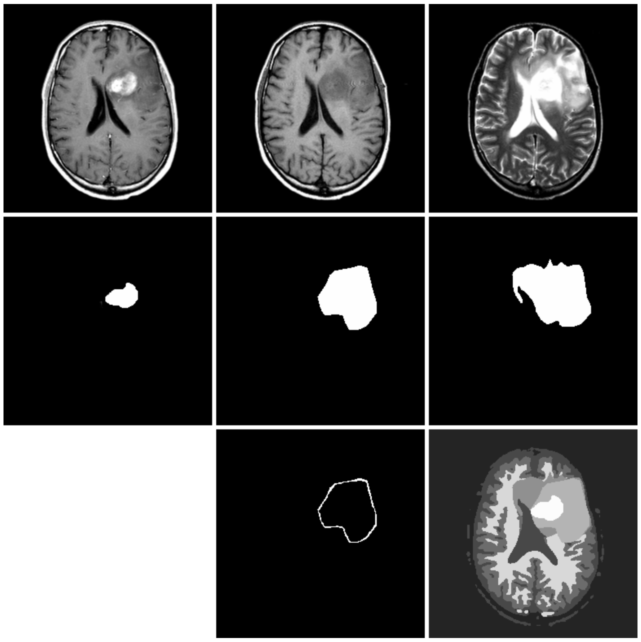


Figure 3.15: Examples of possible post-processing operations. Top row: Multi-spectral image data. Middle row: Segmentations for three different pathological areas. Bottom middle: Gross tumor contour obtained from gross tumor segmentation. Bottom right: The pathology segmentations combined with the results of an Expectation Maximization segmentation algorithm where the abnormal areas were removed from tissue class estimation.

Chapter 4

Instantiation of the Automatic Segmentation Framework

The previous chapter outlined the automatic segmentation framework, and presented the overall purpose and general types methods that are suitable for each step. This chapter will present our specific implementation of this framework, discussing the methods used for each step in the framework (shown in Figure 4.1). This chapter will also discuss techniques that were not explored but could be incorporated into future implementations of the framework to potentially improve results.

4.1 Noise Reduction

The implemented noise reduction stage consists of four steps: two dimensional local noise reduction, inter-slice intensity variation correction, intensity inhomogeneity correction, and finally a three-dimensional local noise reduction step. The methods used follow the guidelines outlined in the previous chapter; They do not require a tissue model or segmentation, and they perform only a small degree of correction to prevent the introduction of additional noise, rather than attempting to determine an optimal correction.

4.1.1 Local Noise Reduction

There are a multitude of methods for reducing the effects of local noise from images. [Smith and Brady, 1997] survey a diverse variety of techniques to perform this task, and a small subset will be examined in this section to motivate our selection. The main assumption underlying most local noise reduction techniques is that noise at a specific pixel location can be reduced by examining the values of neighboring pixels. The algorithms in this section will be discussed with respect to two dimensional image data, but each has a trivial extension to three dimensions.

A simple method of noise reduction is *mean filtering*. In mean filtering, noise is reduced by replacing each pixel's intensity value with the mean of its neighbors, with its neighbors being defined by a square window centered at the pixel. A more popular method of noise reduction is through *Gaussian filtering*. This method is similar to mean filtering, but uses a weighted mean. The weights are determined by a radially symmetric spatial Gaussian function, weighing pixels proportional to their distance from the center pixel. The result of this filtering operation is described at a pixel level by Equation 4.1 (with $I(x, y)$ being the intensity value at location (x, y), and σ being a parameter of the filter).

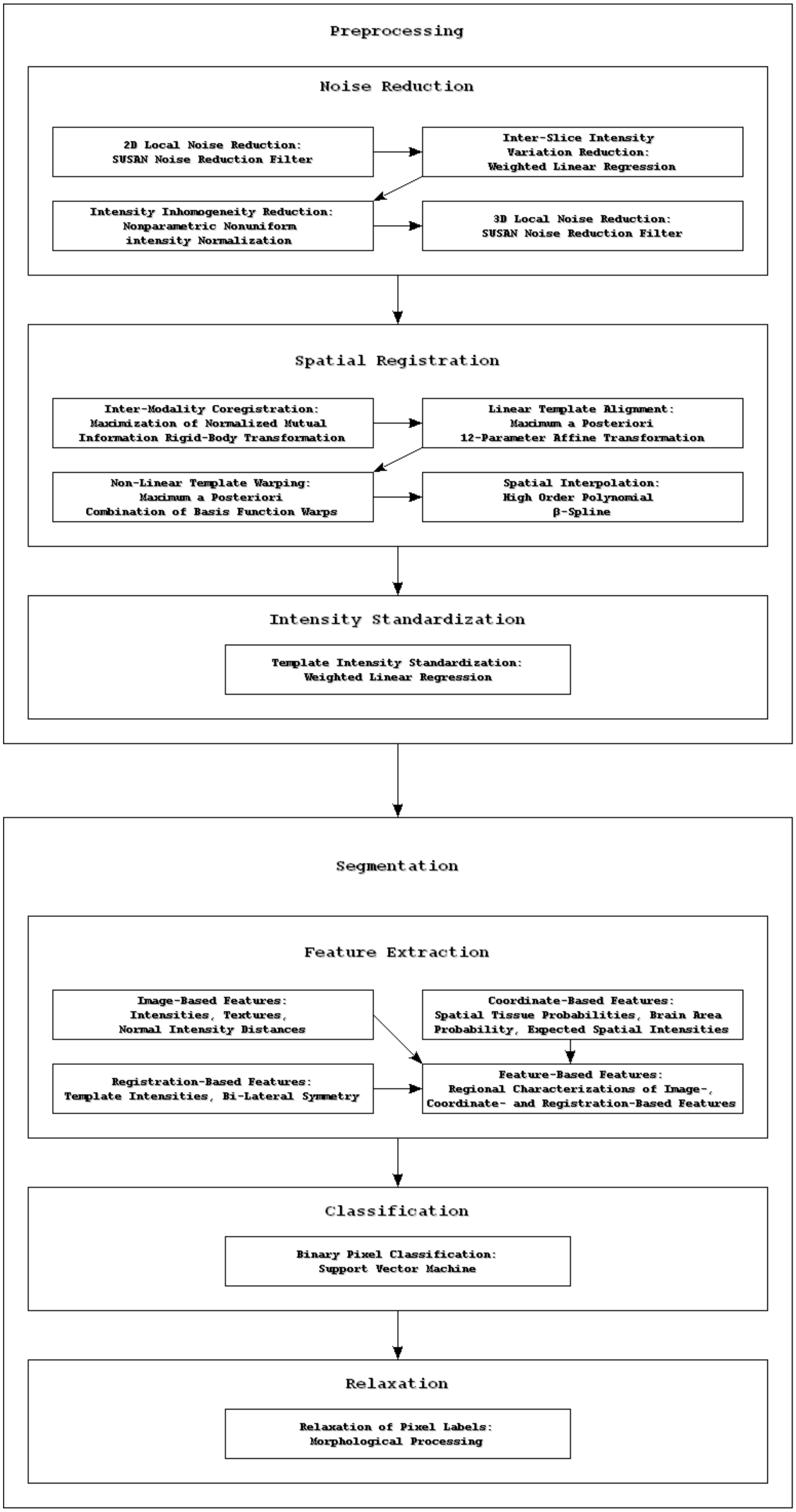

Preprocessing
Noise Reduction
2D Local Noise Reduction: SUSAN Noise Reduction Filter
Inter-Slice Intensity Variation Reduction: Weighted Linear Regression
Intensity Inhomogeneity Reduction: Nonparametric Nonuniform intensity Normalization
3D Local Noise Reduction: SUSAN Noise Reduction Filter
Spatial Registration
Inter-Modality Coregistration: Maximization of Normalized Mutual Information Rigid-Body Transformation
Linear Template Alignment: Maximum a Posteriori 12-Parameter Affine Transformation
Non-Linear Template Warping: Maximum a Posteriori Combination of Basis Function Warps
Spatial Interpolation: High Order Polynomial β-Spline
Intensity Standardization
Template Intensity Standardization: Weighted Linear Regression
Segmentation
Feature Extraction
Image-Based Features: Intensities, Textures, Normal Intensity Distances
Coordinate-Based Features: Spatial Tissue Probabilities, Brain Area Probability, Expected Spatial Intensities
Registration-Based Features: Template Intensities, Bi-Lateral Symmetry
Feature-Based Features: Regional Characterizations of Image-, Coordinate- and Registration-Based Features
Classification
Binary Pixel Classification: Support Vector Machine
Relaxation
Relaxation of Pixel Labels: Morphological Processing

Figure 4.1: Overview of the implementation of this framework.

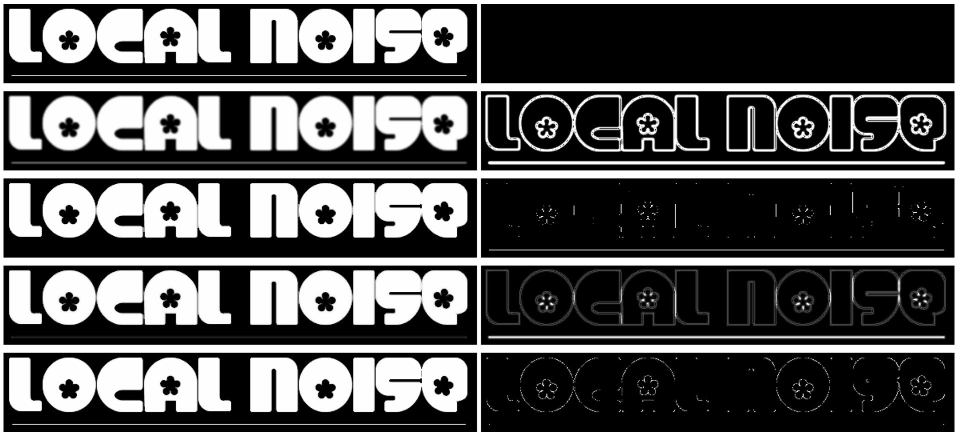


Figure 4.2: Noise reduction results on a toy image of letters and an underline. Left, top to bottom: Original Image, Gaussian filtering, Median filtering, ADF, and SUSAN filtering, respectively. Right: Absolute difference between noise reduced images and the original noise-free image for the different techniques (multiplied by 10, darker is better). Notice that noise has primarily been introduced at edges, and is most prevalent in the Gaussian filtering and ADF methods. Only the original image and the SUSAN filtered image have fully preserved the underline. In this case, the Median filter has altered the letter's edges the least, and the SUSAN filter has most effectively preserved the inter-letter area (with the exception of the original unfiltered image).

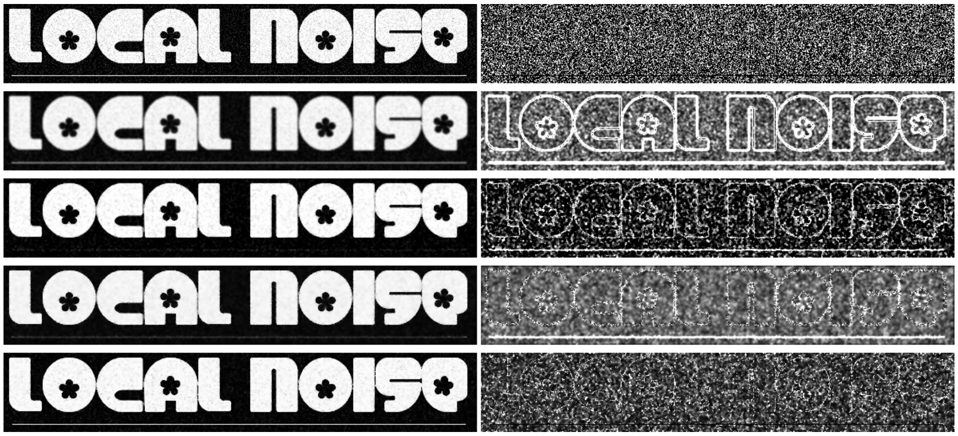


Figure 4.3: Noise reduction results on a toy image of letters and an underline corrupted by Gaussian white noise. Left, top to bottom: Image after no noise reduction, Gaussian filtering, Median filtering, ADF, and SUSAN filtering, respectively. Right: Absolute difference between noise reduced images and the original image for the different techniques (multiplied by 10, darker is better). The top difference image illustrates the white noise model. The Median filter reduced the total amount of noise by the largest amount. However, the SUSAN filter clearly has more desirable behavior near edges, where the other methods have introduced much more visible artifacts.

$$G(x,y) = \frac{\sum\limits_{i}\sum\limits_{j} I(x+i,y+j)e^{-\frac{(x+i)^2+(y+j)^2}{2\sigma^2}}}{\sum\limits_{i}\sum\limits_{j} e^{-\frac{(x+i)^2+(y+j)^2}{2\sigma^2}}} \tag{4.1}$$

Linear filtering methods such as mean filtering and Gaussian filtering unquestionably reduce the effects of local noise through the use of neighborhood averaging. However, high-pass information

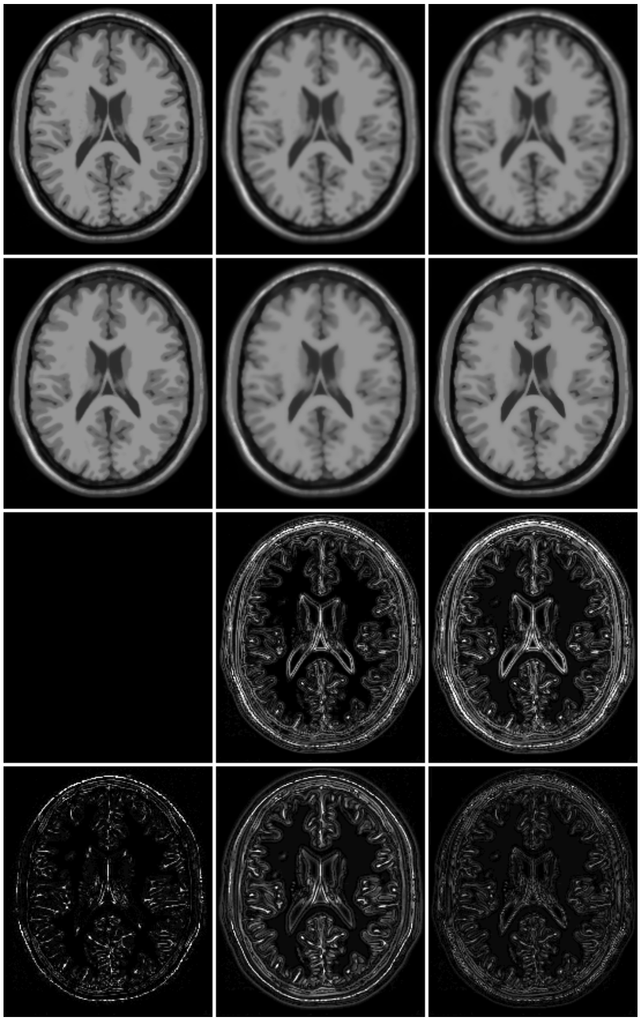


Figure 4.4: Noise reduction results on noise-free simulated MRI [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998]. Top, left to right: Image after no filtering, Mean filtering, Gaussian filtering. Second row, left to right: Image after Median filtering, ADF, and SUSAN filtering. Bottom 2 rows: Absolute differences between noise reduced images and the original noise-free image (multiplied by 10, darker is better). As with the toy example, the Median filtered and SUSAN filtered images have introduced the least amount of additional noise.

is lost, due to averaging across edges. *Median filtering* is an alternative to linear methods. A Median filter replaces each pixel's intensity value with the median intensity value in its neighborhood. In addition to incorporating only intensities that were observed in the original image, median filtering does not blur relatively straight edges. Median filtering is resistant to impulse noise (large changes in the intensity due to local noise), since outlier pixels will not skew the median value. Median filtering and other 'order-statistic' based filters are more appealing than simple linear filters, but have some undesirable properties. Median filtering is not effective at preserving the curved edges [Smith and Brady, 1997] often seen in biological imaging. Median filtering can also degrade fine image features, and can have undesirable effects in neighborhoods where more than two structures are represented. Due to the disadvantages of Median filtering, it is generally applied in low signal to noise ratio situations.

Anisotropic Diffusion Filtering (ADF) is a popular preprocessing step for MR image segmentation, and has been included previously in tumor segmentation systems, including the works of

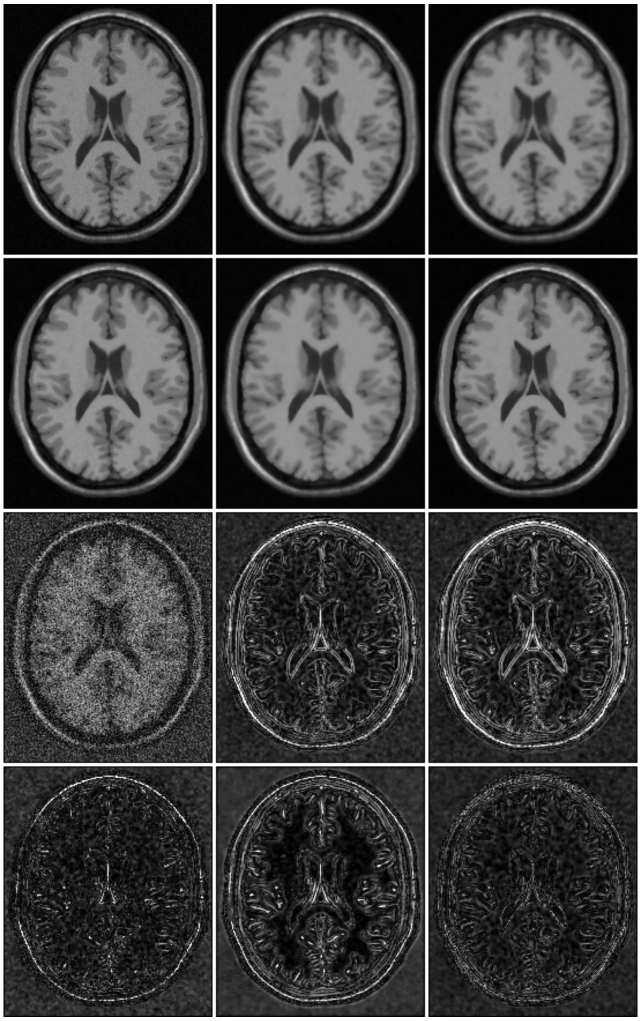


Figure 4.5: Noise reduction results on simulated MRI with simulated noise [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998]. Top, left to right: Image after no noise reduction, Mean filtering, Gaussian filtering. Second row, left to right: Image after Median filtering, ADF, and SUSAN filtering. Bottom 2 rows: Absolute differences between noise reduced images and the original noise-free image (multiplied by 10, darker is better). The difference image after no filtering illustrates the noise model from the simulator (not white). In this scenario, the Median and SUSAN filtered images diverge the least from the noise-free version of the image, while the SUSAN filter in addition has the most desirable behavior near edges.

[Vinitski et al., 1997, Kaus et al., 2001]. This technique was introduced in [Perona and Malik, 1990], and extended to MR images in [Gerig et al., 1992]. As with mean and Gaussian filtering, ADF reduces noise through smoothing of the image intensities. Unlike mean and Gaussian filtering, however, ADF uses image gradients to reduce the smoothing effect from occurring across edges. ADF thus has the goal of smoothing within regions, but not between regions (*edge-preserving smoothing*). Furthermore, in addition to preserving edges, ADF enhances edges since pixels on each side of the edge will be assigned values representative of their structure. This is desirable in MR image segmentation since it reduces the effects of partial volume averaging.

One disadvantage of ADF is that, unlike Median filtering, it is sensitive to impulse noise, and thus can have undesirable effects if the noise level is high. The *Anisotropic Median-Diffusion Filtering* method was developed to address this weakness [Ling and Bovik, 2002], but this method introduces the degradation of fine details associated with Median filtering and so is not used here.

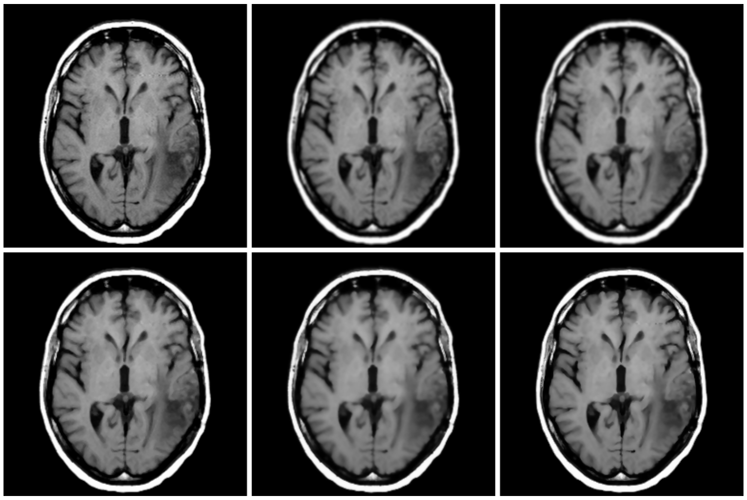


Figure 4.6: Noise reduction results on real T1-weighted MR images. Top, left to right: Original image, Mean filtered image, Gaussian filtered image. Bottom, left to right: Median filtered image, ADF filtered image, SUSAN filtered image.

Another disadvantage of ADF is that regions near thin lines and corners are not appropriately handled, due to their high image gradients [Smith and Brady, 1997].

Another alternative to ADF for edge-preserving smoothing is the *Smallest Univalue Segment Assimilating Nucleus (SUSAN) filter* [Smith and Brady, 1997]. The main idea behind this simple method (with an intimidating name and equation) is that the contribution of neighboring pixels is weighed by a Gaussian in the spatial *and* the intensity domain (see Equation 4.2, τ is an additional parameter of the non-linear filter). The use of a Gaussian in the intensity domain allows the algorithm to smooth near thin lines and corners. For example, rather than ignoring the region around a thin line (due to the presence of a high image gradient), the SUSAN filter weighs pixels on the line more heavily when evaluating other pixels on the line, and weighs pixels off the line according to pixels that are similar in (spatial location and) intensity to them. In addition to addressing this weakness of ADF, the SUSAN filter employs a heuristic to account for impulse noise. If the dissimilarity with neighboring pixels in the intensity and spatial domain is sufficiently high, a median filter is applied instead of the SUSAN filter. Thus, Equation 4.2 is replaced by a median filter if a pixel appears to resemble impulse noise.

$$S(x,y) = \frac{\displaystyle\sum_{i \neq x}\sum_{j \neq y} I(x+i, y+j) e^{-\frac{(x+i)^2+(y+j)^2}{2\sigma^2} - \frac{(I(x+i)-I(y(+j))^2}{t^2}}}{\displaystyle\sum_{i \neq x}\sum_{j \neq y} e^{-\frac{(x+i)^2+(y+j)^2}{2\sigma^2} - \frac{I(x+i)I(y+j))^2}{t^2}}} \tag{4.2}$$

The effects of several of the noise reduction methods discussed in this section on ‘toy’ data are shown in Figures 4.2 and 4.3, demonstrating the effects of these filters with and without added white noise, respectively. The effects of these techniques on a simulated MR image with a known noise component are demonstrated in Figures 4.4 and 4.5. Finally, results on real data (where the noise component of the image is not known) are presented in Figure 4.6. We chose to use the SUSAN filtering method, since it is less sensitive to the selection of the parameters than ADF, and has slightly better noise reduction properties than Median Filtering and ADF.

4.1.2 Inter-Slice Intensity Variation Reduction

The second step in the Noise Reduction phase is the reduction of inter-slice intensity variations. Due to gradient eddy currents and ‘crosstalk’ between slices in ‘multislice’ acquisition sequences, the two-dimensional slices acquired under some acquisition protocols may have a constant slice-by-slice intensity off-set [Leemput et al., 1999b]. It is noteworthy that these variations have different properties than the intensity inhomogeneity observed within slices, or typically observed across slices. As opposed to being slowly varying, these variations are characterized by sudden intensity changes in adjacent slices. A common result of inter-slice intensity variations is an interleaving between 'bright' slices and 'dark' slices [Simmons et al., 1994], (the 'even-odd' effect). Gradually varying intensity changes between slices will be corrected for in the intensity inhomogeneity reduction step, but most methods for intensity inhomogeneity reduction do not consider these sudden changes. This step, therefore, attempts to reduce sudden intensity variations between adjacent slices.

In comparison to the estimation of slowly varying intensity inhomogeneities, correcting inter-slice intensity variations has received little attention in the medical imaging literature. One early attempt to correct this problem in order to improve segmentation was presented in [Choi et al., 1991]. This work presented a system for the segmentation of normal brains using Markov Random Fields, and presented two simple methods to reestimate tissue parameters between slices (after patient-specific training on a single slice). One method thresholded pixels with high probabilities of containing a single tissue type, while the other used a least squares estimate of the change in tissue parameters. A similar approach was used in one of the only systems thus far to incorporate this step for tumor segmentation [Ozkan et al., 1993]. This system first used patient-specific training of a neural network classifier on a single slice. When segmenting an adjacent slice, this neural network was first used to classify all pixels in the adjacent slice. The locations of pixels that received the same label in both slices were then determined, and these pixels in the adjacent slice were used as a new training set for the neural network classifier used to classify the adjacent slice. Each of these approaches require not only a tissue model, but patient-specific training, making them unsuitable for use in our framework at this early stage.

One of the most impressive inter-slice intensity correction methods to date was presented in [Leemput et al., 1999b]. This work presented two methods to incorporate inter-slice variation correction within an EM segmentation framework. The first simply incorporated slice-by-slice constant intensity offsets into the inhomogeneity estimation, while the second method computed a two-dimensional inhomogeneity field in each slice and with these produced a three-dimensional inhomogeneity field that allowed inter-slice intensity variations. [Zijdenbos et al., 1995] presents the method used by the INSECT system for this step to improve the segmentation of Multiple Sclerosis lesions. This method estimates a linear intensity mapping based on pixels at the same location in adjacent slices that are the same tissue type. Unfortunately, despite the lack of patient-specific training, these methods each still require a tissue model (in each slice) that may be violated in data containing significant pathology.

[Vokurka et al., 1999] presented a method that is not dependent on a tissue model. This method used a median filter to reduce noise, and pruned pixels from the intensity estimation by band thresholding the histogram, and removing pixels representing edges. The histogram was divided into bins and the authors fit a parabola to the heights of the 3 central bins (the author's desired an estimation that was symmetric around the central bin), used to determine the intensity mapping. Although model-free, this method makes major assumptions about the distribution of the histogram, that may not be true in all modalities or in images with pathological data. In addition, this method ignores spatial information.

Inter-slice intensity variation correction can be addressed using the same techniques employed in Intensity Standardization, which will be discussed in Section 4.4. However, most methods for Intensity Standardization employ a tissue model or a histogram matching method that will be sensitive to outliers. It was ultimately chosen not to use one of the existing histogram matching methods, since real data may have anisotropic pixels, where the tissue distributions can change significantly between slices. The methods in [Zijdenbos et al., 1995, Ozkan et al., 1993] are more appealing since these methods use spatial information to determine appropriate pixels for use in estimation. However, these methods rely on a tissue model that would violate the framework's guidelines for a noise reduction technique. Although the method of [Vokurka et al., 1999] is a histogram matching method, removing points from the estimation in a model-free way is appealing. We present in this section a simple method to identify good candidates for estimating the intensity between slices as in [Zijdenbos et al., 1995, Ozkan et al., 1993], but in a model-free way.

We will assume that the intensity mapping between adjacent slices can be described by a multiplicative scalar value w, a model commonly used [Zijdenbos et al., 1995, Leemput et al., 1999b]. Although this simple linear transformation is insufficient to completely correct the effect, it allows the admission of straightforward methods that can greatly reduce this effect. If we assume that the slices are exactly aligned such that each pixel in slice X corresponds to a pixel in slice Y of the same tissue type, then the scalar w could be estimated by solving the equation below (where X and Y are vectors of intensities and $X(i)$ has the same spatial location as $Y(i)$ within the image):

$$Xw = Y \tag{4.3}$$

However, since there will not be an exact mapping between tissue types at locations in adjacent slices, an exact value for w that solves this equation will not exist, and therefore the task becomes to estimate an appropriate value for w. One computationally efficient way to estimate a good value of w would be to calculate the value for w that minimizes the squared error between Xw and Y:

$$\min_{w} \sum_{i} (X(i)w - Y(i))^2 \tag{4.4}$$

The optimal value for w in this case can be determined by solving for w in the ‘normal equations’ [Shawe-Taylor and Cristianini, 2004] (employing the matrix pseudoinverse):

$$w = (X'X)^{-1}X'Y \tag{4.5}$$

Unfortunately, this computation is sensitive to areas where different tissue types are not aligned, since these regions are given weight equal to that of pixels where tissue types are appropriately aligned in the adjacent slices. The value w thus simply minimizes the error between the intensities at corresponding locations in adjacent slices, irrespective of whether the intensities should be the same (possibly introducing additional inter-slice intensity variations). The objective must thus be modified to restrict the estimation of w to locations that actually *should* have the same intensity after the intensity transformation w is applied. This is difficult without the use of a tissue model or a segmentation of the image. However, an alternate approach to identifying tissues or performing a segmentation is to weight the errors based on the importance of having a small error between each corresponding location $(X(i), Y(i))$. Given a weighting of the importance for each pixel to have the same intensity between adjacent slices $R(i)$, the calculation of w would focus on computing a value that minimizes the squared error for areas that are likely to be aligned, while reducing the effect of areas where tissues are likely misaligned. Given $R(i)$ for each i, the least squares solution can be modified to use this weight by performing element-wise multiplication of both the vectors X and Y with R [Moler, 2002]. This scaling of both vectors modifies the error function to be proportional to the values in R (using $.*$ to denote element-wise multiplication):

$$\min_{w} \sum_{i} ((X(i). * R(i))w - Y(i). * R(i))^2 \tag{4.6}$$

The value w that minimizes the above relevance-weighted loss function can be computed as before:

$$w = ((X. * R)'(X. * R))^{-1}(X. * R)'(Y. * R) \tag{4.7}$$

If the image was segmented into anatomically meaningful regions, computing $R(i)$ would be trivial, it would be 1 at locations where the same tissue type is present in both slices and 0 when the tissue types differ. Without a segmentation, this can be approximated. An intuitive approximation would be to weight pixels based on a measure of similarity between their regional intensity

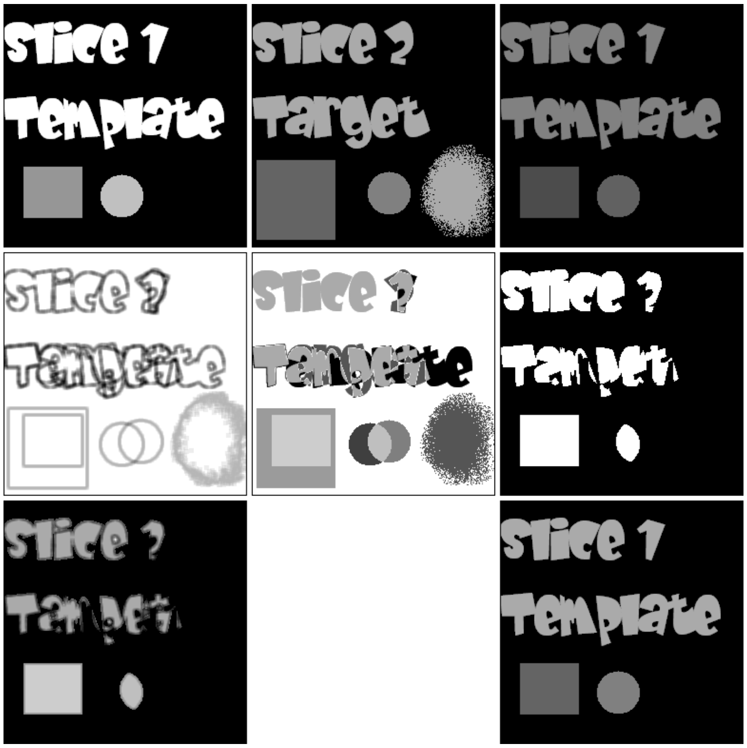


Figure 4.7: Toy example of weighting for inter-slice intensity variation reduction. The goal is to transform slice 1 so that its intensities match those of slice 2. Top row, left to right: Original slice 1, original slice 2, slice 1 scaled from unweighted linear regression. Since the objects in the two slices differ, the unweighted linear regression did not estimate a good transformation. Middle row: elements of the weighting. Left to right: Regional joint entropy, absolute difference, joint foreground pixels. Bottom left: Combined weighting (darker regions receive less weight). Note that the weighting focuses the estimation on regions that should have the same intensity. Bottom right: Linear regression using the combined weighting estimates a near-optimal linear scaling.

distributions. A method robust to intensity-scaling to perform this approximation is to compute the (regional) joint entropy of the intensity values. The (Shannon) joint entropy is defined as follows [Pluim et al., 2003]:

$$H(A_1, A_2) = - \sum_{i \epsilon A_1, j \epsilon A_2} p(i,j) \log p(i,j) \tag{4.8}$$

The value $p(i,j)$ represents the likelihood that intensity i in one slice will be at the same location as intensity j in the adjacent slice, based on an image region. We use a 25 pixel square window to compute the values $p(i,j)$ for a region, and divide the intensities into 10 equally spaced bins to make this computation. The frequencies of the 25 intensity combinations in the resulting 100 bins are used for the $p(i,j)$ values (smoothing these estimates could give a less biased estimate). The joint entropy computed over these values of $p(i,j)$ has several appealing properties. In addition to being insensitive to scaling of the intensities, it is lowest when the pixel region is homogeneous in both slices, will be higher if the intensities are not homogeneous in both slices but are spatially correlated, and will be highest when the intensities are not spatially correlated. After a sign reversal and normalization to the range $[0, 1]$, the regional joint entropy of the image regions could be used as values for $R(i)$, that would encourage regions that are more homogeneous and correlated between the slices to receive more weight in the estimation of w than heterogeneous and uncorrelated regions.

Joint entropy provides a convenient measure for the degree of spatial correlation of intensities, that is not dependent on the values of the intensities as in many correlation measures. However,

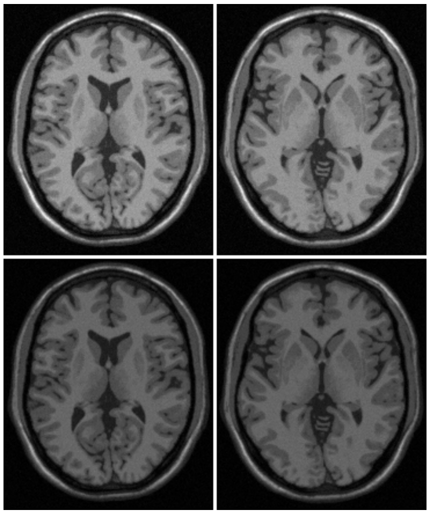


Figure 4.8: Inter-slice intensity variation reduction with simulated MR images [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998] and an applied linear intensity offset. Top left: Simulated slice 1. Top right: Simulated slice 2 ($10mm$ away from slice 1). Bottom left: Slice 1 with an applied linear offset. Bottom right: Slice 2 transformed using the weighted linear regression between slice 2 and the darkened slice 1. The method successfully recovered a scale factor very close to the one applied.

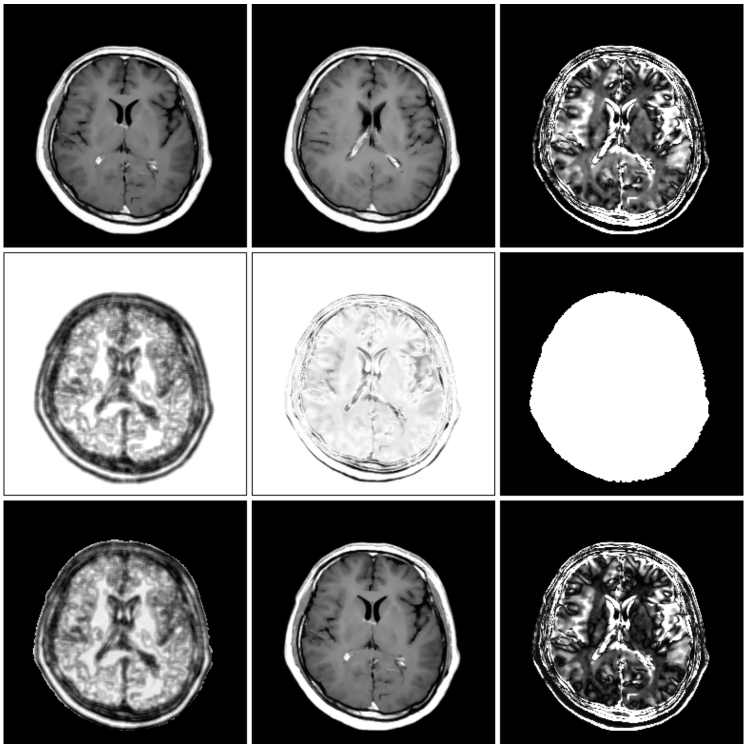


Figure 4.9: Inter-slice intensity variation reduction for real data. Top, left to right: Slice with an unknown intensity offset, adjacent slice, difference between the adjacent slices (multiplied by 10). Middle row, left to right: Entropy weighting, difference weighting, joint foreground pixels. Bottom, left to right: Combined weighting, slice 1 after transformation, difference between slice 1 after transformation and slice 2. The effect has not been completely removed, but has been noticeably reduced.

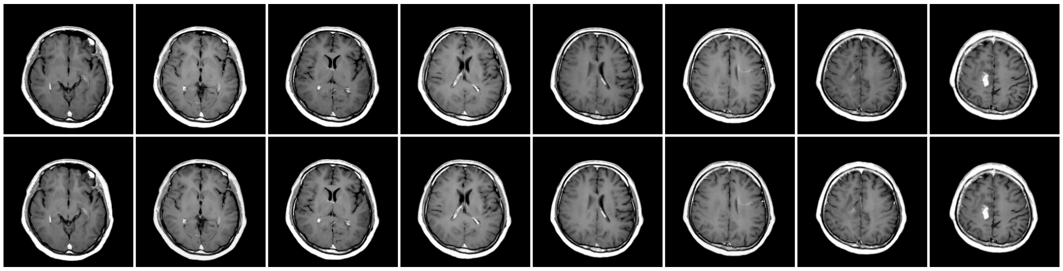


Figure 4.10: Inter-slice intensity variation reduction for an image series. Top, original image series (the even slices are noticeably brighter than the odd slices). Bottom, the same series after reduction of inter-slice intensity variations. The variations have not been completely corrected, but their effects have been reduced.

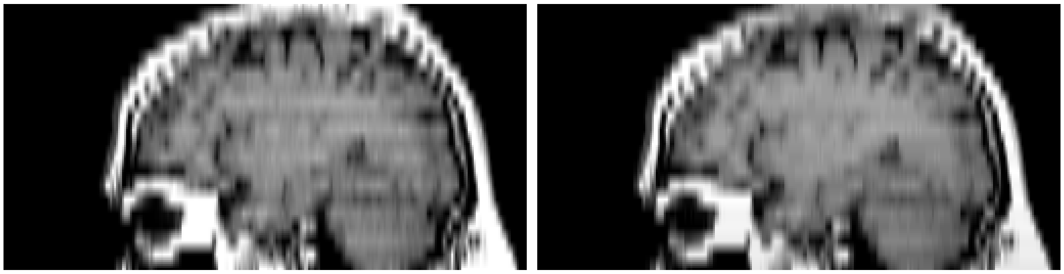


Figure 4.11: Inter-slice intensity variation reduction from a different angle. Left: Sagittal view of the original slices from the series in Figure 4.10. Right: The same view after inter-slice intensity variation reduction. The brain area in the input image clearly shows the ‘even-odd’ offset effect, that has been noticeably reduced in the output image.

the values of the intensities in the same regions in adjacent slices should also be considered, since pixels of very different intensity values should receive decreased weight in the estimation, even if they are both located in relatively homogeneous regions. Thus, in addition to assessing the spatial correlation of regional intensities, higher weight should be assigned to areas that have similar intensity values before transformation, and the weight should be dampened in areas where intensity values are different. The most obvious measure of the intensity similarity between two pixels is the absolute value of their intensity difference. This measure is computed for each set of corresponding pixels between the slices, and normalized to be in the range $[0, 1]$ (after a sign reversal). Values for $R(i)$ that reflect both spatial correlation and intensity difference can be computed by multiplying these two measures. As a further refinement to this measure, the threshold selection algorithm from [Otsu, 1979] (and morphological filling of holes) is used to distinguish foreground (air) pixels from background (head) pixels, and $R(i)$ is set to zero for pixels representing background areas in either slice (since they are not relevant to this calculation). Thus, each value in $R(i)$ is computed as follows (where N_1 and N_2 are normalizing constants, $H(X(i), Y(i))$ is the regional joint entropy centered at i, and P_{fore} is an indicator function that returns 1 if the pixel is part of the foreground and 0 otherwise):

$$r(i) = \frac{(N_1 - H(X(i), Y(i)))}{N_1} \frac{(N_2 - |X(i) - Y(i)|)}{N_2} P_{fore}(X(i)) P_{fore}(Y(i)) \tag{4.9}$$

Figure 4.7 demonstrates the advantage of weighting the estimation on a toy example. Figure 4.8 shows an example of the estimation being applied to simulated MR images to correct for an applied linear offset, while Figure 4.9 presents results on real data.

In our implementation, the weighted least squares estimation computes the linear mapping to the median slice in the sequence from each of the two adjacent slices. The implemented algorithm then proceeds to transform these slices, and then estimates the intensity mappings of their adjacent slices, continuing until all slices have been transformed. Figure 4.10 shows the results of this process, while Figure 4.11 shows the same results viewed from an orthogonal angle. In our data set, this effect was only present in the T1-weighted images, and thus in our experiments this step was not performed for T2-weighted images.

Future implementations could expand on this method by computing a non-linear intensity mapping between the slices. Our experiments with non-linear mappings showed that they were difficult to work with, since non-linear transformations tended to reduce image contrast. This process would thus need to be subject more advanced regularization measures.

4.1.3 Intensity Inhomogeneity Reduction

The third step in the Noise Reduction phase is the reduction of intensity inhomogeneity across the volume. The task in this step is to estimate a three-dimensional inhomogeneity field, of the same size as the volume, that represents the intensity inhomogeneity. This field is often assumed to vary slowly spatially, and to have a multiplicative effect. The estimated field can be used to generate an image where the variances in the intensities for each tissue type are reduced, and differences between tissue types are preserved.

There are an immense number of techniques for post-acquisition intensity inhomogeneity reduction. Discussed in [Gispert et al., 2004] are the causes of the intensity inhomogeneity, and many of the techniques proposed for automatic post-acquisition correction are surveyed in [Gispert et al., 2004, Shattuck et al., 2001], including methods based on linear and non-linear filtering, seed point selection, clustering, mixture models through expectation maximization with and without spatial priors, entropy minimization, hidden Markov models, and many other approaches. The diversity of methods proposed are the result of the difficulty in validating methods on real data sets and the lack of studies that quantitatively compares different methods.

Rather than introducing yet another technique, the utilization of an existing inhomogeneity correction algorithm is appealing, but it is not obvious which correction algorithm should be used. This is primarily due to the fact that new methods are often evaluated by comparing results before and after correction. Although it is clear that corrected volumes can improve segmentation, validating new methods without quantitative comparisons to existing methods makes selecting among the many existing correction algorithms difficult. [Arnold et al., 2001] performed a multi-center evaluation on real and simulated data of six correction algorithms, selected as representative methods for different correction strategies. The methods examined were based on homomorphic filtering (HUM), Fourier domain filtering (EQ), thresholding and Gaussian filtering (CMA), a tissue mixture model approach using spatial priors (SPM99), the Nonparametric Nonuniform intensity Normalization algorithm (N3), and a method comparing local and global values of a tissue model (BFC). Although there was no clear superior method, the BFC and the N3 methods generally provided a better estimation than the other methods. A more recent study compared the performance of four algorithms on the simulated data used in [Arnold et al., 2001], in addition to real data at different field strengths [Gispert et al., 2003]. The four methods examined were the N3 method, the SPM99 method, the SPM2 method (expectation maximization of a mixture model with spatial priors), and the author's NIC method (expectation maximization that minimizes tissue class overlap by modeling the histogram by a set of basis functions). The simulated data indicated that the NIC method was competitive with the N3 and BFC methods. The results on real data indicated that the N3 method outperformed the SPM99 and SPM2 methods, and that the NIC method outperformed the N3 method.

[Velthuizen et al., 1998] examined the effects of intensity inhomogeneity correction on brain tumor segmentation. The segmentation method used was a kNN classifier using the image intensities as features, employing patient-specific training. Although no performance gain was achieved, the methods evaluated were simple filtering methods, which have not performed well in the few comparative studies on correction methods. A study that compared different inhomogeneity correction algorithms in the presence of Multiple Sclerosis lesions was done in [Sled, 1997]. This work com-

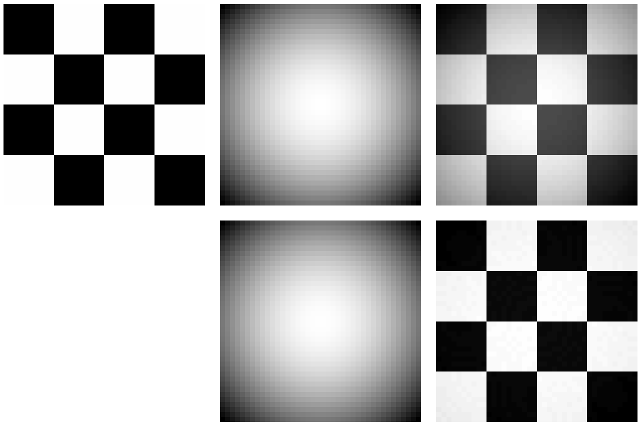


Figure 4.12: The N3 algorithm applied to an ideal image. Top, left to right: Original image, applied inhomogeneity field, image corrupted by the inhomogeneity field. The N3 algorithm is ideal to correct the inhomogeneity in this image, since the inhomogeneity field's intensities follow a Gaussian distribution, and the underlying image is composed of two high frequency components (black tiles and white tiles) that were 'flattenned' in the histogram. Bottom middle: Inhomogeneity field estimated by the N3 algorithm. Bottom right: Corrupted image corrected by the inhomogeneity field estimated by the N3 algorithm. The fields are not identical, and there remains inhomogeneity in the corrected image (visible near the image edges). However, the inhomogeneity has been clearly reduced.

pared the N3 algorithm to an algorithm based on (automatic) seed point selection from segmented white matter regions (WM), and an expectation maximization fitting of a tissue mixture model with and without spatial priors (EM and REM, respectively). Although the REM method performed the best overall on simulated data, both of the EM based algorithms performed poorly on real data. Among these four methods, only the N3 algorithm does not rely on either a tissue model or a segmentation. Based on the quantitative evaluations performed on real data sets, the algorithms with the best performance on real data seem to be the NIC, BFC, and N3 algorithms. Since the NIC method is not automatic, and both the BFC and NIC methods assume a tissue model (that will be violated if large abnormalities are present), the most logical choice for a bias correction strategy would be the N3 method.

The Nonparametric Nonuniform intensity Normalization (N3) algorithm was designed for accurate intensity inhomogeneity correction in MR images, without the need for a parametric tissue model [Sled, 1997, Sled et al., 1999]. One of the main goals of the authors of these works was to remove the dependency of the intensity correction on *a priori* anatomic information, such that inhomogeneity correction could be performed as a preprocessing step in quantitative analysis. As with most inhomogeneity correction methods, this work models the intensity inhomogeneity effect as a slowly varying multiplicative field. The main assumption underlying this method is that an image will consist of several components that occur with a high frequency. In typical brain images, these components might be nuclear gray matter, cortical gray matter, white matter, CSF, scalp, bone, and other tissue types. Under this assumption, the histogram of a noise-free and inhomogeneity-free image should form a set of peaks at the intensities of these tissues (with a small number of partial volume pixels in between the peaks). If an image is corrupted by a slowly varying inhomogeneity field, however, these peaks in the histogram will be 'flattened', since the values that should be at the peaks will vary slowly across the image. The N3 method seeks to estimate an underlying uncorrupted image where the frequency of the histogram is maximized (by 'sharpening' the probability density function of the observed image), while enforcing that the field must be slowly varying (preventing degenerate solutions).

In the N3 method, the probability density function of the values in the inhomogeneity field are assumed to follow a zero-mean Gaussian distribution, which allows tractable computation of

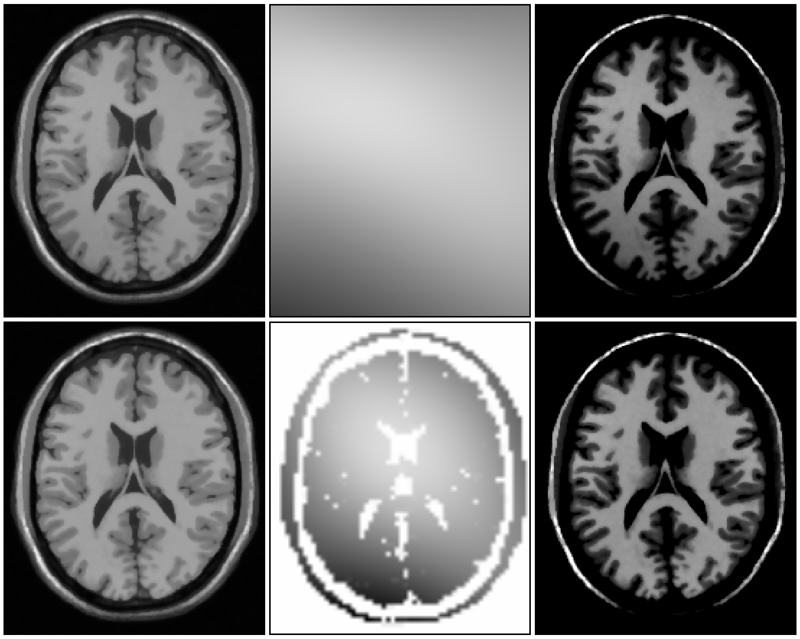


Figure 4.13: The N3 algorithm applied to a simulated MR image [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998] with a known inhomogeneity field. Top, left to right: T1-weighted inhomogeneity corrupted image, applied inhomogeneity field, the image with intensities re-scaled to focus on white matter regions. The inhomogeneity is most clearly visible in the lower left part of the re-scaled image (the white matter here is considerably darker). Bottom, left to right: Image corrected with the N3 algorithm, the inhomogeneity field estimated by the N3 algorithm, the re-scaled corrected image. The holes in the field are due to the thresholding performed by the algorithm to determine foreground pixels. The shape and orientation of the estimated field differ slightly from the true field. Residual inhomogeneity remains visible in the lower left area of the image, but it has been clearly reduced.

the underlying uncorrupted image. Under this assumption, the probability density for the (log) intensities of the underlying data can be estimated by deconvolution of the probability density for the (log) intensities of the observed image with a Gaussian distribution. The procedure iterates by repeated deconvolution of the current estimate of the true underlying data. After each iteration, an inhomogeneity field can be computed based on the current estimate of the underlying data. This field is affected by noise, and is smoothed by modeling it as a linear combination of (β-spline) basis functions. This smoothed field is used to update the estimated underlying probability density, which reduces the effects of noise on the estimation of the underlying probability. The algorithm converges empirically to a stable solution in a small number of iterations (the authors say that it is typically ten). Figure 4.12 presents an example of the results of this algorithm applied to an image where the assumptions made are true, while Figure 4.13 shows the results of applying this technique to simulated data with a known inhomogeneity field where the assumptions are only approximately true.

Consider the case of MR brain images with pathology. Many approaches (such as the Expectation Maximization approaches) rely on the segmentation of the image with a tissue model consisting of a fixed number of class (that are often assumed to have a Gaussian distribution). Inhomogeneity correction in the presence of pathology for these methods is challenging since the model's assumptions are violated in creating the segmentation that will be used to estimate the field. Erratic results have been reported for EM-based approaches in the presence of Multiple Sclerosis Lesions [Sled, 1997], which interfere with the histogram to a lesser extent than do large tumors. Furthermore, since the performance of EM algorithms depends heavily on having a good initialization,

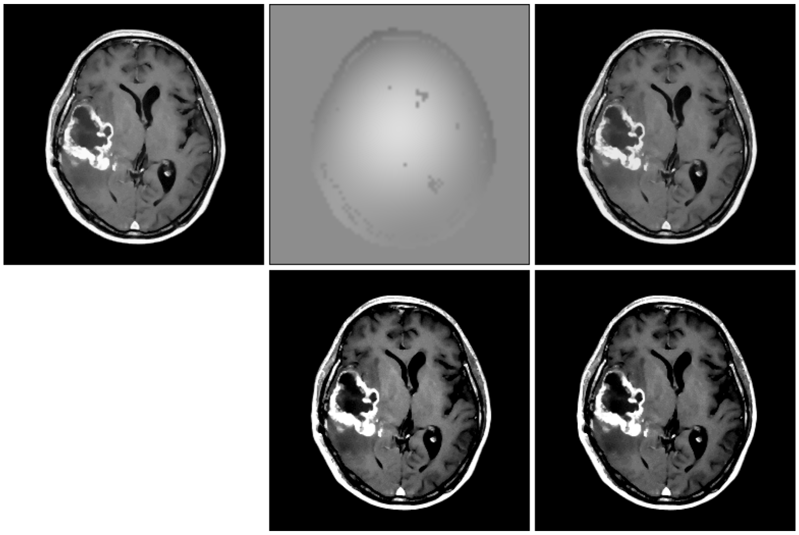


Figure 4.14: The N3 algorithm applied to real MR images with large abnormal regions. Top, left to right: Original image, estimated field, corrected image. Bottom middle: Original image rescaled to highlight white matter regions. Bottom right: Corrected image rescaled to highlight white matter regions. It can be seen that the inhomogeneity within white matter regions has been reduced, but that tissues near the center of the image remain brighter than those near the periphery. It is noteworthy that the large enhancing tumor, necrotic, and edema areas visible on the left hand side of the image have not interfered significantly with the inhomogeneity field estimation.

sensitivity even to normal anatomic variations could cause the algorithm to reach an unsatisfactory local optima [Sled, 1997]. Now consider the N3 approach, that does not rely on a segmentation or tissue model. Although this method does make assumptions about the intensity distribution, these assumptions apply to pathology in addition to normal data. This method has been shown to give effective inhomogeneity correction in the presence of Multiple Sclerosis lesions [Sled, 1997] and large tumors [Likar et al., 2001]. Intuitively, resistance to a small number of outliers interfering with the estimation is done through the same method that confers resistance to noise, the smoothing of the field estimations between iterations. A larger number of outliers will also not interfere in the estimation, since they will comprise an additional high frequency component of the histogram. However, one case where this method could fail is if the abnormal area varies in intensity across the image slowly enough that it mimics an inhomogeneity field effect. Although there is inherent ambiguity in determining tumor boundaries since edges may not be well defined, this does not indicate that the intensity varies slowly over large elements of the structure, and the transition from normal areas to abnormal areas is fast enough that it is has not been confused with inhomogeneity effects in our experiments. Figure 4.14 shows the results of applying the N3 algorithm to real data.

A set of related approaches to the N3 method are methods based on entropy minimization. First proposed in [Viola, 1995], this idea has since been explored and extended in several works including [Mangin, 2000, Likar et al., 2001, Ashburner, 2002, Vovk et al., 2004]. Both the N3 method and the entropy minimization methods assume that the histogram should be composed of high frequency components that have been ‘flattened’ by the presence of an inhomogeneity field, and estimate a field that will result in a well clustered histogram. The main difference is that the N3 method assumes an approximately Gaussian distributed inhomogeneity field, while the entropy minimization methods directly estimate a field that will minimize the (Shannon) entropy. In [Likar et al., 2001], comparative experiments were made between the N3 algorithm, a method that estimates image gradients and normalizes homogeneous regions (FMI), and three entropy minimization methods. The entropy based approaches and the N3 approach all outperformed the FMI method, while the entropy based approaches and the N3 approach performed similarly for images of the brain, and one of the entropy based methods (M4) tended to outperform the other methods on average on images of other areas of the body. [Vovk et al., 2004] compared the M4 method to a new entropy minimization method that estimated entropy based on both intensity information and the results of image convolution with a Laplacian filter (a method to approximate the local image second derivative). The new method outperformed the M4 method on simulated data and a limited set of real data, obtaining nearly optimal performance based on the author's measure.

Another recent approach that outperformed the N3 method in a small scale study was presented in [Studholme et al., 2004]. This method used an aligned template to estimate expected local image statistics in estimating the bias field. This type of method is obviously not appropriate for the task of brain tumor segmentation, since a large tumor will not be present in the template. Although we presented several simple methods to account for outliers in the previous section and will be discussing more sophisticated template-based methods in Section 4.3, these methods are for estimating global effects, rather than the local effects of intensity inhomogeneity. We would prefer to use a method that can use abnormal areas to enhance the bias field estimation, rather than extrapolating over abnormal areas or running the risk of the abnormal area being recognized as an inhomogeneity field effect.

Although the entropy minimization approaches have been introduced as a potential alternative to the N3 method, we chose to use the N3 method. The N3 methods has been involved in a larger number of comparative studies and has been tested for a much larger variety of different acquisition protocols and scanners, consistently ranking as one of the best algorithms. However, the entropy minimization approaches have shown significant potential in the limited number of comparative studies that they have been evaluated in, and these approaches typically require a smaller number of parameters than the N3 method, are slightly more intuitive, and have (arguably) more elegant formulations. Thus, a potential future improvement for this step could be the use of an entropy minimization approach (with the method of [Vovk et al., 2004] being one of the most promising).

4.1.4 Summary

The noise reduction step consists of four stages. The first stage is the reduction of local noise. This is done by using the SUSAN filter, which is a non-linear filter that removes noise by smoothing image regions based on both the spatial and intensity domains. This filter has the additional benefit that it enhances edge information and reduces the effects of partial volume averaging. The second stage is the correction for inter-slice intensity variations. We use a simple least squares method to estimate a linear multiplicative factor based on corresponding locations in adjacent slices. This step uses a simple regularization measure to ensure that outliers do not interfere in the estimation, and to bias the estimation towards a unit multiplicative factor. The third stage is the correction for smooth intensity variations across the volume. This stage uses the N3 method, which finds a three-dimensional correcting multiplicative field that maximizes the frequency content of the histogram, under the assumption that the field varies slowly spatially. This technique is not affected by large pathologies, and does not rely on a tissue model that is sensitive to outliers. Finally, we perform an additional three-dimensional local noise reduction as the fourth stage. This step removes regions that can be identified as noise after the two inhomogeneity correction steps. The SUSAN filter is again used for this step, for the same reasons it was used in the two-dimensional case. A three-dimensional SUSAN filter should be used for this stage, but a two-dimensional SUSAN filter was used for our experiments since the pixel sizes were anisotropic.

The implementation of the SUSAN filter present in the MIPAV package was used [MIPAV, Online]. Default values of 1 for the standard deviation, and 25 for the brightness threshold (the images were converted to an 8-bit representation) were used. The implementation of the N3 algorithm in the MIPAV package was also used. The work in [Sled et al., 1999] was followed in setting the parameters, using a Full Width at Half Maximum of 0.15, a field distance of 200mm, a sub-sampling factor of 4, a termination condition of a change of 0.001, and setting the maximum number of iterations at 50. Adaptive histogram thresholding to distinguish foreground from background pixels was not used as suggested by the author, as it gave erratic results. The inter-slice intensity correction method was implemented using Matlab [MATLAB, Online], taking advantage of the pseudoinverse to compute the matrix inversion (with the smallest norm) in the least squares solution.

4.2 Spatial Normalization

Spatial normalization also consists of four steps: inter-modality coregistration, linear template registration, non-linear template registration, and interpolation. Medical image registration is a topic with an extensive associated literature. A survey of medical image registration techniques is provided in [Maintz and Viergever, 1998]. Although slightly dated due to the effects of several influential recent techniques, this extensive work categorizes over 200 different registration techniques based on 9 criteria. Although these criteria can be used to narrow the search for a registration solution, there remains decisions to be made among a variety of different methods, many of which are close variants with small performance distinctions. The registration methods selected for this phase follow the guidelines presented in the previous chapter; they are fully automatic and do not rely on segmentations, landmarks, or extrinsic markers present in the image. Furthermore, they each utilize three dimensional volumes, and use optimization methods that are computationally efficient.

4.2.1 Coregistration

The first step in the Spatial Normalization phase is the spatial alignment of the different modalities. In our implementation, we define one of the modalities as the target, and compute a transformation for each other modality mapping to this target. This transformation is assumed to be rigid-body, meaning that only global translations and rotations are considered (since pixel sizes are known). Typically, coregistration is used as a method to align MR images with CT images, or MR images with PET images. Techniques that can perform these tasks are also well suited for the (generally) easier task of aligning MR images with other MR images of a different (or the same) modality. The major challenge associated with the coregistration task has traditionally been to define a quantitative measure that assesses spatial alignment. Given this measure, the task is reduced to a search for a set of transformation parameters that optimize this measure. Since direct comparisons of intensities are not meaningful when aligning images of different modalities, various measures have been proposed for coregistration that do not rely on minimizing the difference between images. Examples of these measures can be found in the influential work of [West et al., 1997], which evaluated 16 algorithms for the registration of MR images with CT images, and MR images with PET images. Currently, one of the most popular methods of coregistration in medical imaging is the maximization of the relative entropy measure known as *Mutual Information* (and its variants)

[Pluim et al., 2003] provides an excellent overview of the concepts of entropy and Mutual Information, a brief overview of their history, and provides an extensive survey of medical image registration techniques based on Mutual Information and its variants. Mutual Information requires the computation of the entropy of individual images. This measure utilizes the same formula as Joint Entropy, but the probabilities represent the likelihoods that each intensity will occur at each random pixel in the image I:

$$H(I) = -\sum_{i \epsilon I} p(i) \log p(i) \tag{4.10}$$

Two of the most common methods to calculate these probability include using the (normalized) intensity frequencies as in the previous section, and Parzen Windowing. The entropy of an image characterized in this way can be thought of as computing the ‘predictability’ of the image intensities. Images that have a small number of high probability intensities will have low entropy as their intensities are relatively predictable. Conversely, images that have a more uniform distribution of probabilities will have high entropy, since several intensities are relatively equally predictable. Joint

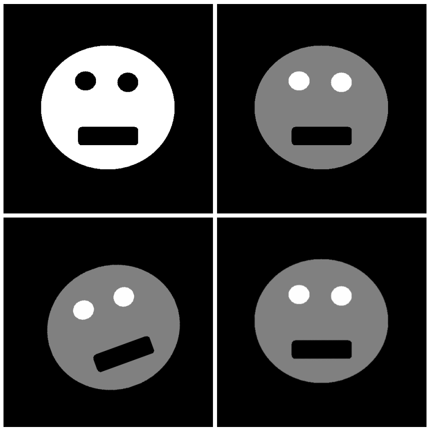


Figure 4.15: Coregistration by Maximization of Mutual Information. Top left: template volume. Top right: original input volume. The structures present in the images are identical and perfectly aligned, but have different intensity properties. Bottom left: The input image translated and rotated. Bottom right: The translated and rotated image coregistered with the template image by maximization of Mutual Information.

entropy for registration is often computed using a slightly different approach than the regional joint entropy used in the previous section. In registration, the joint entropy (Equation 4.11) is computed based on the entire region of overlap between the two volumes after transformation. Joint entropy is an appealing measure in this context to assess the quality of an inter-modality (or intra-modality) alignment. This can be related to the idea of 'predictability'. If two images are perfectly aligned, the intensities of the first image could significantly increase the predictability of the intensities in the aligned second image (and vice versa), since high probability intensity combinations will be present at the many locations of tissues with the same properties. However, if two images are misaligned, then the corresponding intensities in the second image will not be as predictable given the first image, since tissues in the first image will correspond to more random areas in the second image.

$$H(I_1, I_2) = - \sum_{i \epsilon I_1, j \epsilon I_2} p(i,j) \log p(i,j) \tag{4.11}$$

Minimizing joint entropy in general is not an appropriate measure to use unless images are already in fairly good alignment. This is due to the fact that a transformation that aligns background areas alone could achieve a low joint entropy [Pluim et al., 2003]. The Mutual Information measure addresses this shortcoming by including the entropies of each image in the region of overlap, and is commonly formulated as follows (with $H(i)$ being the entropy of the region of overlap from image i, and $H(i,j)$ being the joint entropy in the region of overlap from i and j):

$$MI(I_1, I_2) = H(I_1) + H(I_2) - H(I_1, I_2) \tag{4.12}$$

This measure will be high if the regions of overlap in the individual images have a high entropy (thus aligning background areas will result in a low score), but penalizes if the overlapping region has a high joint entropy (a sign of misalignment). This measure was originally applied to medical image registration by two different groups in [Collignon et al., 1995, Collignon, 1998, Viola, 1995, Wells et al., 1995]. It has gained popularity as an objective measure of an inter-modality alignment since it requires no prior knowledge about the actual modalities used, nor does it make assumptions about the relative intensities or properties of different tissue types in the modalities. The only assumption made is that the intensities of one image will be most predictable, given the other image, when 'interesting' regions of the images are aligned. Figure 4.15 shows a simple example of a registration that maximizes Mutual Information. Mutual information based registration is also appealing because it has well justified statistical properties, and it is computationally simple.

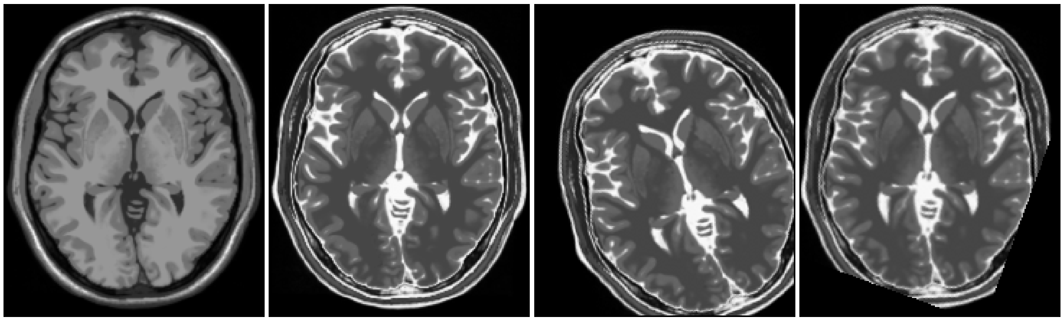


Figure 4.16: Coregistration to correct for a known transformation in a simulated MR image [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998]. Left to right: T1-weighted template image, corresponding T2-weighted image, T2-weighted image after translation and rotation, translated and rotated T2-weighted image coregistered with T1-weighted image. The missing areas in the coregistered images were cropped during the translation/rotation, but do not prevent the method from recovering the transformation.

One criticism of the Mutual Information metric is that in special cases it can decrease with increasing misalignment when images only partially overlap [Studholme et al., 1998]. In order to address this, the *Normalized Mutual Information* measure was proposed in [Studholme et al., 1998]:

$$NMI(I_1, I_2) = \frac{H(I_1) + H(I_2)}{H(I_1, I_2)} \tag{4.13}$$

This measure offers improved results over Mutual Information based registration, and is the measure we use for coregistration. We perform coregistration as a rigid-body transformation, and align each modality with a single target modality (the T1-weighted volume), the same modality that will be used in template registration. Modalities were typically well aligned in our test data, and this step was primarily included as a preventative measure in case data was encountered where the modalities were not aligned. Figure 4.16 demonstrates an example where Normalized Mutual Information was used to recover a known rigid-body transformation on simulated data of different modalities, while Figure 4.17 illustrates the minor improvement that was achieved through this step

Although Mutual Information based methods are very effective at the task of coregistration, their use of spatial information is limited to the intensities of corresponding pixels after spatial transformation. Although this allows accurate registration among a wide variety of both MR and other types of modalities, there are some modalities, such as ultrasound, where maximizing Mutual Information based on spatially corresponding intensity values may not be appropriate [Pluim et al., 2003]. Future implementations could utilize methods such as those discussed in [Pluim et al., 2003] that incorporate additional spatial information to improve robustness and allow the coregistration of a larger class of image modalities.

4.2.2 Template Registration

Linear template registration is the task of aligning the modalities with a template image in a standard coordinate system. Coregistration of the different modalities has already been performed, simplifying this task, since the transformation needs to only be estimated from a single modality. The computed transformation from this modality can then be used to transform the images in all modalities into the standard coordinate system. In the implemented system, linear template registration is included primarily as a preprocessing step for non-linear template registration. Computing the transformation needed for template registration is simpler than for coregistration, since the intensities between the template and the modality to be registered will have similar values. As with coregistration, there is a wealth of literature associated with this topic, and we refer to [Maintz and Viergever, 1998] for a review of methods. However, linear template registration is a relatively ‘easy’ problem com-

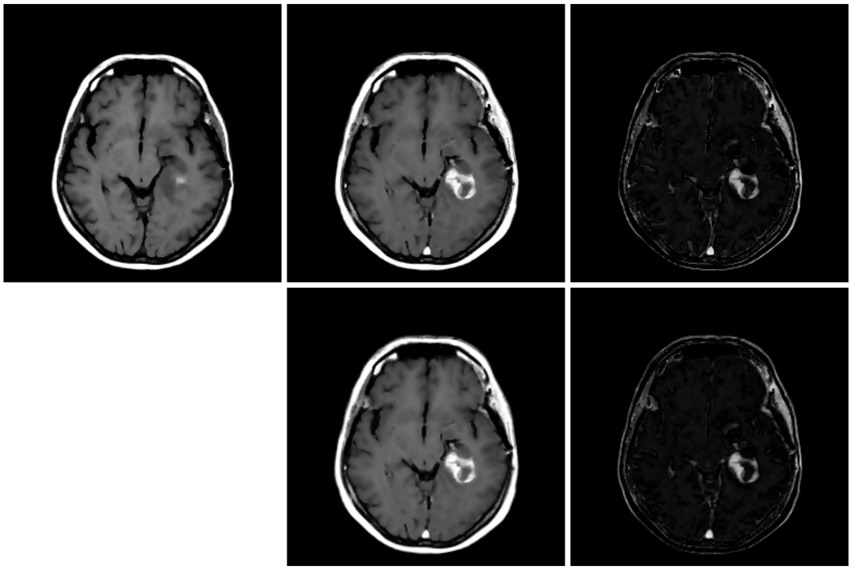


Figure 4.17: Coregistration of real data. Top, left to right: T1-weightd image, T1-weighted image after contrast injection, contrast absolute difference image. Bottom middle: T1-weighted image after contrast injection coregistered to the T1-weighted image. Bottom right: Contrast absolute difference image after coregistration. The difference is difficult to display in a stationary two-dimensional format, but the alignment between the images has been increased.

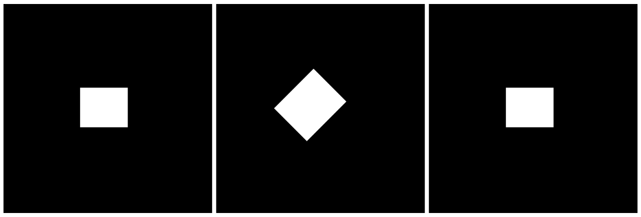


Figure 4.18: Linear registration to recover a known affine transformation. Left: Original image. Middle: Original image after translation, rotation, and resizing. Right: Translated/rotated/resized image after affine registration.

pared to coregistration and non-linear template registration, since straightforward metrics can be used to assess the registration (as opposed to coregistration), and the number of parameters to be determined is relatively small (as opposed to non-linear registration).

We have chosen to use the linear template registration method outlined in [Friston et al., 1995, Ashburner et al., 1997]. This method uses the simple mean squared error between the transformed image and the template as a measure of registration accuracy. It computes a linear 12-parameter *affine* transformation minimizing this criteria. This consists of one parameter for each of the three dimensions with respect to translation, rotation, scaling, and shearing. An additional parameter is used to estimate a linear intensity mapping between the two images, making the method more robust to intensity non-standardization. The method operates on smoothed versions of the original images to increase the likelihood of finding the globally optimal parameters, and uses the *Gauss-Newton* optimization method from [Friston et al., 1995] to efficiently estimate the 13 parameters. The result of applying this technique to a toy volume is shown in Figure 4.18.

The main contribution of [Ashburner et al., 1997] was the introduction of a (statistically sound) method of regularization into the registration process. As discussed earlier, regularization during parameter estimation can be thought of as introducing a ‘penalty’ for parameters that are determined to be less likely by some criteria. An alternate, but equivalent, way to interpret regularization is

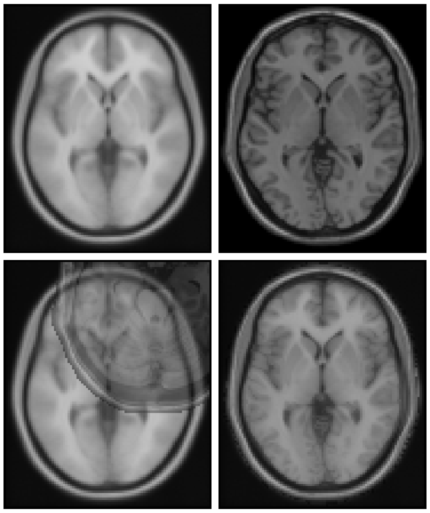


Figure 4.19: *Maximum a posteriori* linear registration of simulated data [BrainWeb, Online, Cocosco et al., 1997, Kwan et al., 1999, Kwan et al., 1996, Collins et al., 1998] with average data [ICBM View, Online]. Top left: Central slice of an average T1-weighted volume. Top right: Central slice of a simulated T1-weighted volume. Bottom left: Initial overlay of the simulated T1-weighted image over central slice of the average T1-weighted image. Bottom right: Overlay of the simulated T1-weighted image over the central slice of the average T1-weighted image after three-dimensional *Maximum a posteriori* registration.

that it considers not only the performance of the parameters (in this case the mean squared error resulting from using the parameters), but also simultaneously considers the likelihood of the parameters given prior knowledge of what the parameters should be. [Ashburner et al., 1997] uses a *maximum a posteriori* (MAP) formulation (Equation 4.14), an approach based on this interpretation of regularization. It assesses a registration based on both an assessment of the mean squared error after transformation with the parameters $p(b|a_p)$, and the prior probabilities for the parameter values $p(a_p)$. The advantages of assessing the prior probabilities of the transformations are that the algorithm converges in a smaller number of iterations, and the parameter estimation is more robust in cases where the data is not ideal. Examples where the data is not ideal for registration include cases with large inter-slice gaps exit or where anisotropic pixels are used. The disadvantage of including prior probabilities is that *meaningful* prior probabilities or expected values must be known. The parameters from the registration of 51 high-quality MR images were used to estimate the prior probabilities in [Ashburner et al., 1997]. Interesting results of this included that shearing should be considered unlikely in two of the image planes, and that the template used was larger than a typical head since zooms of greater than 1 in each dimension are expected. The result of applying this technique to align a simulated volume with an ‘average intensity’ volume are shown in Figure 4.19, while the results of aligning several real images with anisotropic pixels are shown in Figure 4.20.

$$p(a_p|b) = \frac{p(b|a_p)p(a_p)}{\int_q p(b|a_q)p(a_q)dq} \tag{4.14}$$

The method of [Friston et al., 1995, Ashburner et al., 1997] has several appealing properties, such as a method to account for intensity standardization, fast convergence to a regularized mean squared error solution, and robustness to anisotropic pixels, inter-slice gaps and other situations where the data is not ideal. Future implementations could use a cost function that is based on a measure of Mutual Information, rather than simply the mean squared error, this could confer additional robustness to intensity non-standardization.

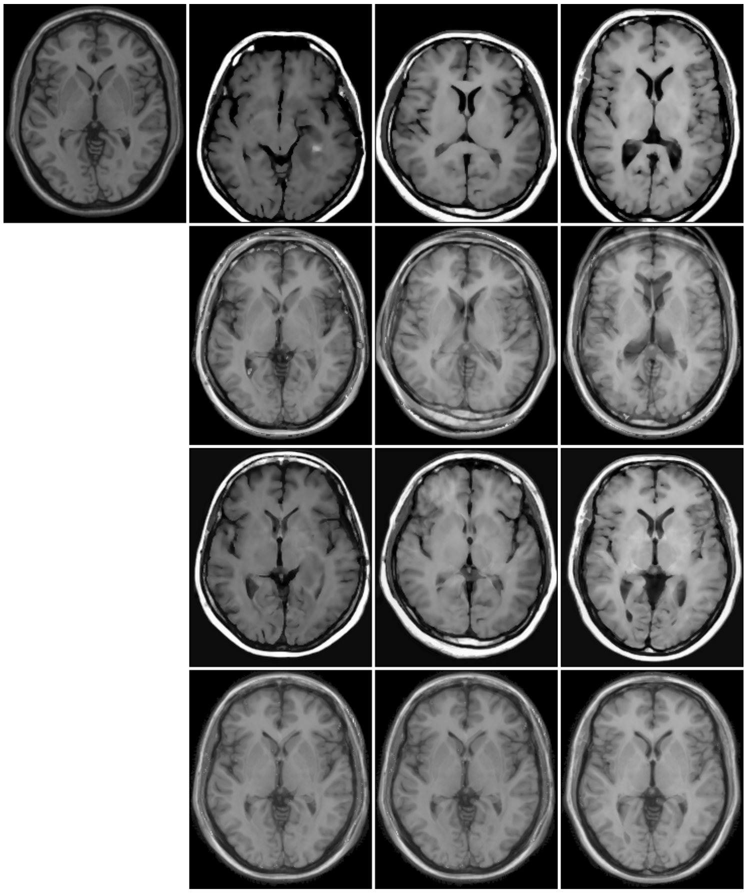


Figure 4.20: *Maximum a posteriori* linear registration of real volumes with anisotropic pixels. Top: Template used in registration (left) and the three input volumes. Second row: Overlay of the input volumes on the template after initial alignment of the image centers. Third row: Central slice after *Maximum a posteriori* registration. Bottom row: Overlay after registration. A linear transformation is clearly insufficient to exactly align the structures in the images, but the correspondence has been significantly increased.

4.2.3 Non-linear Warping

Non-linear warping to account for inter-patient anatomic differences after linear registration is becoming an increasingly researched subject. However, as with intensity inhomogeneity field estimation, it is difficult to assess the quality of a non-linear registration algorithm, since the optimal solution is not available (nor is optimality well-defined in this case). For our purposes, we would like a method that has shown to perform well in comparative studies based on objective measures, but we have an additional important constraint that must be placed on the registration method used. Since non-linear warping can cause local deformations, it is essential that a non-linear warping algorithm is selected that has an effective method of regularization.

[Hellier et al., 2001] performed an evaluation of four non-linear registration methods, and a linear Mutual Information based registration method for registering images of the brain. It was found that the non-linear methods improved several global measures (such as the overlap of tissue types) compared to the linear method. However, the non-linear methods gave similar results to the linear method for the local measures used (distances to specific cortical structures). A more recent study by the same group [Hellier et al., 2002] evaluated the method of [Ashburner and Friston, 1999], which is included in the SPM99 software package [SPM, Online] and is extremely popular in the neuroscience community (see [Hellier et al., 2002] for statistics). This method performed similarly to the

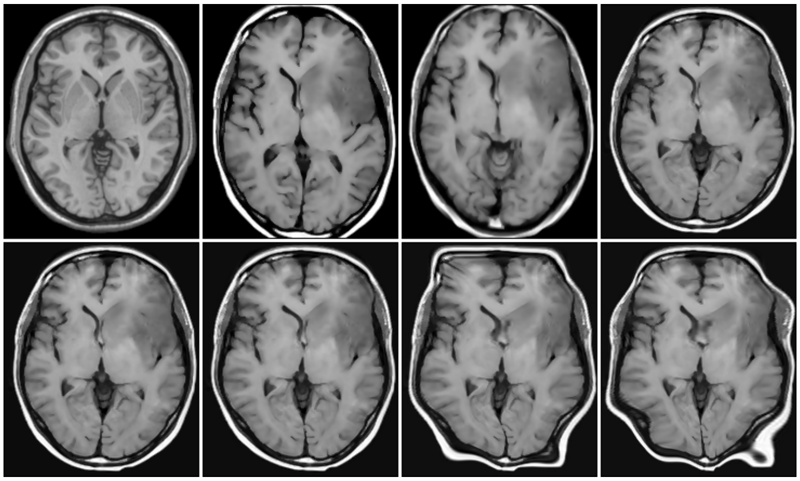


Figure 4.21: Registration with different degrees of freedom and different degrees of regularization. Top, left to right: Template image, central slice of input volume, input volume after 6-parameter rigid-body registration, input volume after 12-parameter affine registration. Bottom: Non-linear warping of the input image after affine registration. The degree of regularization decreases from left to right. Note that only brain pixels were used to estimate the error after warping. Warping clearly increases the correspondence between the input image and the template, but regularization is essential in order to prevent excessive deformation of the image.

other non-linear methods for the global measures. However, it performed significantly better than the linear and each of the non-linear methods for the local measures.

As with the linear method of registration discussed in the previous section, the method outlined in [Ashburner and Friston, 1999] works on smoothed images and uses a MAP formulation that minimizes the mean squared error subject to regularization in the form of a prior probability. The parameters of this method consist of a large number of non-linear spatial basis functions that are combined to define warps (392 for each of the three dimensions), in addition to four parameters that model intensity scaling and inhomogeneity. The basis functions used are the lowest frequency components of the Discrete Cosine Transform. The non-linear 'deformation field' is computed as a linear combination of these spatial basis functions. As opposed to linear template registration that use only a small number of parameters, the large number of parameters in this model means that regularization is necessary in order to ensure that spurious results do not occur. Without regularization over such an expressive set of parameters, the image to be registered could be warped to exactly match the template image (severely over-fitting). The prior probabilities are thus important to ensure that the warps introduced decrease the error enough to justify introducing the warp. Unlike in the linear method, this method does not compute the prior probabilities based on empirical measures for each parameter. Instead, the prior probability is computed based on the smoothness of the resulting deformation field, assessed by computing the Laplacians of the deformation field (referred to in this work as the 'membrane energy'). This form of prior biases the transformation towards a globally smooth deformation field, rather than a set of sharp local changes that cause an exact fit. Note that this does not exclude local changes, but it discourages changes from uniformity that do not result in a sufficiently lower squared error. The effects of varying this degree of regularization are seen in 4.21.

On top of its elegant formulation and its computational efficiency, the method presented in [Ashburner and Friston, 1999] has the advantage that it is trivial to vary the degree of regularization. For the task of non-linearly registering images that could potentially have large tumors, a higher degree of regularization is likely needed than for registering images of normal brains. The algorithms appealing properties, wide use, and performance in comparative studies make the method of [Ashburner and Friston, 1999] an obvious choice for a non-linear registration algorithm. The fi-

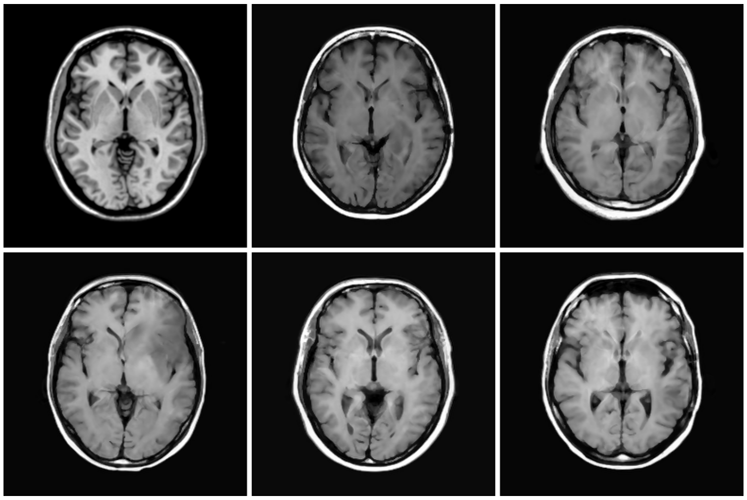


Figure 4.22: Template warping of different subjects. Top left: Template image. Others: Images non-linearly warped to the template image.

nal results of registering several of the images in our experimental data with a template are seen in Figure 4.22. As with the linear template registration method, future implementations could examine methods that use Mutual Information based measures for this step, in order to improve robustness to intensity non-standardization. For references regarding non-linear Mutual Information based registration, we again refer to [Pluim et al., 2003]. An alternate (or parallel) direction to explore for future implementations would be to use multiple modalities to enhance the registration. Finally, although the regularization method used is effective at preventing excessive warping, it is not necessarily biologically motivated. Future implementations could explore methods that take advantage of the image information in order to provide more appropriate regularization.

4.2.4 Spatial Interpolation

After a transformation has been computed, an interpolation method is required to assign values to pixels of the transformed volume at the new pixel locations. This involves computing, for each new pixel location, an interpolating function based on the intensities of pixels at the old locations. The choice of an effective interpolation algorithm is important, since some methods will introduce more interpolation artifacts into the image than others (see Figure 4.23). Interpolation is an interesting research problem, that has a long history over which an immense amount of variations on similar themes have been presented. We highly recommend [Meijering, 2002] for an extensive survey and history of data interpolation. This article also references a large number of comparative studies of different methods for medical image interpolation. The conclusion drawn based on these evaluations is that *β-spline* kernels are in general the most appropriate interpolator.

Interpolation with β-splines involves modeling the image as a linear combination of piecewise polynomial (β-spline or Haar) basis functions, which are defined recursively as follows [Hastie et al., 2001]:

$$B_{i,1}(x) = 1 \ \ if \ \ \tau_i \leq x < \tau_{i+1}, \ \ otherwise \ \ 0. \tag{4.15}$$

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x) \tag{4.16}$$

In this formulation, $B_{i,m}$ is the ith β-spline of order m, with knots τ. Given an image represented as a linear combination of such basis functions, the intensity value of the image at any real-valued location can be determined. This approach is related to interpolation using radial basis

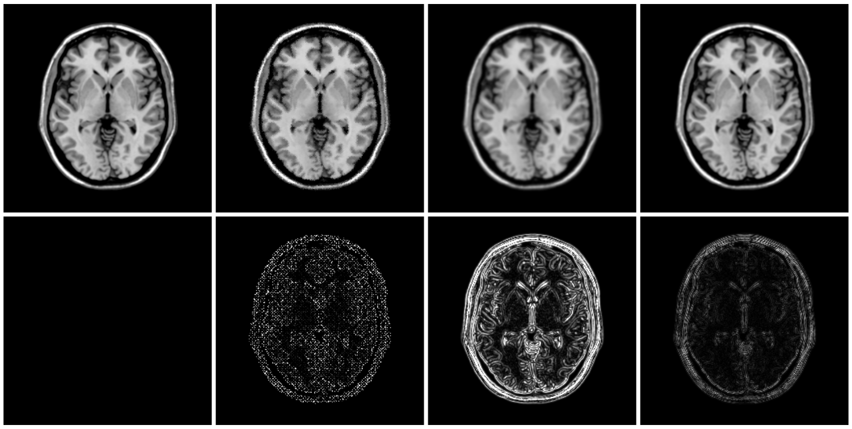


Figure 4.23: A test of interpolation algorithms. The top left image was rotated 4 times and resized 4 times. The other top images show the final results of using different interpolation algorithms to interpolate the intermediate volumes. Left to right: Nearest neighbor, linear, and cubic interpolation. Bottom: the absolute difference (multiplied by ten) between the interpolated image and the original volume (darker is better). Note that some interpolation strategies inherently accumulate more error compared to the original image.

functions such as thin-plate splines and Hardy's multiquadratics [Lee et al., 1997]. For higher-order β-splines, as with radial basis functions, modeling the image as a linear combination of spatially varying basis functions results in an interpolating surface that fits the known data points exactly, has continuous derivatives, and appropriately represents edges in the image. The coefficients of the β-spline basis functions that minimize the mean squared error can be determined in cubic time using the pseudoinverse as is typically done with radial basis function interpolation schemes. Unfortunately, cubic time is computationally intractable for the data set sizes being examined (even small three-dimensional volumes may have over one million data points). However, computing the β-spline coefficients can be done using an efficient algorithm based on recursive digital filtering [Unser et al., 1991]. This results in an interpolation strategy that is extremely efficient given its high accuracy. Figure 4.24 shows the results of different interpolation strategies for interpolating after non-linear registration, including β-splines.

A noteworthy point with respect to MRI interpolation with β-splines is that it has been shown that β-splines converge to the ideal Sinc interpolating filter as their degree increases (see [Ashburner and Friston, 2003b]). Convolution by a Sinc function is the closest 'real space' approximation to Fourier interpolation [Ashburner and Friston, 2003b], and 'windowed sinc' approximations of the Sinc function are thus a popular interpolator for MRI data. However, windowed sinc interpolation is not necessarily the ideal method for interpolating data that may have anisotropic voxels, and β-splines have outperformed windowed sinc methods in comparative studies based on MRI and other data types (see [Meijering, 2002].

We naturally chose to use a high-order β-splines for spatial interpolation, given their high accuracy and low computational expense. Future implementations could examine radial basis function interpolation methods such as thin-plate and polyharmonic splines, and Hardy's multiquadratics. Recently, efficient methods have been proposed for determining the coefficients for these basis functions [Carr et al., 1997, Carr et al., 2001]. Another method that could be explored in the future is Kriging, a Bayesian interpolation method [Leung et al., 2001].

4.2.5 Summary

The registration phase consists of four steps: coregistration of the different modalities, linear affine template registration, non-linear template warping, and spatial interpolation. The coregistration step registers each modality with a target modality by finding a rigid-body transformation that maxi-

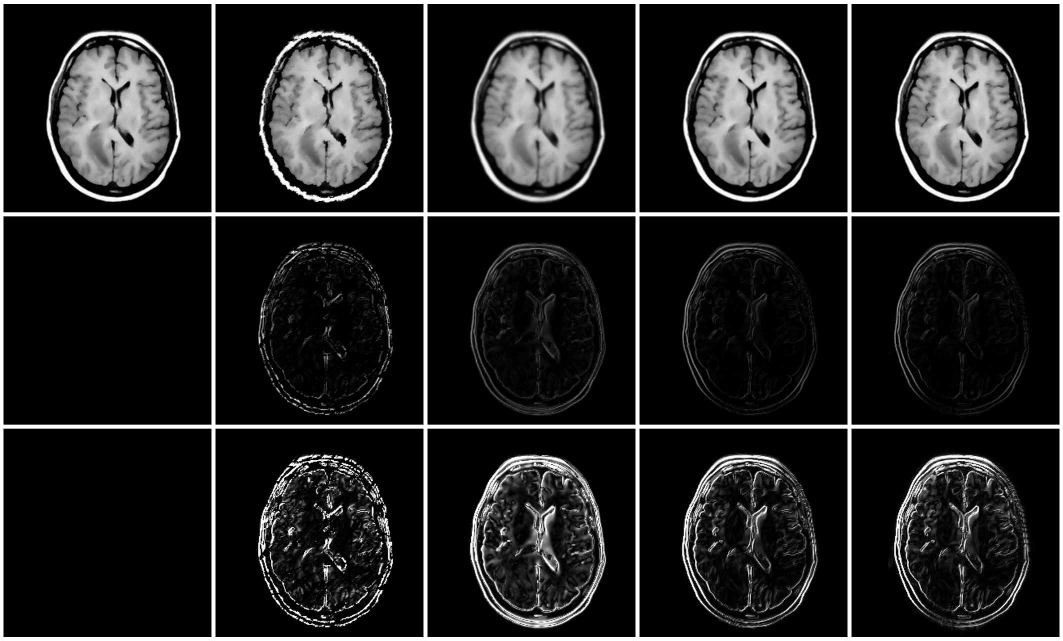


Figure 4.24: A test of interpolation algorithms on real data for linear and non-linear template registration. The top left image went through 6 affine transformations and 6 non-linear warpings using different interpolation algorithms to interpolate the intermediate volumes. Top, left to right: Original image, final interpolation using Nearest Neighbor, Trilinear, low-order β-Spline, and high-order β-spline interpolation. Middle: Absolute difference between the final interpolated images and the original image. Bottom: Absolute difference between the final interpolated images and the original image multiplied by ten. In this test, the β-spline methods have the smallest accumulated error.

mizes the Normalized Mutual Information measure. The T1-weighted image before contrast injection is used as the target modality. Linear template registration is then performed by computing a MAP transformation that minimizes the regularized mean squared error between the target modality and the template. The T1-weighted image before contrast injection is used to compute the parameters, and transformation of each of the coregistered modalities is performed using these parameters. As a template, the averaged single subject T1-weighted image from the ICBM View software was used [ICBM View, Online], which is related to the 'colin27' (or 'ch2') template from [Holmes et al., 1998]. Non-linear template warping refines the linear template registration by allowing warping of the image with the template to account for global differences in head shape and other anatomic variations. The warping step computes a MAP deformation field that minimizes the (heavily) regularized mean squared error. The regularization ensures a smooth deformation field rather than a large number of local deformations. The same template is used for this step, and as before the T1-weighted pre-contrast image is used for parameter estimation and the transformation is applied to all modalities. The final step is the spatial interpolation of pixel intensity values at the new locations. High-order polynomial β-spline interpolation is used, modeling the image as a linear combination of β-spline basis functions. This technique has attractive Fourier space properties, and has proved to be the most accurate interpolation strategy given its low computational cost. To prevent interpolation errors from being introduced between registration steps, spatial interpolation is *not* performed after coregistration or linear template registration. The transformations from these steps are stored, and interpolation is done only after the final (non-linear) registration step.

Each of the four registration steps are implemented in the most recent Statistical Parametric Mapping software package (SPM2) from the Wellcome Department of Imaging Neuroscience [SPM, Online]. For coregistration, we used the Normalized Mutual Information measure and default parameter values. For template registration, several modifications were made to the default behavior. The template image was smoothed with an 8mm full width at half maximum kernel to decrease the likelihood of reaching a local minimum. The regularization (λ) value was increased to 10 (heavy). The pixel resolution of the output volumes was changed to be ($1mm$ by $1mm$ by $2mm$), and the bounding box around the origin was modified to output volumes conforming to the template's coordinate system. The interpolation method was changed from trilinear interpolation to β-spline interpolation, and used polynomial β-splines of degree 7. The transformation results were stored in .mat files, allowing transformations computed from one volume to be used to transform others, and allowing interpolation to be delayed until after non-linear registration.

4.3 Intensity Standardization

Intensity standardization is the vital step that allows the intensity values to approximate an anatomical meaning. This subject has not received as significant of a focus in the literature as intensity inhomogeneity correction, but research effort in this direction has grown in the past several years. This is primarily due to the fact that it can remove the need for patient specific training or the reliance on tissue models, which may not be available for some tasks or for some areas of the body. This section will survey the literature relating to intensity standardization, before presenting our approach. Although EM-based methods that use spatial priors are an effective method of intensity standardization, they will not be revisited here, since they have been covered elsewhere, and they are not appropriate for this step in the framework.

The intensity standardization method used by the INSECT system was (briefly) outlined in [Zijdenbos et al., 1995], in the context of improving Multiple Sclerosis lesions segmentation, and was discussed earlier in this document in the context of inter-slice intensity variation reduction. This method estimates a linear coefficient between the image and template based on the distribution of 'local correction' factors. Another study focusing on intensity standardization for Multiple Sclerosis lesion segmentation was presented in [Wang et al., 1998], that compared four methods of intensity standardization. The first method simply normalized based on the ratio of the mean intensities between images. The second method scaled intensities linearly based on the average white matter intensity (with patient-specific training). The third method computed a global scale factor using a "machine parameter describing coil loading according to reciprocity theorem", computing a transformation based on the voltage needed to produce a particular 'nutation angle' (that was calibrated for the particular scanner that was used). The final method examined was a simple histogram matching technique based on a non-linear minimization of squared error applied to 'binned' histogram data, after the removal of air pixels outside the head (this outperformed the other three). In [Nyul and Udupa, 1999], another histogram matching method was presented (later made more robust in [Nyul et al., 2000]), that computed a piecewise intensity scaling based on 'landmarks' in the histogram. Similar to previous works on intensity standardization, this study also demonstrated that intensity standardization could aid in the segmentation of Multiple Sclerosis lesions. This method was later used in a study that evaluated the effects of inhomogeneity correction and intensity standardization [Madabhushi and Udupa, 2002], finding that these steps complemented each other, but that inhomogeneity correction should be done prior to intensity standardization. Another method of intensity standardization was presented in [Christenson, 2003], that normalized white matter intensities using histogram derivatives. One method, mentioned in Chapter 2, that was used as a preprocessing step in a tumor segmentation system was presented in [Shen et al., 2003]. This method thresholded background pixels, and used the mean and variance of foreground pixels to standardize intensities. A similar method was used in [Collowet et al., 2004], comparing it to no standardization, scaling based on the intensity maximum, and scaling based on the intensity mean.

The methods discussed above are relatively simple and straightforward. Each method (with the exception of [Zijdenbos et al., 1995]) uses a histogram matching method that assumes either a simple distribution or at least a close correspondence between histograms. These assumptions can be valid for controlled situations, where the protocols and equipment used are relatively similar, and only minor differences exist between the image to be standardized and the template histogram. However, in practice this may not be the case, as histograms can take forms that are not well characterized by simple distributions, in addition to potential differences in the shapes of the input and template image histograms. This relates to the idea that a term like ‘T1-weighted’ does not have a correspondence with absolute intensity values, since there are a multitude of different ways of generating a T1-weighted image, and the resulting images can have different types of histograms. Furthermore, one ‘T1-weighted’ imaging method may be measuring a slightly different signal than another, meaning that tissues could appear with different intensity properties on the image, altering the histogram.

A more sophisticated recent method was presented in [Weisenfeld and Warfield, 2004]. This method used the Kullback-Leibler (KL) divergence as a measure of relative entropy between an image intensity distribution and the template intensity distribution. This method computed an inhomogeneity field that minimized this entropy measure, and thus simultaneously corrected for intensity inhomogeneity and performed intensity standardization. Relative entropy confers a degree of robustness to the histogram matching, but even this powerful method fundamentally relies on a histogram matching scheme and ignores potentially relevant spatial information. Without the use of spatial information to ‘ground’ the matching by using the image-specific characteristics of tissues, standardizing the histograms does not necessarily guarantee a standardization of the intensities of the different tissue types. The EM-based approaches (that use spatial priors) can perform a much more sophisticated intensity standardization, since the added spatial information in the form of priors allows individual tissue types to be well characterized. By using spatial information to locate and characterize the different tissue types, the standardization method is inherently standardizing the intensities based on actual tissue characteristics in the image modalities, rather than simply aligning elements of the histograms.

To date, most intensity standardization methods rely on a tissue model, or a relatively close match between the histograms of the input image and the template image. The presence of potentially large tumors makes it difficult to justify these assumptions, and the only work we are aware of that used an explicit intensity standardization step as a preprocessing phase for tumor segmentation assumes a simple uni-modal intensity distribution (after subtraction of background pixels) [Shen et al., 2003], while bi-modal (or higher) distributions are evident in typical T1-weighted and T2-weighted images even when viewed at courser scales. Furthermore, we are not aware of any existing methods that incorporate a means to reduce the effects of tumors and edema pixels (that are not present in the template image) on the estimation of the standardization parameters without the use of a tissue model. Thus, for this implementation, a simple method of intensity standardization was developed that is related to the proposed approach for inter-slice intensity variation reduction discussed earlier.

Our method for inter-slice intensity variation reduction uses spatial information between adjacent slices to estimate a linear mapping between the intensities of adjacent slices, but used simple measures to weight the contribution of each corresponding pixel location to this estimation. For intensity standardization, the problems that complicate the direct application of this approach are determining the corresponding locations between the input image and the template image, and accounting for outliers (tumor, edema, and areas of mis-registration) that will interfere in the estimation. Determining the corresponding locations between the input image and the template was trivial for inter-slice correction, since we assumed that adjacent slices would in general have similar tissues at identical image locations. Although typically not trivial for intensity standardization, non-linear template registration has been performed, thus the input image and template will already be aligned, and it can be assumed that locations in the input image and the template will have similar tissues.

In inter-slice intensity correction, the contribution of each pixel pair was weighted in the parameter estimation based on a measure of regional spatial correlation and the absolute intensity difference, which made the technique robust to areas where the same tissue type was not aligned. Since the input image will not be exactly aligned with the template image in the case of intensity standardization, these weights can also be used to make the intensity standardization more robust.

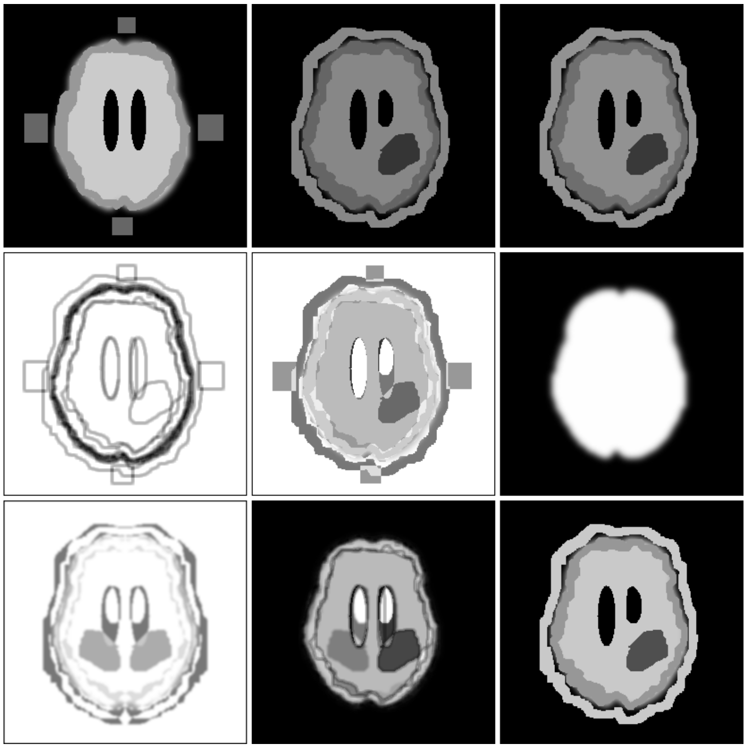


Figure 4.25: Intensity Standardization of a toy volume. Top left: Toy template image (consisting of gray matter, white matter, CSF, and fiducial markers. Top middle: Toy image to be standardized, that is slightly different anatomically, has fat visible outside of the skull, a large tumor, and no fiducial markers. Top right: The (poor) results obtained by unweighted linear regression. Middle row: Different elements of the pixel weighting. Left to right: Regional joint entropy, absolute difference, and brain area prior probability. The entropy and absolute difference have the same effect as before, but the brain probability allows restriction of the estimation to the brain area, rather than all foreground pixels. Bottom left: Symmetry weighting (note the low weight assigned to the tumor). Bottom middle: Combined weighting, indicating the estimation will place the largest weight on common gray matter, white matter, and csf regions. Bottom right: The results of weighted linear regression with the combined weighting for intensity standardization.

However, intensity standardization is complicated by the presence of tumors and edema, which are areas that may be homogeneous and similar in intensity to the corresponding region in the template, but that should not significantly influence the estimation. To account for this, we use a measure of regional symmetry as an additional factor in computing the weights used in the regression. The motivation behind this is that regions containing tumor and edema will typically be asymmetric [Gering, 2003a, Joshi et al., 2003]. Thus, giving less weight to asymmetric regions reduces the influence that abnormalities will have on the estimation.

A simple measure of symmetry is used, since the images have been non-linearly warped to the template where the line of symmetry is known. The first step in computing symmetry is computing the absolute intensity difference between each pixel and the corresponding pixel on the opposite side of the known line of symmetry. Since this estimation is noisy and only reflects pixel-level symmetry, the second step is to smooth this difference image with a 5 by 5 Gaussian kernel filter (the standard deviation is set to 1.25), resulting in a smoothly varying regional characterization of symmetry. Although symmetry is clearly insufficient to distinguish normal from abnormal tissues since normal areas may also be asymmetric, this weighting is included to decrease the weight of potentially bad areas from which to estimate the mapping, and thus it is not important if a small number of tumor pixels are symmetric or if a normal area is asymmetric.

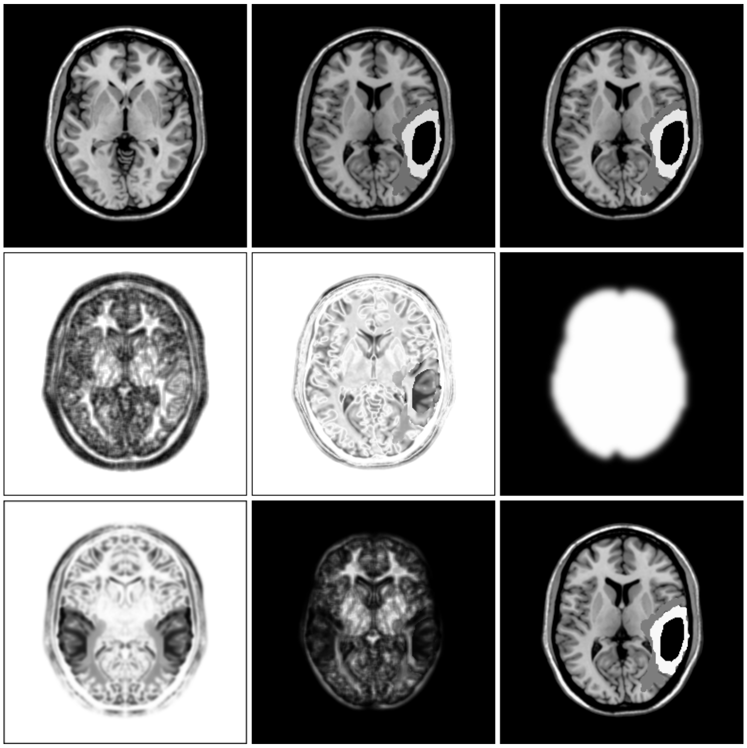


Figure 4.26: Intensity Standardization with a synthetic tumor to recover a known intensity offset. Top, left to right: Template image, image to be standardized with a synthetic tumor and an applied intensity offset, results of unweighted linear regression. Middle, left to right: Regional joint entropy weighting, absolute difference weighting, spatial brain prior weighting. Bottom, left to right: Symmetry weighting, combined weighting, result of weighted linear regression for intensity standardization. Note that the weighting makes the estimation primarily based on shared white matter regions, and reduces the tumor area's effect on the estimation.

The final factor that is considered in the weighting of pixels for the intensity standardization parameter estimation is the spatial prior 'brain mask' probability in the template's coordinate system (provided by the SPM2 software [SPM, Online]). This additional weight allows pixels that have a high probability of being part of the brain area to receive more weight than those that are unlikely to be part of the brain area. This additional weight ensures that the estimation focuses on areas within the brain, rather than standardizing the intensities of structures outside the brain area, that are not as relevant to the eventual segmentation task.

The weighted linear regression is performed between the image and the template in each modality. The different weights used are the regional joint entropy, the absolute difference in pixel intensities, the regional symmetry measured in each modality, and the brain mask prior probability. These are each normalized to be in the range [0,1], and the final weight is computed by multiplying each of the weights together (assuming independence). The values for $R(i)$ are thus computed as follows (with $S(X(i))$ denoting the measure of symmetry at pixel i in image X, $P_{brain}(i)$ being the spatial prior probability that the pixel is part of the brain, and $H(X(i), Y(i))$ defined as before):

$$r(i) = \frac{(N_1 - H(X(i), Y(i)))}{N_1} \frac{(N_2 - |X(i) - Y(i)|)}{N_2} \frac{(N_3 - S(X(i)))}{N_3} P_{brain}(i) \tag{4.17}$$

This method was implemented in Matlab [MATLAB, Online], and is applied to each slice rather than computing a global factor to ease computational costs. The results of applying this technique to toy data and data with a synthetic tumor to recover a known intensity offset are shown in Fig-

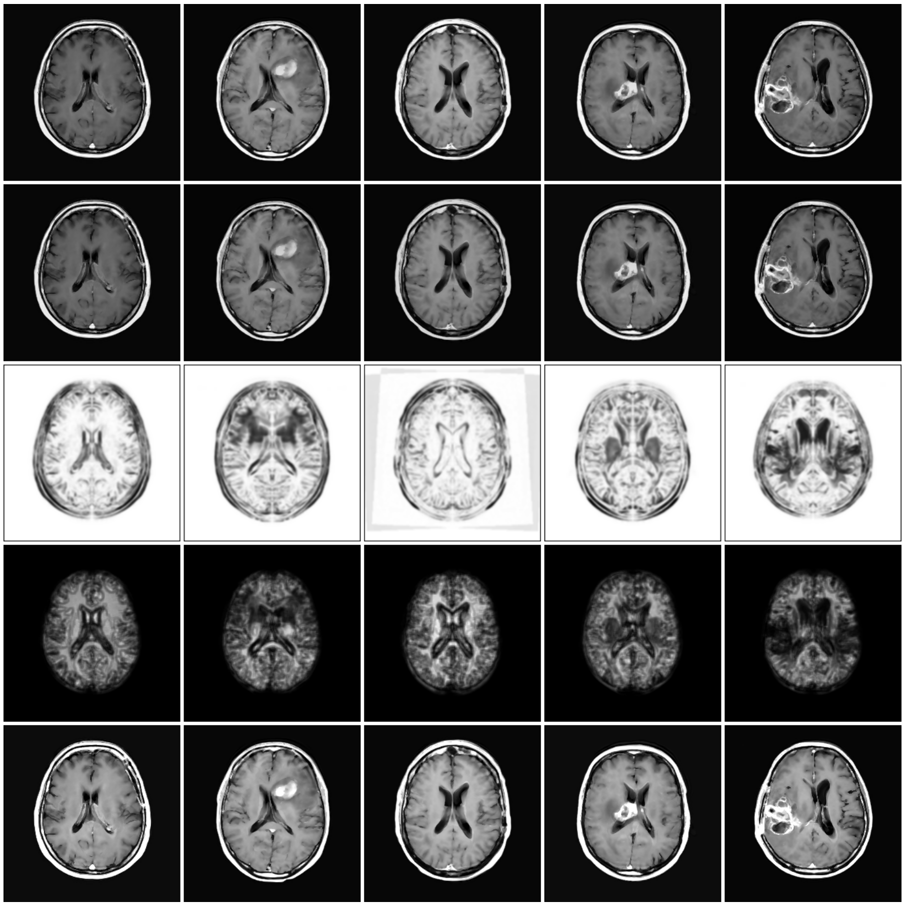


Figure 4.27: Intensity Standardization of real data. Top row: T1-weighted images from 5 patients. Second row: Intensity Standardization based on unweighted linear regression. Third row: Symmetry weighting based on T1-weighted and (coregistered) T2-weighted image. The three abnormal regions have clearly had their weight reduced. Fourth row: Combined weighting. The estimation for most of the images is primarily based on white matter regions, although some images also have high weights assigned to csf and gray matter regions. Bottom row: The results of Intensity Standardization. It is obvious that the differences in intensity between images have been significantly reduced.

ures 4.25 and 4.26, respectively. The application of this technique to real data (from different sites) to standardize the intensities between images is demonstrated in Figure 4.27.

There are several methods that could be explored to improve this step in future implementations. Different loss functions could be examined, since loss functions such as the absolute error and the Huber loss are more robust to outliers than the squared error measure used here [Hastie et al., 2001], though at a higher computational expense. In general, we found that non-linear transformations could further reduce the average error between the images, but this came at the cost of reduced contrast in the images. This occurred even when using a simple additive factor in addition to the linear scale factor. Future work could further explore non-linear methods that incorporate regularization to allow non-linear intensity standardization that is constrained to preserve image contrast. Although methods based on tissue models have been purposely avoided up to this point in the implemented system, this may be a step that could benefit from a tissue model, especially if the technique will be applied for large data sets where intensity standardization will be a larger problem. One direction to explore with respect to this idea could be to use a method similar to the tissue estimation performed in [Prastawa et al., 2004], that used spatial prior probabilities for gray matter, white matter, and CSF to build a tissue model, but used an outlier detection scheme to make the estimation more robust. The weighting methods discussed in this section, and symmetry in particular, could be incorporated into an approach similar to this strategy to potentially achieve more effective intensity standardization.

4.4 Feature Extraction

At this point, the images have been non-linearly registered with an atlas in a standard coordinate system, and have undergone significant processing to reduce intensity differences within and between images. However, the intensity differences have only been reduced, not eliminated. If the intensity differences were eliminated, then a multi-spectral thresholding might be a sufficiently accurate pixel-level classifier to segment the images, and feature extraction would not be necessary. Since the differences have only been reduced and ambiguity remains in the multi-spectral intensities, we cannot rely on a simple classifier that solely uses intensities. This section will highlight the many pixel features that have been implemented to improve an intensity based pixel classification. Experimentation with these different features will be explored in the next chapter, and it should be noted that not all of the features presented in this section were included in the final implementation.

4.4.1 Image-Based Features

Since considerable effort has been spent towards normalizing the intensities, the most obvious features to use are the standardized pixel intensities in each modality. Apart from including these features in some form, there can remain considerable uncertainty as to what are the best features. A simple set of additional features to use for a pixel-level classifier are the intensities of neighboring pixels, as was used in [Dickson and Thomas, 1997, Garcia and Moreno, 2004].

Including neighborhood intensities introduces a large amount of redundancy and added complexity to the feature set, that may not aid in discrimination. A more compact representation of the intensities within a pixel's neighborhood can be obtained through the use of multi-scale features. Multiple scales of an image are often obtained by repeated smoothing of the image with a Gaussian filter and reductions of the image size. This is typically referred to as the *Gaussian Pyramid* [Forsyth and Ponce, 2003], and produces a set of successively smoothed images of decreasing size. Higher layers in the pyramid will represent larger neighborhood aggregations. The images constructed from a Gaussian pyramid for different patients and modalities are demonstrated in Figure 4.28. This forms a more compact representation of neighborhood properties, since a small amount of features capture a similar amount of information (though some information is inevitably lost with this representation). The use of multi-resolution features also has the advantage that it implicitly encourages neighboring pixels to be assigned the same label (with many classifiers), since the feature values of neighboring pixels at low resolutions will be very similar. Conversely, a similarity between the feature values of neighboring pixels is not necessarily guaranteed if the intensities of neighboring pixels are used as features.

One drawback of using a Gaussian Pyramid approach is that a significant amount of information is lost at the lower resolutions. This is especially evident when viewing the upper layers of the pyramid at the same size as the original image. Depending on the interpolation algorithm used, the higher layers can have a blocky appearance (in the case of nearest neighbor interpolation), or can introduce spurious spatial information (in the case of linear or cubic methods). In considering that these features will be used to classify individual pixels, it is clear that discarding the small differences between neighboring pixels when decreasing the image size will not help in discrimination. The primary motivation for performing re-sampling has traditionally (and even recently) been cited as computational cost [Forsyth and Ponce, 2003]. Although this is essential in some applications where Gaussian pyramids are used (such as hierarchical Markov Random Fields), in this application there is no benefit in computational expense to a smaller image size at coarse resolutions, if we consider convolution as a computationally reasonable operation. We thus explored the use of a *Gaussian Cube* multi-scale feature representation, where each layer in the cube is computed by convolution

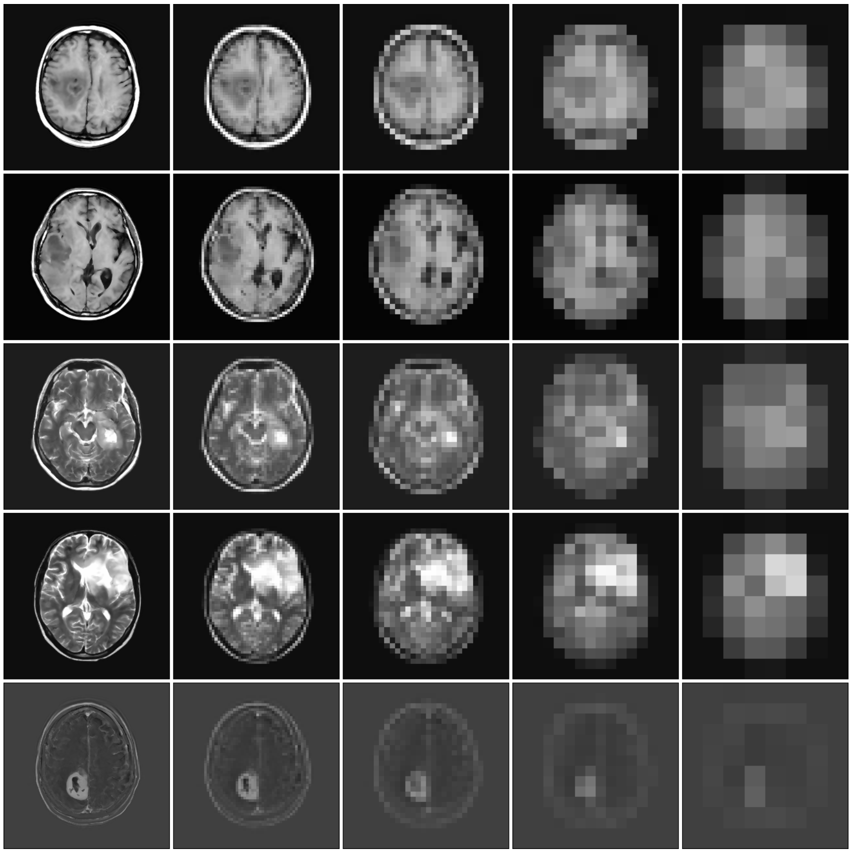


Figure 4.28: Mutli-scale Gaussian Pyramid for different patients and modalities. Left to right: Original image, layer 2, layer 3, layer 4, layer 5. The first two rows are constructed from T1-weighted images, the next two from T2-weighted images, and the final row from the T1-weighted contrast agent difference image. Each row is from a different patient. Although these images represent regional information, significant spatial information is lost at upper layers of the pyramid, resulting in a 'blocky' effect.

of the original image with a Gaussian kernel of increasing size and variance. This approach is illustrated in Figure 4.29, and is similar to that used to compute several of the texture features in [Leung and Malik, 2001].

An additional advantage of using a Gaussian Cube representation for multi-resolution features is that linear combinations of these features will form differences of Gaussians, a traditional method of edge detection similar to the Laplacian of Gaussian operator [Forsyth and Ponce, 2003] (Equation 4.18 is the 2D Laplacian function $\Delta^2 f(x, y)$). Thus the Gaussian cube explicitly encodes low-pass information but can also implicitly encode high-pass information if linear combinations of the features are considered. Also explored was using different sizes of the Laplacian of Gaussian filter to form a Laplacian Cube, as illustrated in Figure 4.30.

$$\Delta^2 f(x, y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{4.18}$$

Two methods were explored for incorporating intensity distribution based information into the features. The first method of incorporating histogram information we explored was to calculate a simple measure of the histogram density at the pixel's location. This was inspired by the 'density screening' operation used in [Clark et al., 1998], and is a measure of how common the intensity is within the image. The density was estimated by dividing the multi-spectral intensity values into equally sized bins (cubes in the multi-spectral case). This feature was computed as the (log of)

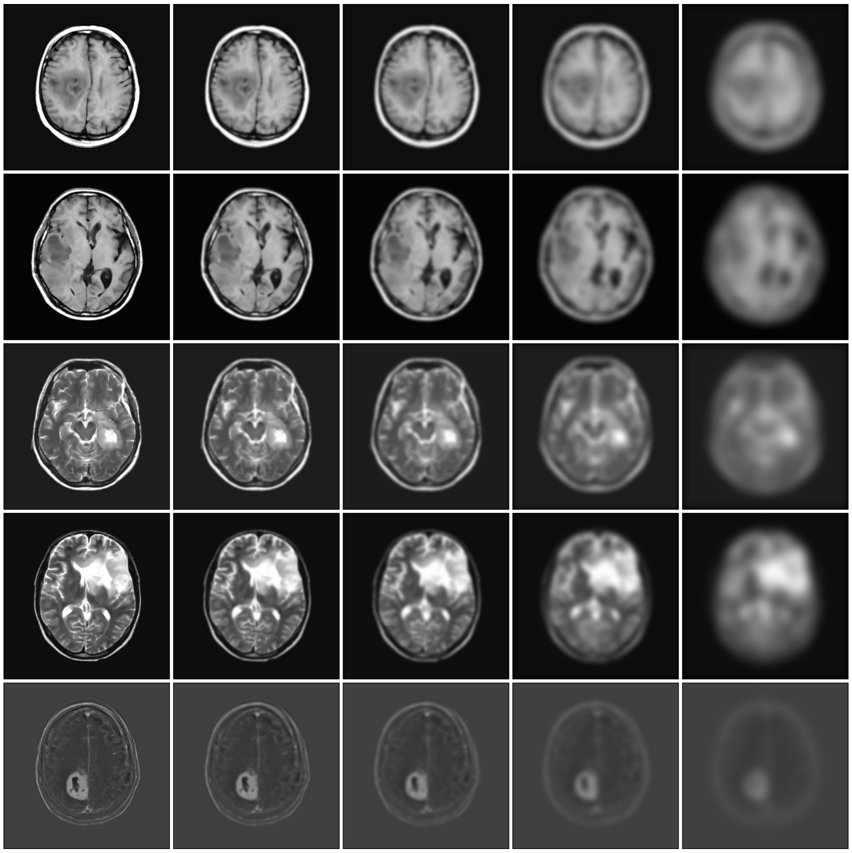


Figure 4.29: Mutli-scale Gaussian Cube of the images from Figure 4.28. The upper (rightmost) layers of the cube still represent regional intensity information, but the images smoothly vary spatially and still closely resemble the coarse-scale structures present in the original image.

the number of intensities within the pixel's intensity's bin. The second type of feature explored to take advantage of intensity distribution properties was computing a 'distance to normal intensity tissue' measure. These features computed the multi-spectral Euclidean distance from the pixel's multi-spectral intensities to the (mean) multi-spectral intensities of different normal tissues in the template's distributions (that the images have been standardized to). Figure 4.31 shows examples of the multi-spectral *histogram density* and the *distance to normal intensity* features.

A set of important image-based features are texture features. There are a large variety of methods to compute features that characterize image textures. Reviews of different methods can be found in [Materka and Strzelecki, 1998, Tuceryan and Jain, 1998], and more recent surveys can be found in [Forsyth and Ponce, 2003, Hayman et al., 2004]. In the medical image processing literature, the most commonly used features to characterize textures are the 'Haralick' features, a set of statistics computed from a gray-level spatial coocurrence matrix [Haralick et al., 1973]. The coocurrence matrix is an estimate of the likelihood that two pixels of intensities i and j (respectively) will occur at a distance d and an angle of θ within a neighborhood. The matrix is often constrained to be symmetric, and the original work proposed 14 statistics to compute based on this matrix, including measures such as angular second momentum, contrast, correlation, and entropy. The statistical values computed from the coocurrence matrix represent the texture parameters, and are typically calculated for a pixel by considering a square window around the pixel. Variations on these types of texture features, that are often combined with 'first order' features, and have shown to provide increased discrimination between tumor pixels and normal pixels compared to purely intensity-based methods [Schad et al., 1993, Kjaer et al., 1995, Herlidou-Meme et al., 2003, Mahmoud-Ghoenim et al., 2003].

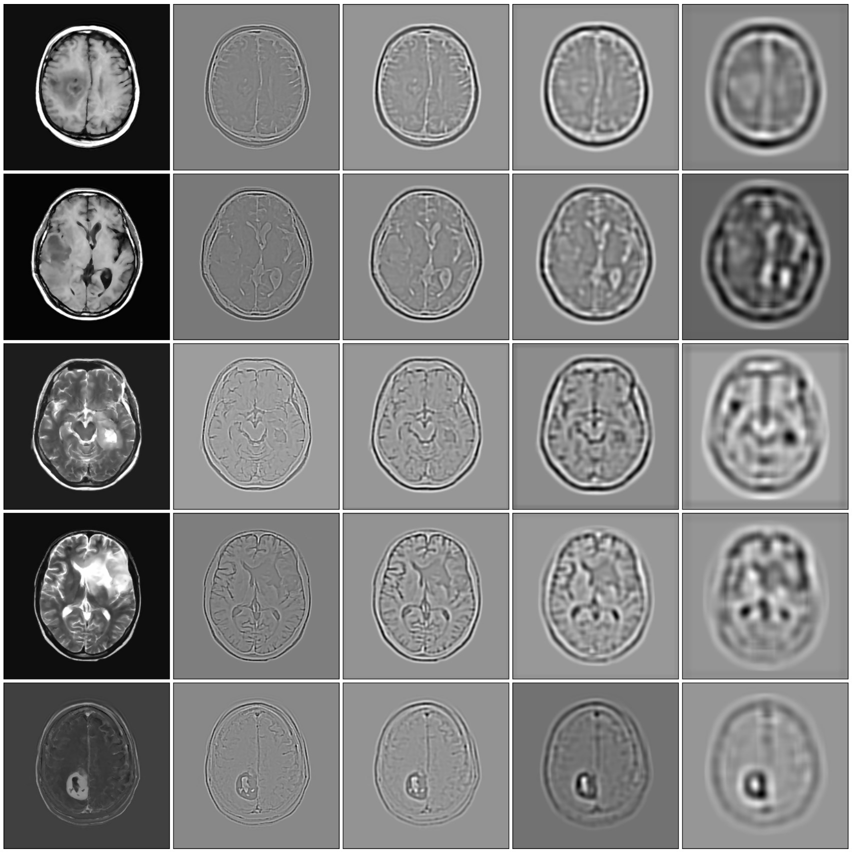


Figure 4.30: Mutli-scale Laplacian Cubes of the images from the two previous Figures. The Laplacian of Gaussian filter outputs have large magnitudes of responses near edges, and are also indicative of whether a region is ‘darker’ or ‘brighter’ than its surroundings. Note that the bottom (leftmost) layer of the Laplacian Cube is still the original images. The filter response images have been scaled to the range [0,1] before display.

A major factor to be considered in computing these features is the method through which the coocurrence matrix is constructed. In our experiments, the coocurrence matrix was constructed by considering only pixels at a distance of exactly 1 from each other, and computing the estimate between intensity i and j at this distance independent of the angle. The intensities were divided into equally sized bins to reduce the sparsity of the coocurrence matrix. More sophisticated methods that could have been used include evaluating different distances or angles, smoothing the estimates, or weighting the contribution of pixel pairs to the coocurrence (a radially decreasing weighting of the neighbors could be used). The statistical features of the coocurrence matrix that we measured follow [Muzzolini et al., 1998], and are angular second momentum (ASM), contrast (CON), entropy (ENT), cluster shade (CS), cluster prominence (CP), inertia (IN), and local homogeneity (LH). We removed correlation from the set in [Muzzolini et al., 1998] rather than working around division by zero valued standard deviations (for entropy, we assumed that zero multiplied by the log of zero is zero), and added the absolute value (ABS) from [Materka and Strzelecki, 1998]. The final set of spatial coocurrence texture parameters (defined for the G by G neighborhood surrounding the pixel) were as follows:

$$ASM = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j)^2 \tag{4.19}$$

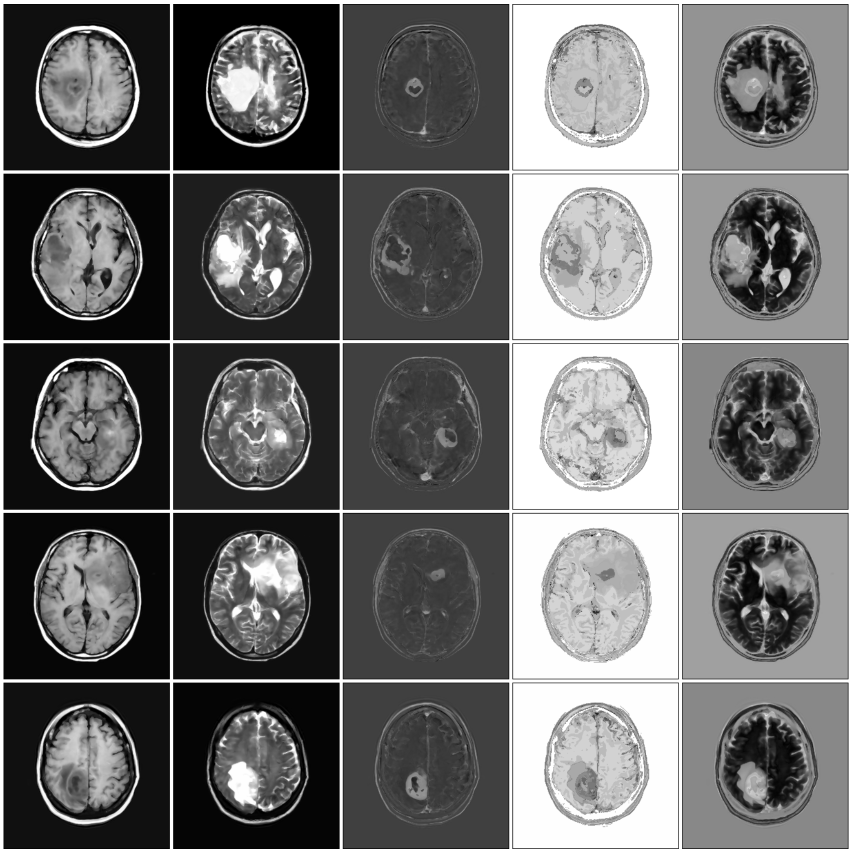


Figure 4.31: Examples of multi-spectral *histogram density* and *distance to normal intensity* features. First three columns: intensity standardized images for different patients (T1-weighted, T2-weighted, and contrast difference image). Fourth column: Multi-spectral histogram density. Fifth column: Multi-spectral Euclidean intensity distance to the multi-spectral mean intensity of gray matter in the template. In some cases, these can clearly represent excellent features that are based on the multi-spectral intensity distribution.

$$CON = \sum_{i=0}^{G-1} n^2 \sum_{|i-j|=n} P(i,j) \tag{4.20}$$

$$ABS = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} |i-j| P(i,j) \tag{4.21}$$

$$ENT = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} -P(i,j) \ln P(i,j) \tag{4.22}$$

$$CS = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i+j-\mu_x-\mu_y)^3 P(i,j) \tag{4.23}$$

$$CP = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i+j-\mu_x-\mu_y)^4 P(i,j) \tag{4.24}$$

$$IN = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i-j)^2 P(i,j) \tag{4.25}$$

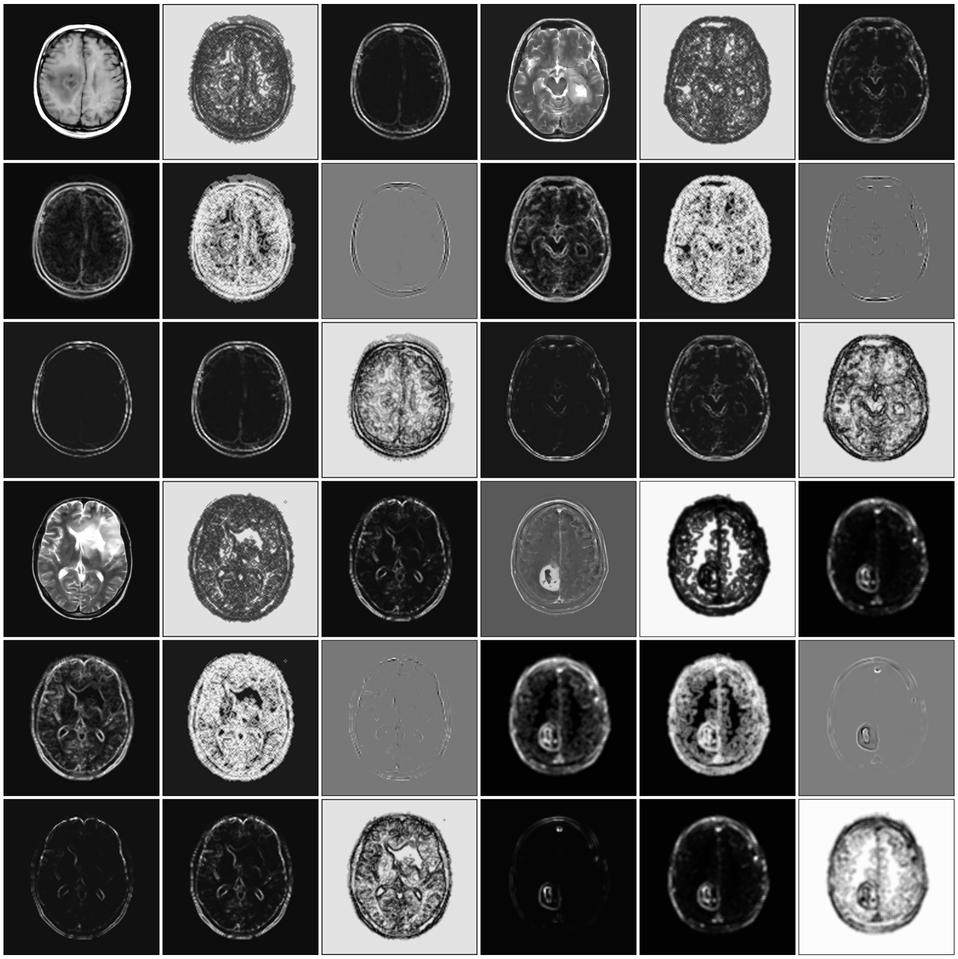


Figure 4.32: Examples of second-order (cooccurrence or ‘Haralick’) textures for 4 different images. Each quadrant contains the original image and the corresponding 8 texture images. The quadrants are organized as follows: Top, left to right: Original image, angular second momentum, contrast. Middle, left to right: absolute value, entropy, cluster shade. Bottom, left to right: Cluster prominence, inertia, local homogeneity.

$$LH = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{1}{1+(i-j)^2} P(i,j) \tag{4.26}$$

Examples of the images resulting from these feature computations are seen in Figure 4.32. We also explored first-order texture parameters (statistical moments). These parameters ignore spatial information and are essentially features that characterize properties of the local histogram. We calculated the parameters from [Materka and Strzelecki, 1998], which are mean, variance, skewness, kurtosis, energy, and entropy. Note that we converted the variance values to standard deviations. The first-order texture parameters used are demonstrated in Figure 4.33 and are defined as follows:

$$Mean: \ \ \mu = \sum_{i=0}^{G-1} iP(i) \tag{4.27}$$

$$Variance: \ \ \sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^2 P(i) \tag{4.28}$$

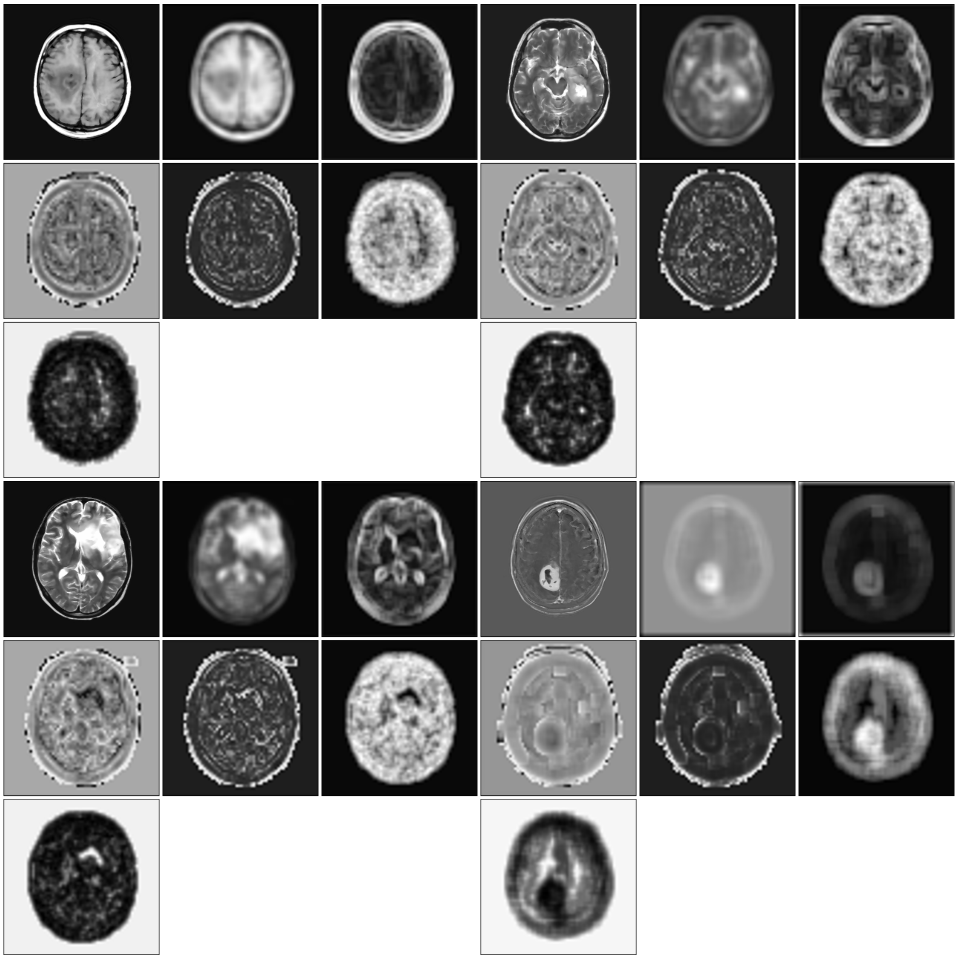


Figure 4.33: Examples of first-order (statistical moment) features for 4 different images. Each quadrant contains the original image and the corresponding 6 texture images. The quadrants are organized as follows: Top, left to right: Original image, mean, standard deviation. Middle, left to right: Skewness, kurtosis, entropy. Bottom left: Energy.

$$Skewness: \quad \mu_3 = \sigma^{-3} \sum_{i=0}^{G-1} (i-\mu)^3 P(i) \tag{4.29}$$

$$Kurtosis: \quad \mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i-\mu)^4 P(i) \tag{4.30}$$

$$Energy: \quad E = \sum_{i=0}^{G-1} P(i)^2 \tag{4.31}$$

$$Entropy: \quad H = -\sum_{i=0}^{G-1} P(i) \log P(i) \tag{4.32}$$

Within the Computer Vision literature, a currently popular technique for computing texture features is through linear filtering [Forsyth and Ponce, 2003], which represents a different approach than the Haralick features. The intuition behind using the responses of linear filters for texture parameters is that (balanced) filters respond most strongly to regions that appear similar to the filter [Forsyth and Ponce, 2003]. Thus, convolving an image with a variety of linear filters can assess how well each image region matches a set of filter 'templates', and the results can be used as a characterization of texture. There are considerable variation between methods based on this general concept. This includes the *Gaussian cube* multi-scale feature representation mentioned above, but also includes techniques based on Gabor filters, eigenfilters, Discrete Cosine Transforms, and Wavelet and other optimized methods. [Randen and Husoy, 1999] performed a comparative study of a large variety of texture features based on linear filtering, but added Haralick features and another type of 'classical' method of representing textures (autoregressive models). Although the study concluded that several of the linear filtering methods generally performed better than most others and that the classical methods generally performed poorly (as did several of the linear filtering approaches), it was also stated that the best methods depended heavily on the data set used and that the classical methods may be more appropriate in specific instances. Based on this conclusion (and several similar ones from related studies discussed in [Randen and Husoy, 1999], it is clear that evaluating a classical approach is still worthwhile, though we would also like to evaluate an approach based on linear filtering. An influential recent approach was proposed in [Leung and Malik, 2001], which used a set of filters consisting of Gaussians, Laplacian of Gaussians, and oriented Gaussian derivatives to form a *filter bank*, whose outputs used as features offered a relative invariance to changes in illumination and viewpoint. A recent comparative study, [Varma and Zesserman, 2002], evaluated four state of the art filter banks for the task of texture classification including the approach of [Leung and Malik, 2001]. This study found that the rotation invariant version of the Maximum Response filter bank (MR8) generally proved to be the best set of texture features for classification. This Maximum Response filter bank is derived from the Root Filter Set filter bank (Figure 4.34), consisting of a single Gaussian filter, a single Laplacian filter, and 36 Gabor filters (6 orientations each measured at 3 resolutions for both the symmetric and the antisymmetric kernel). A Gabor filter is a Gaussian filter that is multiplied element-wise by a one-dimensional cosine or sine wave to give the symmetric and antisymmetric filters, respectively (this filter has analogies to early visual processing in mammals [Forsyth and Ponce, 2003]).

$$G_{symmetric}(x,y) = \cos(k_x x + k_y y)e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.33}$$

$$G_{antisymmetric}(x,y) = \sin(k_x x + k_y y)e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.34}$$

The Maximum Response filter banks selectively choose which of the Gabor filters in the Root Filter Set should be used for each pixel based on the filter responses (the Gaussian and Laplacian are always included). The MR8 filter bank selects the asymmetric and symmetric filter at each resolution that generated the highest response. This makes the filter bank, which is already (relatively) invariant to illumination, also (relatively) invariant to rotation. Another appealing aspect of the MR8 filter bank is that it consists of only 8 features, giving a compact representation of regional texture. Since Gaussians and Laplacians were already being explored in this work, only the 6 additional (Gabor) features were required to take advantage of this method for texture characterization. We implemented the MR8 texture features (using the Root Filter Set code from the author's webpage), as an alternate (or possibly complementary) method of taking into account texture in the features. The features resulting from the MR8 filter bank are illustrated in Figure 4.35.

The fourth type of Image-based features discussed in section 3.4 was structure-based features. These features are based on performing an initial unsupervised segmentation to divide the image into homogeneous connected regions, and computing features based on the regions to which the pixels were assigned. These types of features are commonly referred to as shape or morphology based features, and include measures such as compactness, area, perimeter, circularity, moments, and many others. A description of many features of this type can be found in [Dickson and Thomas, 1997, Soltanian-Zadeh et al., 2004]. Beyond morphology based features, features can also be computed that describe the relationship of the pixel to or within its assigned region, such as the measure used in [Gering, 2003b] to assess whether pixels were in a structure that was too thick to be normal. Another set of features that could be computed after performing an initial unsupervised segmentation would

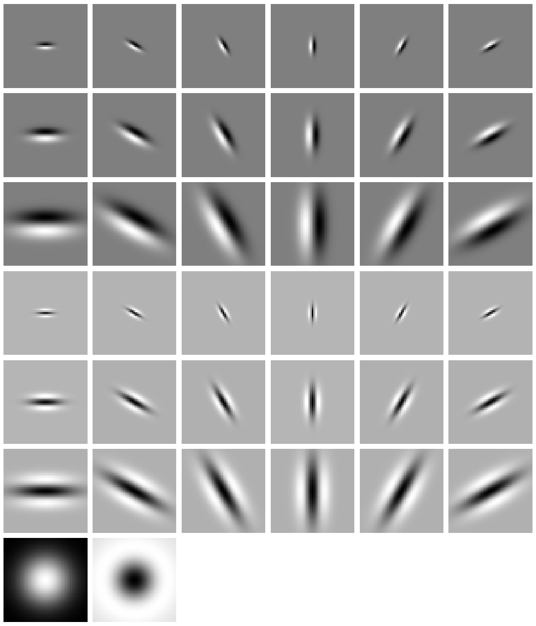


Figure 4.34: The elements of the *Root Filter Set* filter bank. First three rows: Antisymmetric Gabor filters of different scales and orientations. Next three rows: Symmetric Gabor filters of different scales and orientations. Bottom row: Gaussian filter (left) and Laplacian of Gaussian filter (right). The MR8 *Maximum Response* filter bank uses, for each pixel, the value of the response for the symmetric and asymmetric filter with the highest response at each scale (ie, the largest absolute value from each of the first six rows).

be to calculate texture features of the resulting region. The Haralick features or statistical moments would be more appropriate than linear filtering in this case, due to the presence of irregularly shaped regions. When evaluating structure based features, an unsupervised segmentation method should be used that can produce a hierarchy of segmented structures. Since the abnormal classes will not likely fall into a single cluster, evaluating structure based features at multiple degrees of granularity could give these features increased discriminatory ability. Structure-based features were not tested in this implementation, but represent an interesting direction of future exploration.

The features discussed in this section (multi-model pixel intensities, neighboring pixel intensities, Gaussian pyramids and cubes, Laplacian cubes, multi-spectral densities, multi-spectral distances to normal intensities, first order texture parameters, coocurrence based texture features, and the MR8 filter bank) were implemented in Matlab, and their performance will be examined in chapter 5. In addition to structure-based features, future directions to examine include performing multi-modality linear filtering or computing multi-modality textures. Given that the multi-spectral distances to normal intensities proved to be useful features, another direction of future research could be to incorporate the variance and covariance of the normal tissue intensities in the template intensity distribution into the ‘distance from normal intensity’ measure (possibly through the use of the Mahalanobis distance as suggested in [Gering, 2003b]). A final direction of future research with respect to image-based features is the evaluation of texture features based on generative models (such as those that use Markov Random Fields), that are currently regaining popularity for texture classification, and have shown to outperform the MR8 filter bank by one group [Varma and Zisserman, 2003].

4.4.2 Coordinate-Based Features

The simplest form of coordinate-based feature is obviously spatial coordinates. However, these features were not examined, as they are only applicable to inter-patient scenarios if the tumor is highly localized, or a sufficiently large training set is available. Even though many of our experiments

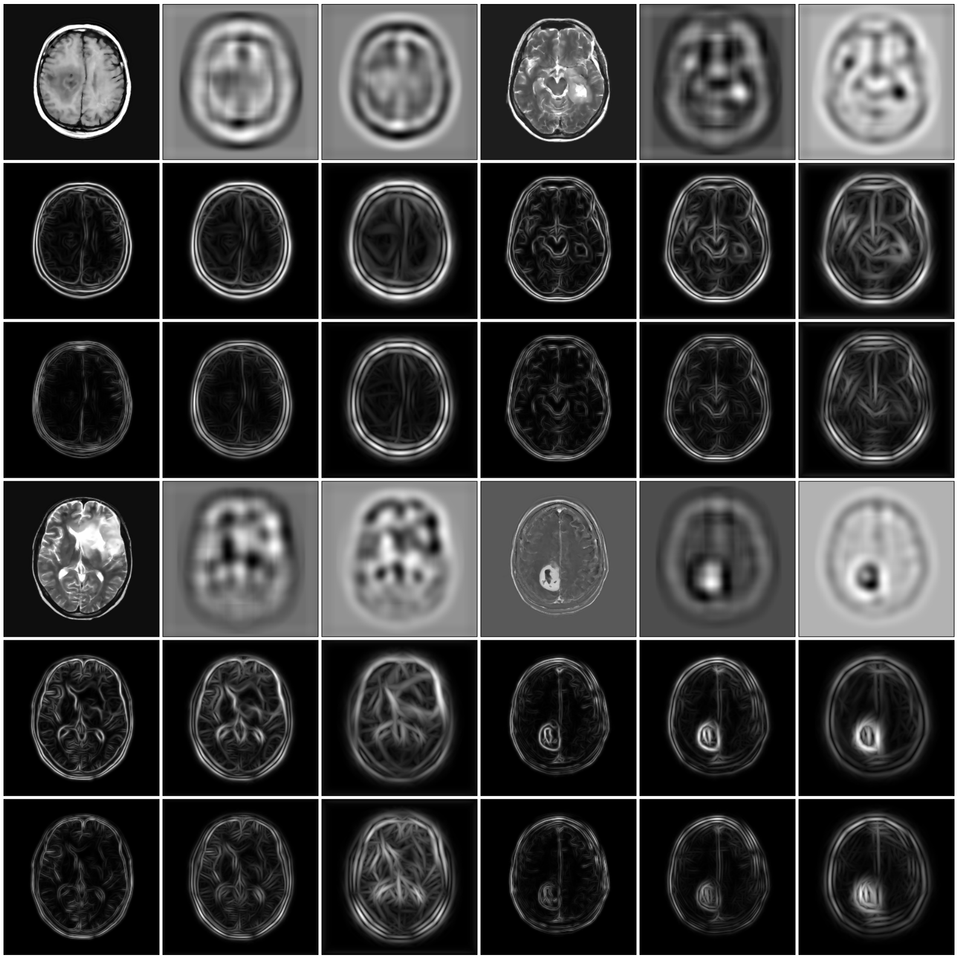


Figure 4.35: Examples of the MR8 texture features for 4 different images. Each quadrant contains the original image and the corresponding 8 texture images. The quadrants are organized as follows: Top, left to right: Original image, (balanced) Gaussian respone, Laplacian of Gaussian response. Middle, symmetric filter maximum responses at different scales. Bottom, asymmetric filter maximum responses at different scales.

in Chapter 5 may have benefited from the use of coordinates, it was decided not to evaluate these features since in general they will not be used.

The coordinate-based features that have been used in other systems are the spatial prior probabilities for gray matter, white matter, and CSF. The probabilities most commonly used are those included with the SPM package [SPM, Online]. The most recent version of this software includes templates that are derived from the ‘ICBM152’ data set [Mazziotta et al., 2001] from the Montreal Neurological Institute, a data set of 152 normal brain images that have been linearly aligned and where gray matter, white matter, and csf regions have been defined. The SPM versions of these priors mask out non-brain areas, reduce the resolution from $1mm^3$ isotropic pixels to $2mm^3$ isotropic pixels, and smooth the results with a Gaussian filter. Rather than use the SPM versions, we chose to use the ‘tissue probability models’ from the ICBM152 data set obtained from the ICMB View software [ICBM View, Online]. These were chosen since the system has a separate prior probability for the brain (removing the need for masking), since these have a higher resolution ($1mm$ by $1mm$ by $2mm$ pixels), and since these probabilities can be measured at multiple resolutions, allowing the use of both the original highly detailed versions and smoothed versions. A comparison of the SPM priors and the ICBM152 priors is shown is Figure 4.36. For a brain mask prior probability, the prior included with SPM2 was used, which is derived from the MNI305 average brain

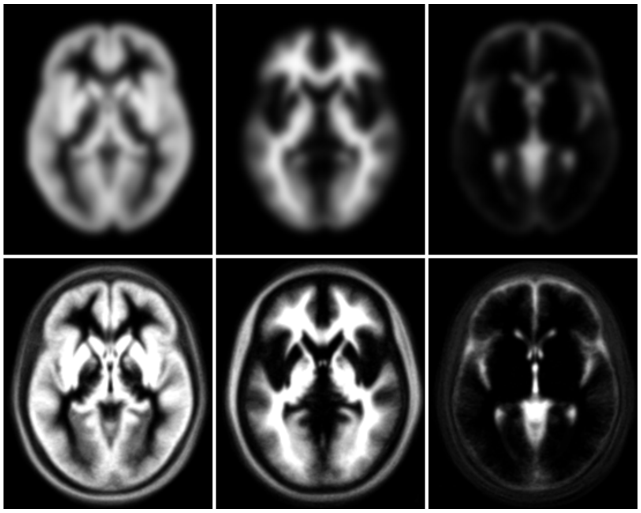


Figure 4.36: Comparison of the SPM [SPM, Online] tissue priors (top) and those from the ICBM152 data set (bottom) [ICBM View, Online]. Left to right: gray matter, white matter, and CSF. The ICBM152 priors have a higher resolution, but do not have non-brain areas masked.

[Evans and Collins, 1993], and is re-sampled to $2mm^3$ isotropic pixels and smoothed as with the other SPM prior probabilities. Figure 4.37 compares registered image to the 4 priors at the corresponding slices in the template coordinate system. Figure 4.38 visualizes the spatial alignment between the image and these priors.

For a simple measure of anatomic variability, the average images constructed from the ICBM152 data set (also obtained from the ICBM View tool) were used. These consist of average T1-weighted and T2-weighted images from the 152 linearly aligned patient. This represents a measure of the average intensity expected at each pixel in the coordinate system, and is an important feature since a large divergence from average may indicate abnormality. This average measure is only a crude measure of variability, and future implementations could examine more sophisticated probabilistic brain atlases, such as those discussed in [Toga et al., 2003].

In addition to more sophisticated measures of anatomic variability, future implementations could examine the effectiveness of additional prior probabilities. This could include spatial prior probabilities for tissues such as bone, connective tissues, glial matter, fat, nuclear gray matter, muscle, or skin.

4.4.3 Registration-Based Features

If the registration stage performed perfect registration, a regional similarity metric between the image and template could be used to find areas of abnormality. However, as with intensity standardization, the registration stage is not expected to be perfect and thus a regional similarity measure will not be a sufficient feature for abnormality segmentation. However, registration-based features have only begun to be explored to enhance segmentation, and they represent potentially very useful features to include alongside other features to enhance segmentation.

The only system to date that has used a registration-based feature for brain tumor segmentation was that in [Kaus et al., 2001] (we consider the use of spatial prior probabilities to be a coordinate-based feature). This system used a distance transform that computed the distance to labeled regions in the template. We did not examine distance transforms, since spatial prior probabilities measured at different resolutions can represent similar information in a more elegant way. Under the same logic, we also did not examine other features that are based on labeled regions in a template.

Rather than using features based on template labels, we chose to explore features that used the template image data directly, which encodes significantly more information (and does not require

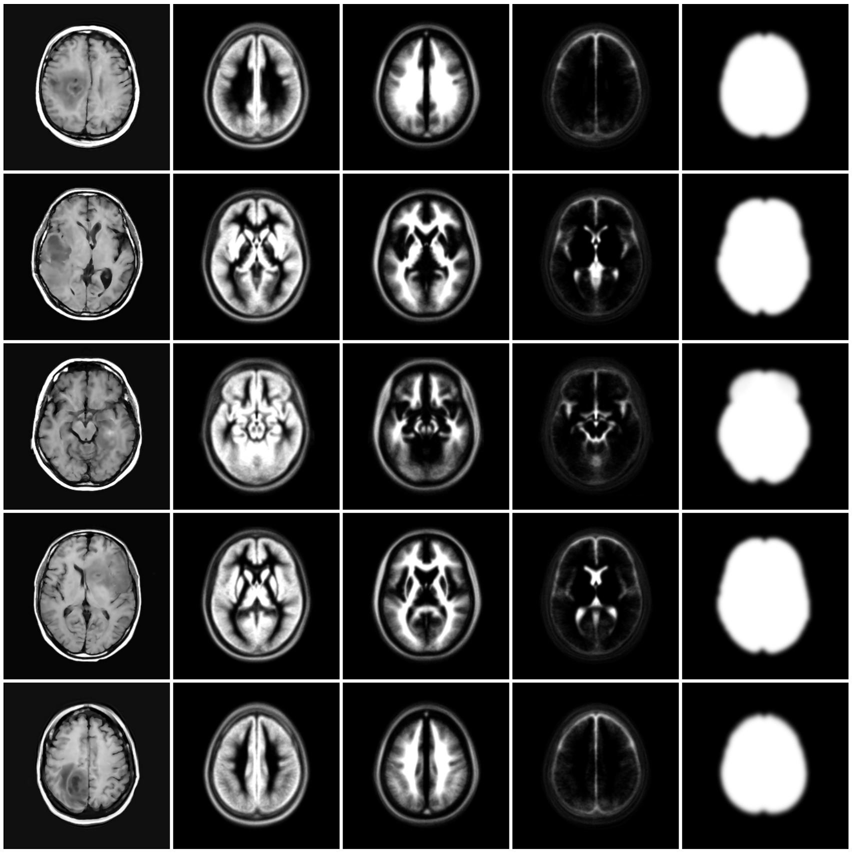


Figure 4.37: Spatial prior probabilities as coordinate-based features for slices from different areas of the brain and different patients. Left to right: Original image, gray matter prior, white matter prior, CSF prior, brain area prior.

manual labeling of structures). The simplest way to incorporate template image data as a feature is to include the intensity of the pixel at the corresponding location in the template. This feature has an intuitive use, since normal regions should have similar intensities to the template while dissimilarity could be an indication of abnormality. This may seem to represent a misleading feature: while the heavily regularized non-linear registration corrects for overall differences in head shape, the registration algorithm might not find an exact pixel-level correspondence. However, the use of a multi-scale feature representation allows for an imperfect correspondence, since regions of slight misalignment and even regions of higher anatomic variability between individuals will often be similar at coarser scales. This feature is illustrated in Figure 4.40, while it is shown at a coarser scale in Figure 4.41. Although we only explored this direct measurement of intensities (at multiple scales), texture features could have been used as an alternative or in addition to intensities. Measuring local difference, correlation, or information measures such as entropy were considered to utilize the template image data, but were not explored in this work.

It has been proposed that bi-lateral symmetry could represent an excellent patient-specific template of a normal brain for use in tumor detection [Gering, 2003b]. However, the use of symmetry is complicated by (i) the problem of locating the line of symmetry (especially in brains with large tumors), (ii) the fact that the normal hemisphere of brains with pathology will also be asymmetric (in addition to other normal areas), and (iii) the ability of tumors and edema to cross the line of symmetry. We have defined symmetry as a *registration-based* feature, since we use the (non-linearly registered) template's known line of symmetry to approximate the line of symmetry in the patient to be labeled. Using this line, we consequently assess local symmetry as a pixel feature by subtracting the intensity value of the contra-lateral pixel from the pixel's own intensity value.

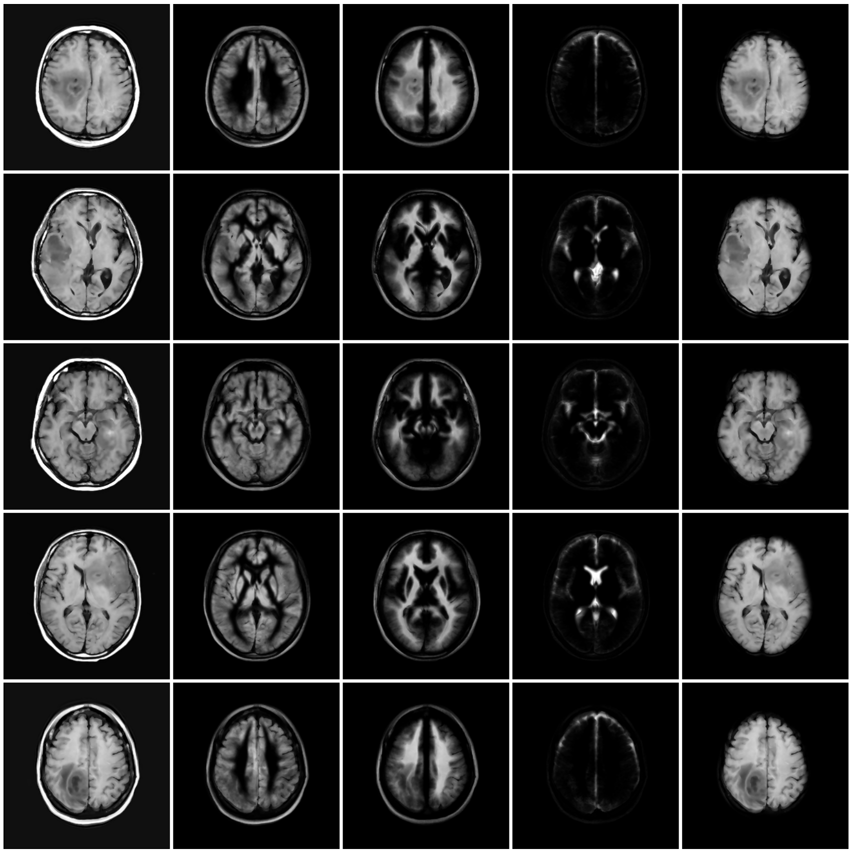


Figure 4.38: The images from the previous Figure multiplied element-wise by the four priors. Left to right: Original image, gray matter prior, white matter prior, CSF prior (the T2-weighted image was used in the calculation to enhance visualization), brain area prior. These images indicate that the spatial priors will likely be able to enhance classification, since they are relatively reliable measures of the expected locations of normal structures (especially relevant is the normal brain location and the locations of normal CSF)

Once again, a multi-scale feature representation is essential for symmetry features, and allows the coarse scale features to make classification more robust to minor asymmetries. This strategy for taking advantage of symmetry does not directly address asymmetry in the normal hemisphere nor symmetric tumor regions, however the other features used can be combined by the classifier to help disambiguate such cases. As with utilizing the template's image information, texture features or other more sophisticated measures could have been computed to assess symmetry.

We did not examine morphometry or other features that take advantage of how the image was warped during registration, and this is an interesting future direction to explore. Another interesting future direction could be the incorporation of registration-based features from multiple templates, since the registration-based features we explored used only a single template. Additionally, the use of a template that is bi-laterally symmetric might improve the discriminatory power of the symmetry features.

4.4.4 Feature-Based Features

Our exploration of feature-based features was limited. We examined multi-scale versions of image-, coordinate-, and registration-based features. However, we did not examine including neighborhood feature values or computing texture values based on features other than the intensities. We did not explore automatic feature selection nor dimensionality reduction, which represent future directions

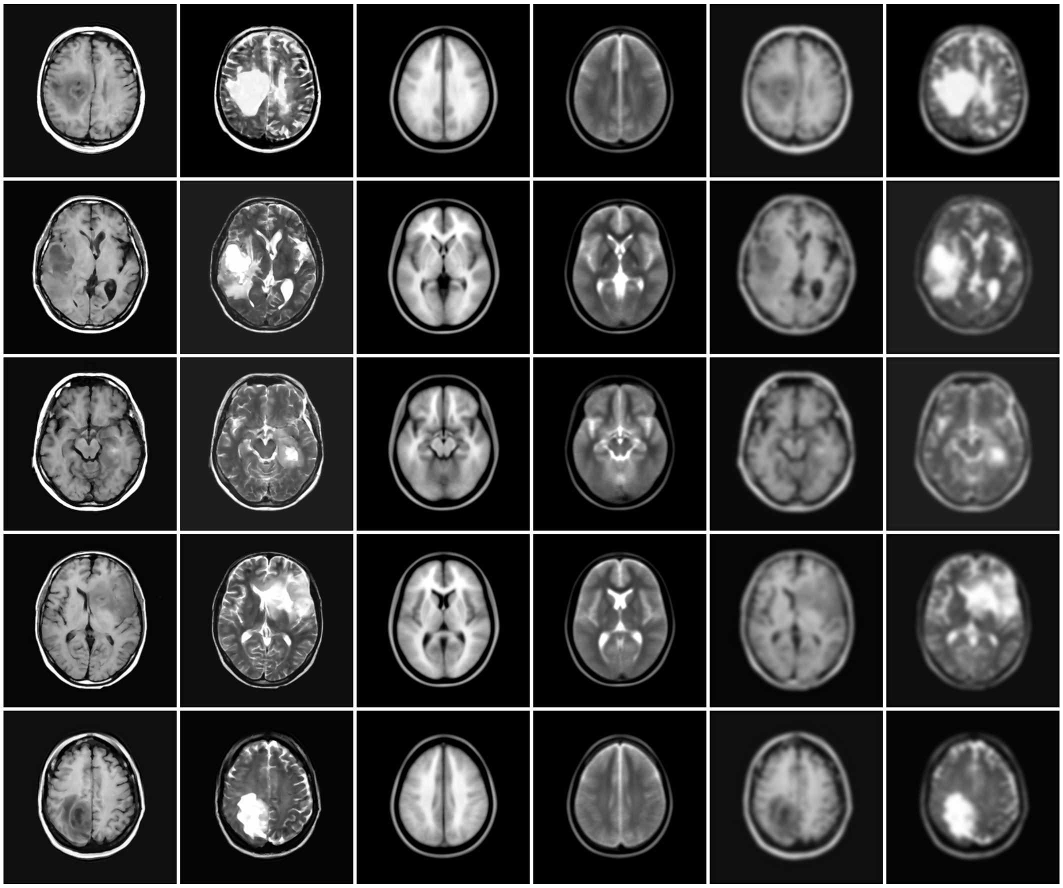


Figure 4.39: Average intensities from a population to characterize expected spatial intensity values. Left to right: T1-weighed image, T2-weighted image, average T1-weighted image from a population, average T2-weighted image from a population, smoothed T1-weighted image, smoothed T2-weighted image. The average intensities represent expected spatial intensities, and therefore could be potentially relevant features. Notice that smoothed versions of the images (as in a Gaussian Cube feature representation) structurally resemble the average images (in normal regions).

of research that could improve results. Also, we did not examine sequences of feature extraction operations, that could improve results but whose meanings are not necessarily intuitive and would require automated feature extraction. This would include, for example, computing the Haralick features of a low resolution version of the filter bank results (or simply recursively computing the filter outputs).

4.4.5 Summary

It is important to note that it is often the combination of different types of features that allows a more effective classification. For example, knowing that a pixel is asymmetric on its own is relatively useless. Even with the additional knowledge that the pixel has a high T2 signal and a low T1 signal would not allow differentiation between CSF and edema. However, consider the use of the additional information that the pixel's region has a high T2 signal and low T1 signal, that CSF is unlikely to be observed at the pixel' spatial location, that a high T2 signal is unlikely to be observed at the pixel's location, that the pixel has a significantly different intensity than the corresponding location in the template, that the pixel's intensities are more distant in the standardized multi-spectral intensity space than normal CSF, and finally that the texture of the image region is not characteristic of CSF. It is clear from this additional information that the pixel is likely edema rather than CSF. It is also

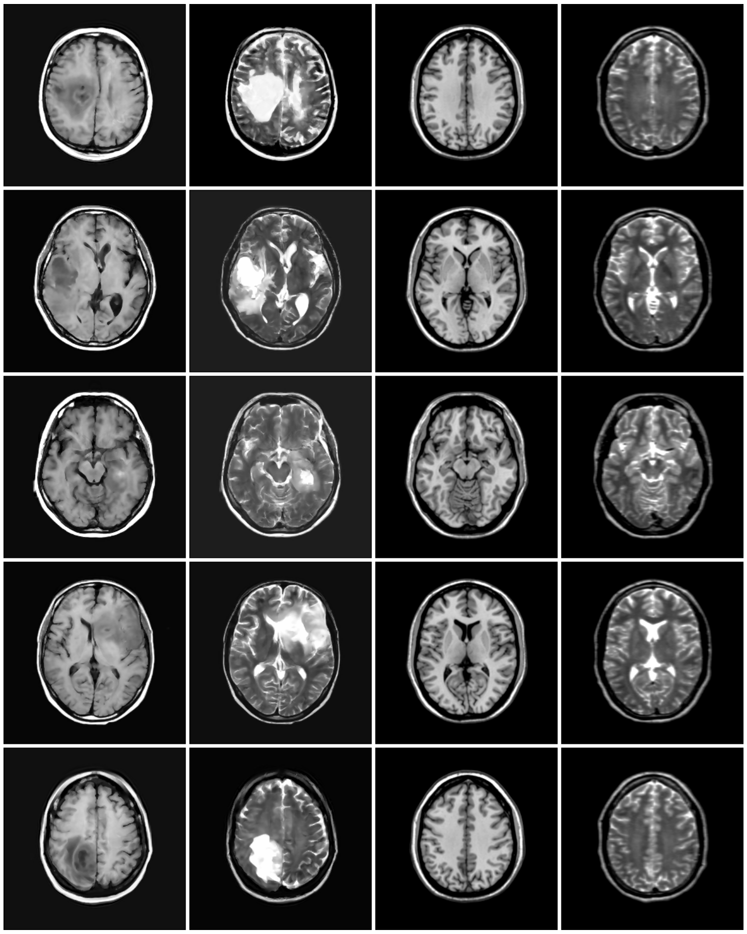


Figure 4.40: Employing the template intensities directly as features. Left to right: T1-weighted image, T2-weighted image, T1-weighted template, T2-weighted template. These features perform a similar function to the average intensities. However, since the images have been non-linearly registered and intensity standardized with the template, finer comparisons can potentially be used.

clear that the use of this additional information will add robustness to classification, since each of the features can be simultaneously considered and combined in classifying a pixel as normal or tumor. It is hoped that reading this section has given the impression that simply using the intensities as features or converting a neighborhood of intensities into a feature vector does not fully take advantage of even the image data, much less the additional information that can be gained through registration.

Not all of the features implemented were included in the final system. Based on our experiments, many of which will be outlined in Chapter 5, the final system included the multi-spectral intensities, the spatial tissue prior probabilities, the multi-spectral spatial intensity priors, the multi-spectral template intensities, the distances to normal tissue intensities, and our characterization of bi-lateral symmetry, all measured at multiple scales using a Gaussian Cube representation. In addition, the final system measured several Laplacian of Gaussian filter outputs and the Gabor responses from the MR8 filter bank for the multi-spectral intensities (although ideally these would be measured for all features and an automated feature selection algorithm would be used to determine the most relevant features).

Although examples of various different types of features were explored in this work, it should be emphasized that there remains a considerable amount of exploration that can be made with respect to feature extraction. More sophisticated coordinate-based and registration-based features in particular

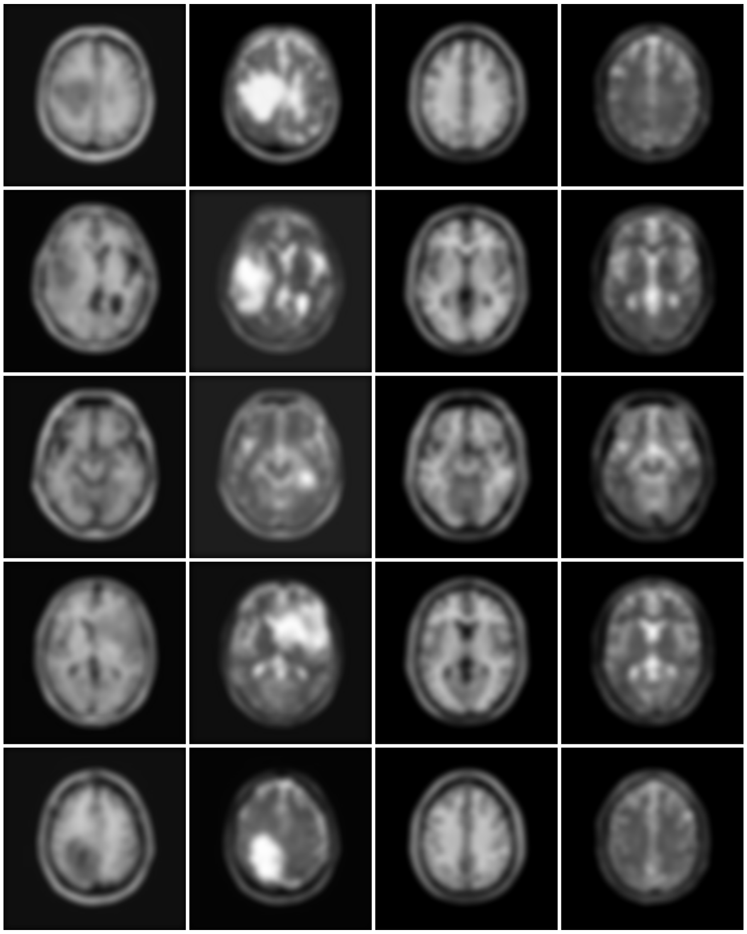


Figure 4.41: Coarse-scale comparison of images with the corresponding template slices. This Figure is a smoothed version of the previous Figure, as in a Gaussian Cube feature representation. Notice that normal regions that did not align exactly in the previous image can still be very similar at coarse scale.

represent areas with significant future potential, as this was the first work we are aware of that explores the use of more than one type of feature from these classes. Automated feature selection or dimensionality reduction also represent areas of exploration that could improve results, as these operations were not explored in this work, and it is clear that there is considerable redundancy and correlation in the features. The computation of each of the features discussed in this section was implemented in Matlab [MATLAB, Online].

4.5 Classification

Supervised classification of data from a set of measured features is a classical problem in Machine Learning and Pattern Recognition. Given the features extracted in the previous section, the task in classification is to use the features measured at a pixel to decide whether the pixel represents a tumor pixel or a normal pixel. Manually labeled pixels will be used to learn a model relating the values of the features to the labels, and this model will then be used to classify pixels for which the label is not given (from the same or a different patient). As discussed in Chapter 2, there have been a variety of different classification methods proposed to perform the task of brain tumor segmentation using image-based features (although most of the previous work has assumed patient-specific training). ANN models have been used by several groups [Dickson and Thomas, 1997,

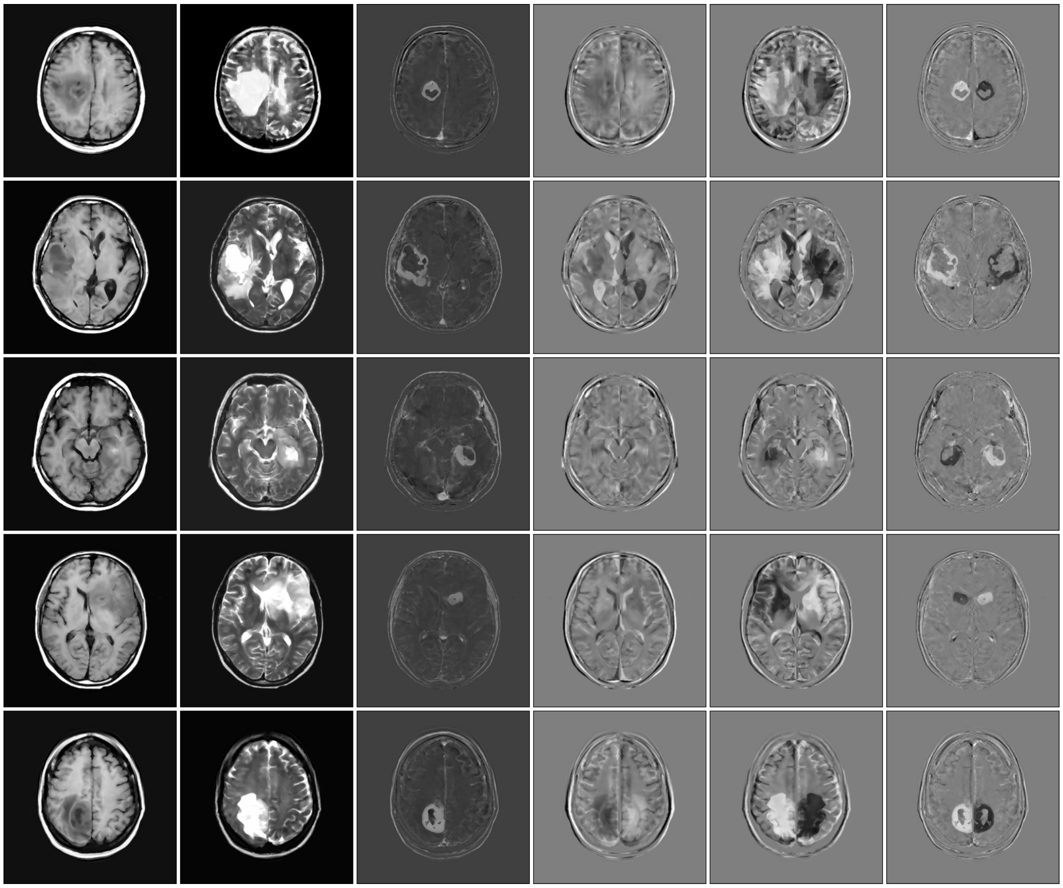


Figure 4.42: Assessing bi-lateral symmetry using the template's known line of symmetry. Left to right: T1-weighted image, T2-weighted image, contrast agent difference image, T1-weighted image symmetry, T2-weighted image symmetry, contrast agent difference image symmetry. It is clear that abnormal regions tend to be more asymmetric than normal regions.

Alirezaie et al., 1997, Ozkan et al., 1993], and are appealing since they allow the modeling of non-linear dependencies between the features, and minimize an objective measure of classification performance. However, training ANN models is problematic due to the large amount of parameters to be tuned and the abundance of local optimums. Classification with Support Vector Machines (SVM) has recently been explored by two groups for the task of brain tumor segmentation [Zhang et al., 2004, Garcia and Moreno, 2004], and represent a more appealing approach than ANN models for the task of binary classification since they have more robust (theoretical and empirical) generalization properties, achieve a globally optimal solution, and also allow the modeling of non-linear dependencies in the features [Shawe-Taylor and Cristianini, 2004].

In the task of binary classification, if the classes are linearly separable in the feature space (and assuming approximately equal covariances and numbers of training instances from both classes), then the logic behind Support Vector Machine classification is that the best linear discriminant for classifying new data will the be the one that is furthest from both classes. Binary classification with (2-class hard) Support Vector Machines is based on this idea of finding the linear discriminant (or hyperplane) that maximizes the *margin* (or distance) to both classes in the feature space. The use of this margin maximizing linear discriminant in the feature space provides statistical bounds on how well the learned model will perform on pixels outside the training set [Shawe-Taylor and Cristianini, 2004]. The task of finding the margin maximizing hyperplane can be formulated (in its dual form) as the maximization of the following expression [Russell and Norvig, 2002]:

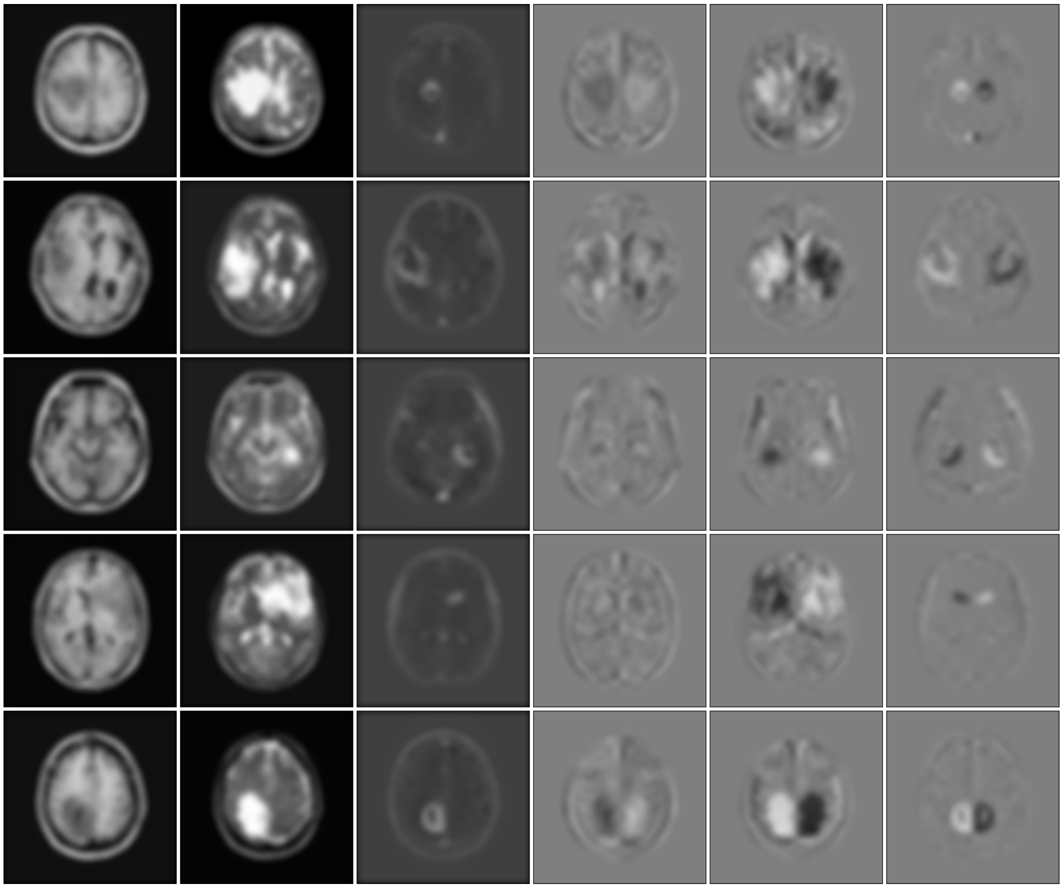


Figure 4.43: Assessing bi-lateral symmetry using the template's known line of symmetry at a coarser scale. This Figure is a smoothed version of the previous Figure, as in a Gaussian Cube feature representation. At a coarser scale, minor asymmetries (ie. near the skull) are less prominent, while large asymmetric regions (as in tumor/edema) can clearly be seen.

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{4.35}$$

Subject to the constraints that $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$, where x_i is a vector of the features extracted for the ith training pixel, y_i is 1 if the ith training pixel is tumor and -1 otherwise, and α_i are the parameters to be determined. This formulation under the above constraints can be formulated as a *Quadratic Programming* optimization problem whose solution is guaranteed to be optimal and can be found efficiently. A new pixel with features x for which the label is not known can be classified using the following expression [Russell and Norvig, 2002]:

$$h(x) = sign(\sum_i \alpha_i y_i (x \cdot x_i)) \tag{4.36}$$

In the optimal solution, most of the α_i values will be zero, except those close to the hyperplane. The training vectors with non-zero values are referred to as Support Vectors. If the classification rule above is examined, it can be seen that only these Support Vectors are relevant in making the classification decision, and that other training points can be discarded since their values can be easily predicted given the Support Vectors (and associated weight values). This sparse representation allows efficient classification of new pixels, and leads to efficient methods of training for large training and features sets. A simple two-dimensional example of a linearly separable dataset, the margin maximizing linear discriminant, and the corresponding Support Vectors is seen in Figure 4.44.

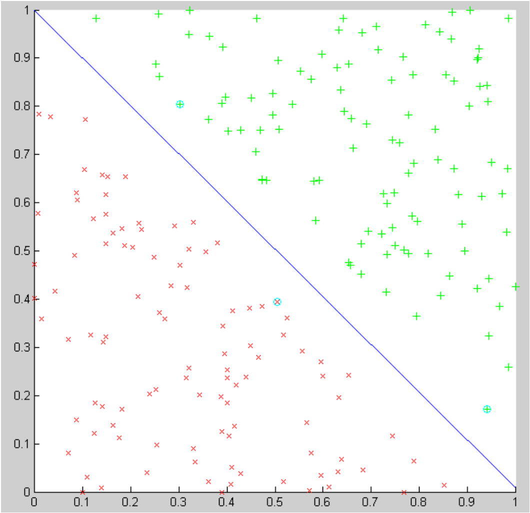


Figure 4.44: Support Vector Machine classification of a two-dimensional linearly separable data set. This data consists of two classes (the X's and the $+$'s) that can clearly be discriminated by a linear function. The line between the classes is the (margin maximizing) linear discriminant found by a hard-margin Support Vector Machine. The three circled data points are the *Support Vectors*. The Support Vectors and associated weights are used to define the line, and these values are sufficient to classify any point in the feature space.

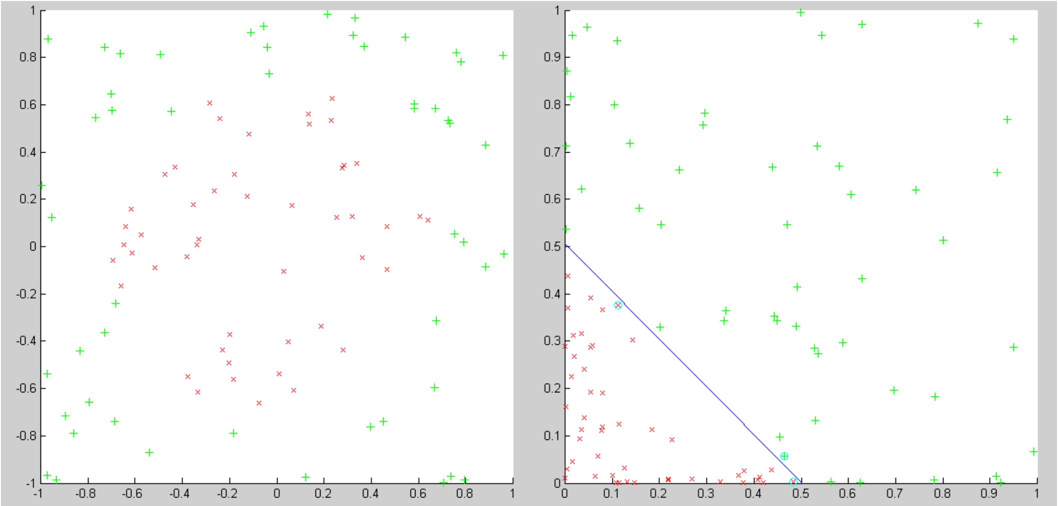


Figure 4.45: An example of a data set that is not linearly separable in the feature space, but can be made linearly separable with the Polynomial kernel. Left: Two-dimensional data set in the original feature space. It is not possible to separate the two classes using a linear function. Right: A two-dimensional projection of the feature space defined by the Polynomial kernel. In this feature space, the classes are now separable by a linear function.

The Support Vector Classification formulation above learns only a linear classifier, while previous work on brain tumor segmentation indicates that a linear classifier may not be sufficient. However, the fact that the training data is represented solely as an inner (or 'dot') product allows the use of the *kernel trick*. The kernel trick can be applied to a diverse variety of algorithms (see [Shawe-Taylor and Cristianini, 2004]), and consists of replacing the inner product with a different measure of similarity (kernel) between feature vectors:

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{4.37}$$

The idea motivating this transformation is that the data can be *implicitly* evaluated in a different feature space, where the classes may be linearly separable. This is similar to including combinations of features as additional features, but removes the need to actually compute or store these combinations (of which there can be an exponential, or infinite number represented through the kernel). The kernel function used needs to have very specific properties and thus arbitrary similarity metrics can not be used, but research into kernel functions has revealed many different types of admissible kernels, that can be combined to form new kernels [Shawe-Taylor and Cristianini, 2004]. Although the classifier still learns a linear discriminant, the linear discriminant is in a different feature space, and will form a non-linear discriminant in the original feature space.

The two most popular non-linear kernels are the Polynomial and the Gaussian kernel (sometimes referred to as the Radial Basis Function kernel). The Polynomial kernel simply raises the inner product to the power of a scalar value d (other formulations add a scalar value R to the inner product before raising to the power of d):

$$K(x_i, x_j) = (x_i \cdot x_j)^d \tag{4.38}$$

This kernel corresponds to embedding the data points in a feature space that includes all monomials (multiplications between features) up to degree d. Since there are an exponential amount of these monomials, without the use of kernels it would not be feasible to use these additional features for higher values of d, nor for large feature sets. Figure 4.45 illustrates an example of a data set that is not linearly separable in the original feature space, but is linearly separable in the feature space defined by the Polynomial kernel. The Gaussian kernel replaces the inner product with a Gaussian distance measure between the feature vectors. This kernel space is thus defined by (exponentially decaying) distances to the training pixels in the feature space (not to be confused with image distances). More complicated feature spaces can allow more effective discrimination of the training data, but at the cost of increased model complexity. More Support Vectors are needed to define a hyperplane in complicated feature spaces, and increasingly complicated feature spaces will eventually overfit the training data without providing better generalization on unseen test data. For example, when using the Polynomial kernel, higher values of d will lead to feature spaces where the classes are increasingly separable in the training data, but the linear separators found will be more heavily biased by the exact values of the training pixel features and will not necessarily accurately classify new pixels. In selecting a kernel, it is thus important to select a kernel that allows separation in the feature space, but to note that increased separability of the training data in the feature space does not guarantee higher classification accuracy of the learned model on test data.

It is possible that a linear discriminant in a given kernel feature space embedding will accurately classify most of the pixels in the training data, but noise and outliers may mean that the classes are not linearly separable in the feature space. There are natural methods of regularization for Support Vector Machines that can overcome cases of non-separability, one of the most popular of which is the use of slack variables, variables added to the Quadratic Programming formulation that allow a specified amount of margin violation [Shawe-Taylor and Cristianini, 2004]. An example of a linear discriminant found for a non-linearly separable data set using slack variables is seen in Figure 4.46. An equivalent but slightly more intuitive method of regularization is the ν-SVM formulation, which regularizes the number of Support Vectors in the learned model [Shawe-Taylor and Cristianini, 2004].

Although it has been stated that the Quadratic Programming formulation can be solved efficiently (for its size), the formulation can still involve solving an extremely large problem, especially in the case of image data where a single labeled image can contribute tens of thousands of training points. Fortunately, optimization methods such as Sequential Minimal Optimization [Platt, 1999], the SVM-Light method [Joachims, 1999], and many others exist that can efficiently solve these large problems.

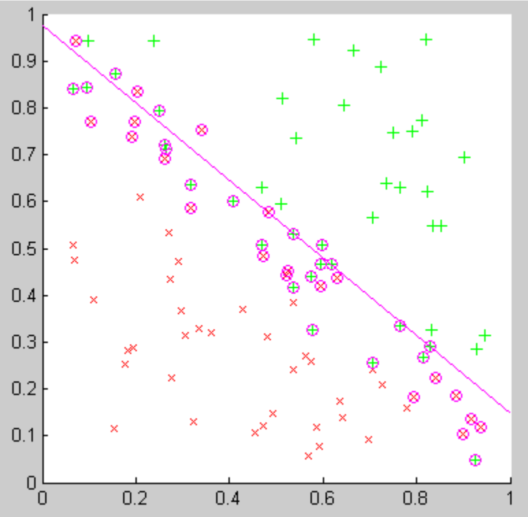


Figure 4.46: Support Vector Machine classification of an approximately linearly separable data set with slack variables. Although the two classes cannot be discriminated with a linear function, the use of slack variables allows a suitable linear discriminant to be found. The circled points are the Support Vectors in this 'soft-margin' Support Vector Machine.

In this implementation, the Linear and Polynomial kernels, slack variables for regularization, and the SVM-Light optimization method were used. A degree of 2 was used in the Polynomial kernel, which performed slightly better on average than higher degree kernels. The Gaussian kernel outperformed these kernels in some experiments, but it proved to be sensitive to the selection of the kernel parameters and performed much worse than the linear and polynomial kernels in other cases. In our experiments, each of the features was scaled to be in the range [-1,1], to decrease the computational cost of the optimization and to increase separability in the case of the Polynomial kernel. Sub-sampling of the training data was also implemented to reduce computational costs. The sub-sampling method used allowed different sub-sampling rates depending on properties of the pixel. The three different cases used for this purpose were: tumor pixels, normal pixels that had non-zero probability of being part of the brain, and normal pixels that had zero probability of being part of the brain. It was found that the latter case could be sub-sampled heavily or even eliminated with minimal degradation in classifier accuracy.

Examples of the classification results (on training data) obtained by a Support Vector Machine classifier on a synthetic tumor are shown in Figure 4.47. The corresponding features computed for the synthetic tumor images are shown in Figure 4.48. Figures 4.49, 4.50, and 4.51 demonstrate (test data) results on real data for three different tasks. The full set of features extracted for one of the real images is seen in Figure 4.52.

There remains several directions of future exploration with respect to the classification step. Firstly, we examined only 2 non-linear kernels, and both were general purpose kernels. Specific kernels for image and graph data exist, and some kernels such as the Hausdorff kernel have been applied to related tasks [Barla et al., 2002]. With respect to the classifier used, our experiments indicated that model averaging is a promising approach, and could be examined further. We ultimately selected to use SVMs rather than a model averaging approach due to the time required for both training and testing and the model, since the SVMs were significantly faster than the model averaging methods. Future implementations could also examine techniques such as the *Bayes Point Machine*, that have better generalization properties than Support Vector Machines [Herbrich et al., 2001]. Finally, this work did not examine classification with more than 2 classes. Future implementations could examine this case, where Support Vector Machines may not be the best choice.

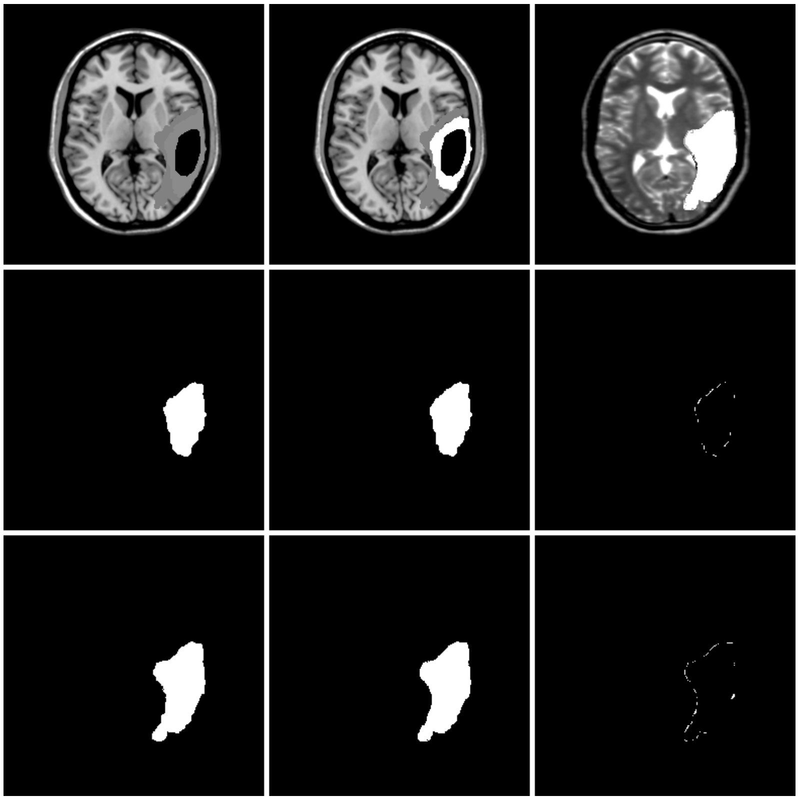


Figure 4.47: The results of a Support Vector Machine pixel classifier for segmenting a synthetic tumor/edema region. Top row: Image data with synthetic enhancing tumor, necrotic, and edema areas. Middle row: Manual label for enhancing tumor area (that is not error free), enhancing tumor pixels predicted by the SVM, difference between the manual and predicted pixel labels. Bottom row: Manual label for edema area (that is not error free), edema pixels predicted by the SVM, difference between the manual and predicted pixel labels. A soft-margin SVM was used with the linear kernel. These images represent re-substitution results, and thus the classifier has seen the manual label for each pixel. The extracted feature set is seen in Figure 4.48.

4.6 Relaxation

Unfortunately, the classifier will not correctly predict the labels for all pixels in new unseen test data. However, the classifier evaluated the label of each pixel individually, and did not explicitly consider the dependencies between the labels of neighboring pixels. The goal of the relaxation phase is to correct potential mistakes made by the classifier by considering the labels of spatial neighborhoods of pixels, since neighboring pixels are likely to receive the same value.

Morphological operations such as dilation and erosion are a simple method to do this. We utilized a related technique, which was to apply a median filter to the image constructed from the classifier output. This median filter is repeatedly applied to the discrete pixel labels until no pixels change label between applications of the filter. The effect of this operation is that pixel's labels are made consistent with their neighbors, and boundaries between the two classes are smoothed. After this operation, 'holes' of pixels labeld as normal that are completely surrounded by tumor pixels are filled with a morphological operation. Figure 4.53 demonstrates the effects of the implemented relaxation operations.

Repeated application of a median filter followed by morphological hole filling can only be considered a crude method of relaxation, but it consistently improved or did not change the accuracy of the system in our experiments. There is a diverse variety of directions to be explored for relaxation in future implementations. One simple improvement could be the selection of the largest connected

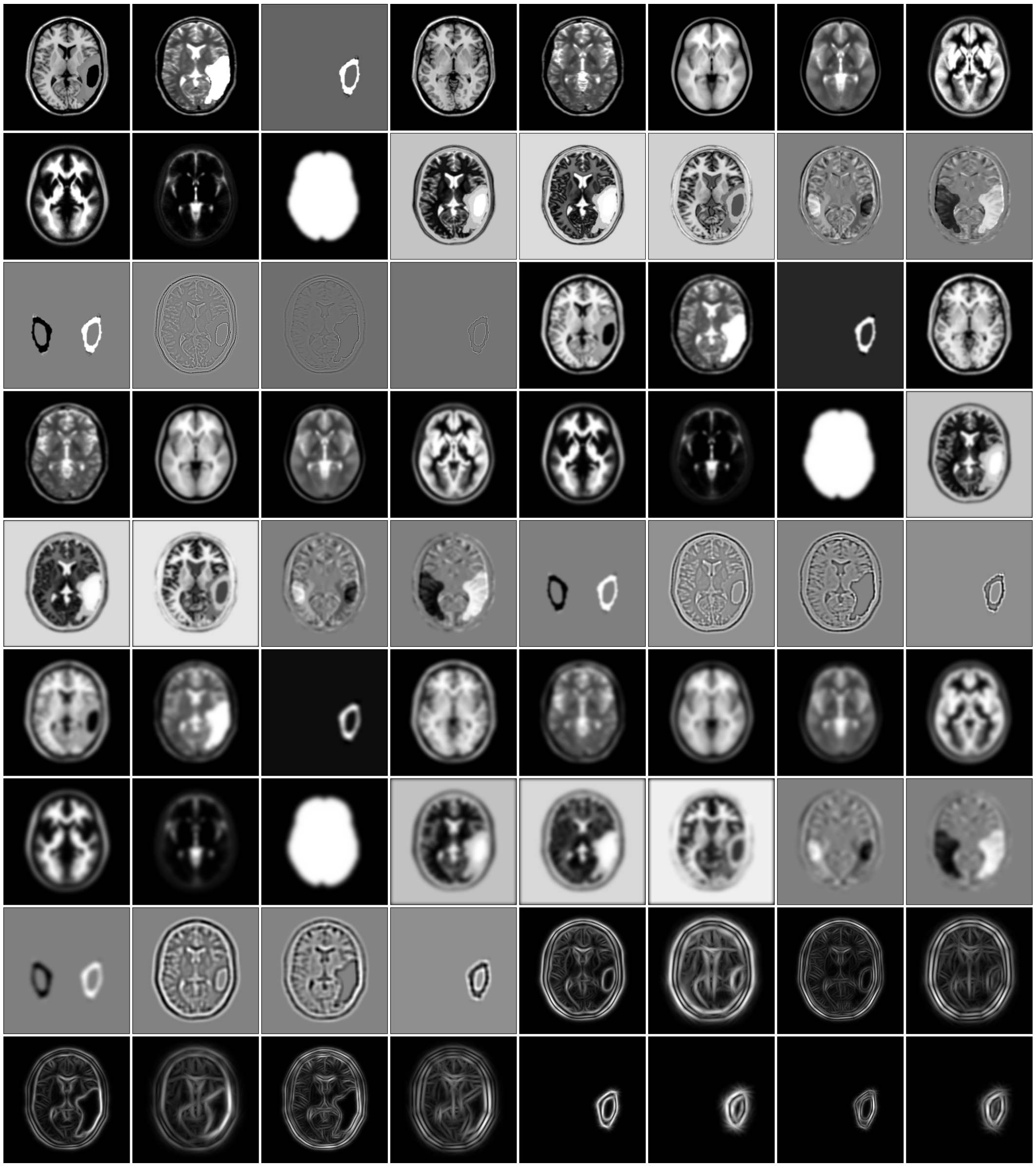


Figure 4.48: The full feature set of 75 features computed for the synthetic tumor data in Figure 4.47 (25 features at 3 scales).

component as in [Mazzara et al., 2004] (we specifically avoided this operation since our data contained multi-focal tumors). Other cluster selection methods could also potentially improve results, since occasionally a small area of misclassified pixels was large enough to not be removed through relaxation.

Another promising direction to explore for label relaxation are methods that relax ‘soft’ probabilistic labels as opposed to discrete binary labels. These methods obviously depend on having a classifier that outputs probabilistic labels. A simple way to generate probabilistic labels, in the case of Support Vector Machines, is to fit a logistic model to the classifier’s decision function [Platt, 2000].

Given probabilistic labels, several relaxation methods could be explored. The simplest relaxation method would be to smooth the probabilistic labels with a low-pass linear filter before assigning discrete labels. A more sophisticated method could be to use the probabilities to initialize a

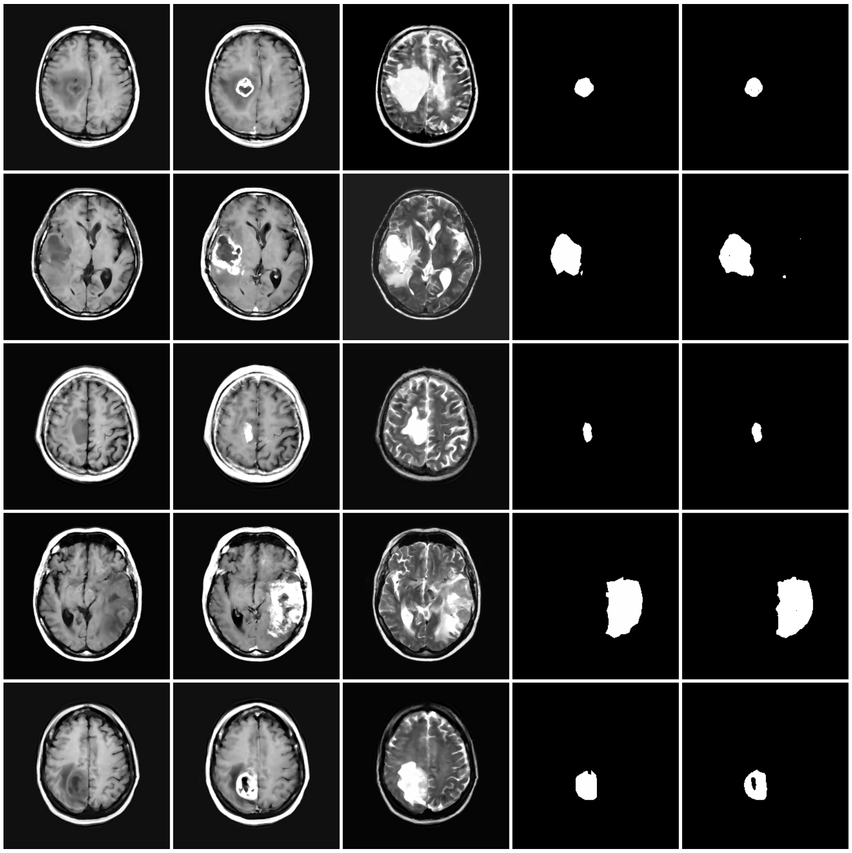


Figure 4.49: The results of a Support Vector Machine pixel classifier for segmenting enhancing tumor regions. Left to right: T1-weighted image, T2-weighted image, T1-weighted image after contrast agent injection, manual label of the enhancing tumor area, enhancing tumor area pixels predicted by the SVM. The Polynomial kernel was used, and the training data came from distant slices in the same volume. Thus, these results utilized *patient-specific* training, but the classifier did not have access to the labels for the slices shown.

Level Set active contour to model and constrain the tumor shape, similar to the methods applied in [Ho et al., 2002, Prastawa et al., 2004]. Markov Random Fields can also be used to relax probabilistic class estimates, and were applied previously in the task of tumor segmentation in [Gering, 2003b], which explored the ICM and the mean-field approximation algorithms. Assuming a Gaussian tissue model for the association potential (as in commonly done) would likely not give accurate results, and employing a logistic or non-parametric model may be more appropriate.

Conditional Random Fields are a relatively new formulation of Markov Random Fields that seek to model the data using a *discriminative* model as opposed to a *generative* model [Lafferty et al., 2001]. This simplification of the task allows the modeling of more complex dependencies in the labels, and the use of more powerful parameter estimation and inference methods. Several groups have recently formulated versions of Conditional Random Fields for image data, including [Kumar and Hebert, 2003]. Future implementations could explore methods such as these (that would simultaneously perform classification and relaxation).

Yet another method of relaxation that could be explored is to use a sequence of classifiers that train on both the features and the output of previous classifiers (including the predictions for neighboring pixels, or aggregations of these predictions). This would allow dependencies in the labels to be captured by a powerful classification model, while still considering dependencies in the features (in a much more computationally efficient way than in Conditional Random Fields). The disadvantage of this method is that it would require more training data, and its results may still need to be

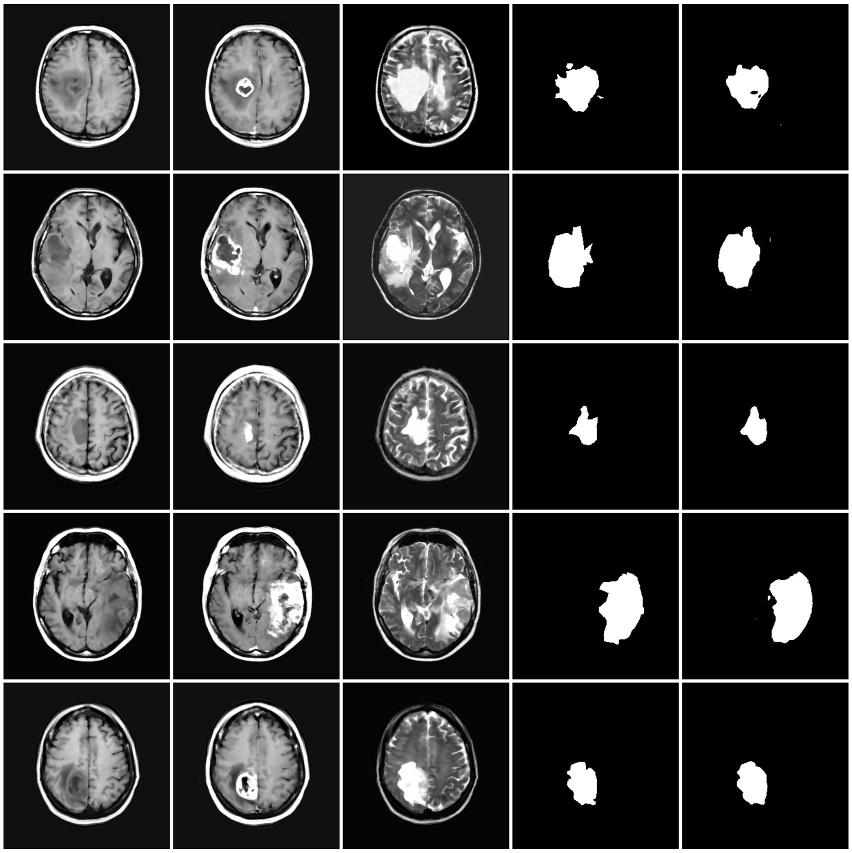


Figure 4.50: The results of a Support Vector Machine pixel classifier for segmenting tumor and edema regions. Left to right: T1-weighted image, T2-weighted image, T1-weighted image after contrast agent injection, manual label of the tumor and edema area, tumor and edema pixels predicted by the SVM. The Polynomial kernel was used, and the training data came from distant slices in the same volume. Thus, these results utilized *patient-specific* training, but the classifier did not have access to the labels for the slices shown.

relaxed.

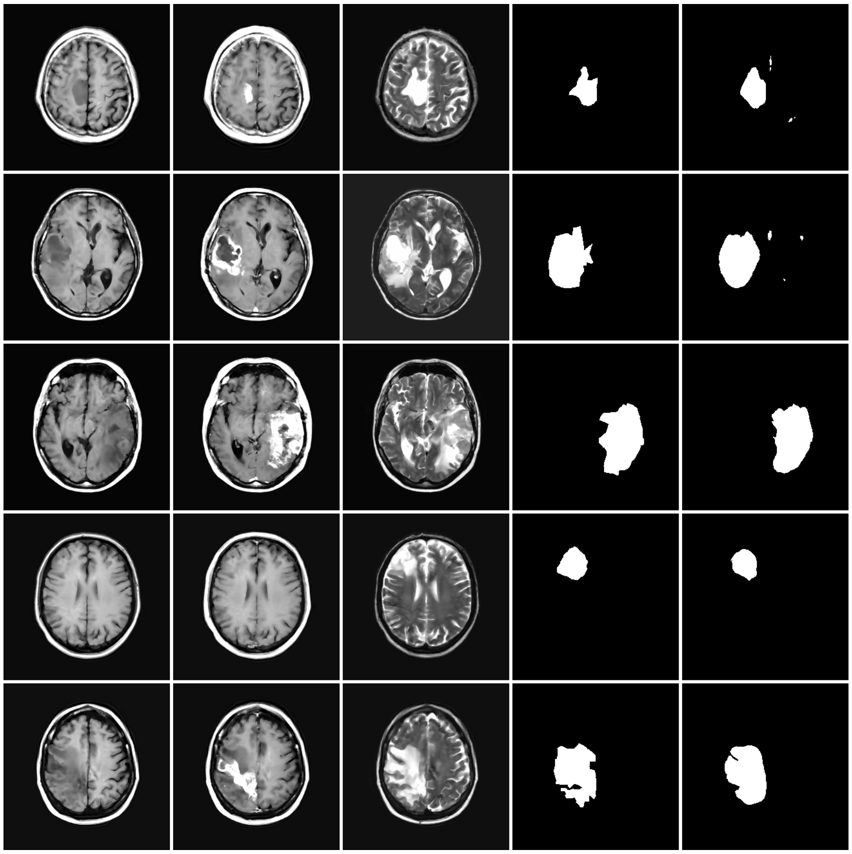


Figure 4.51: The results of a Support Vector Machine pixel classifier for segmenting tumor and edema regions with *inter-patient* training. Left to right: T1-weighted image, T2-weighted image, T1-weighted image after contrast agent injection, manual label of tumor and edema area, tumor and edema pixels predicted by the SVM. The Linear kernel was used, and the training data came from other patients. Thus, these results utilized *inter-patient* training, and did not have access to labels for any pixels from the patients classified.

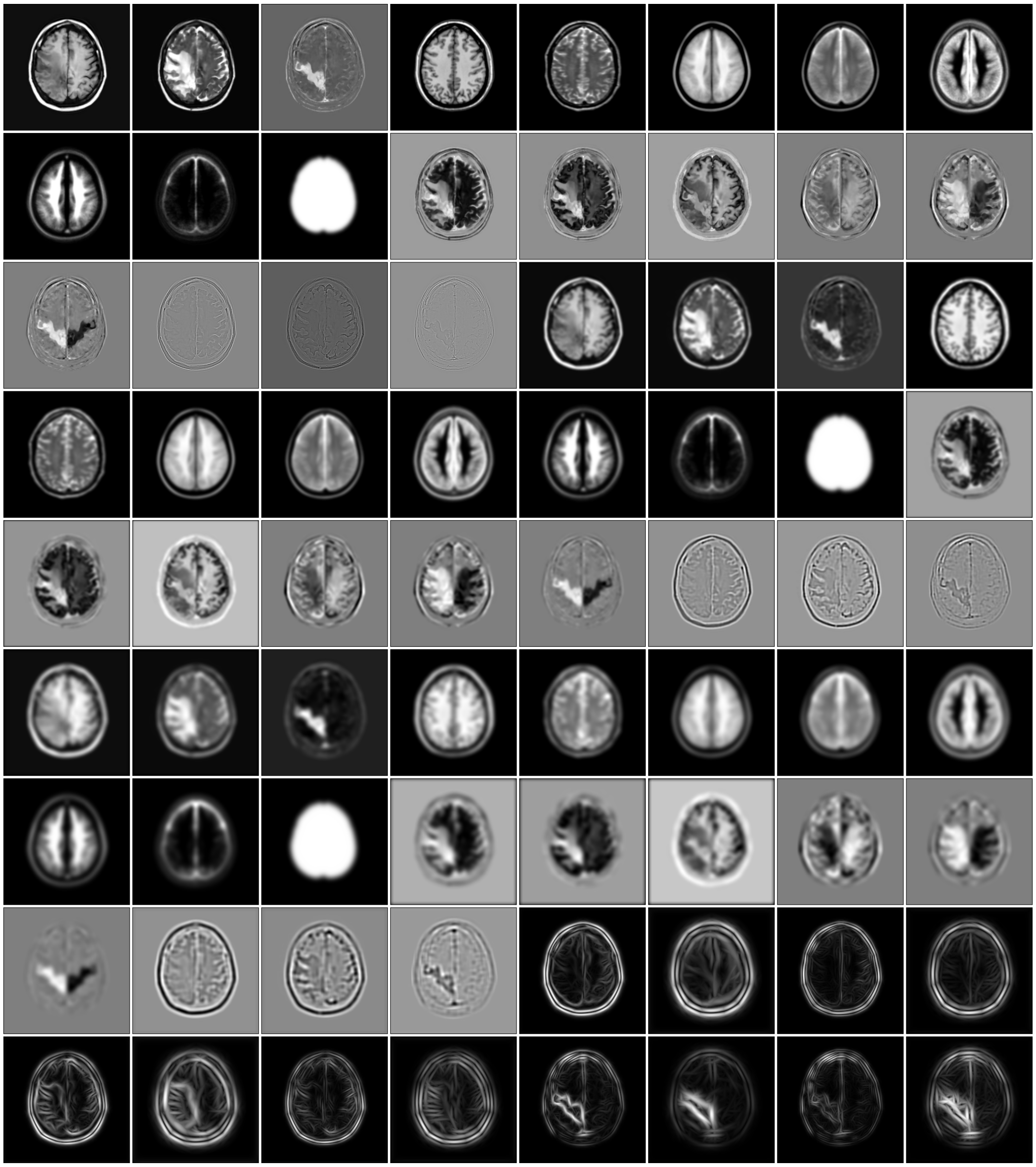


Figure 4.52: The full feature set extracted for the images in the last row of Figure 4.51 (25 features measured at 3 resolutions).

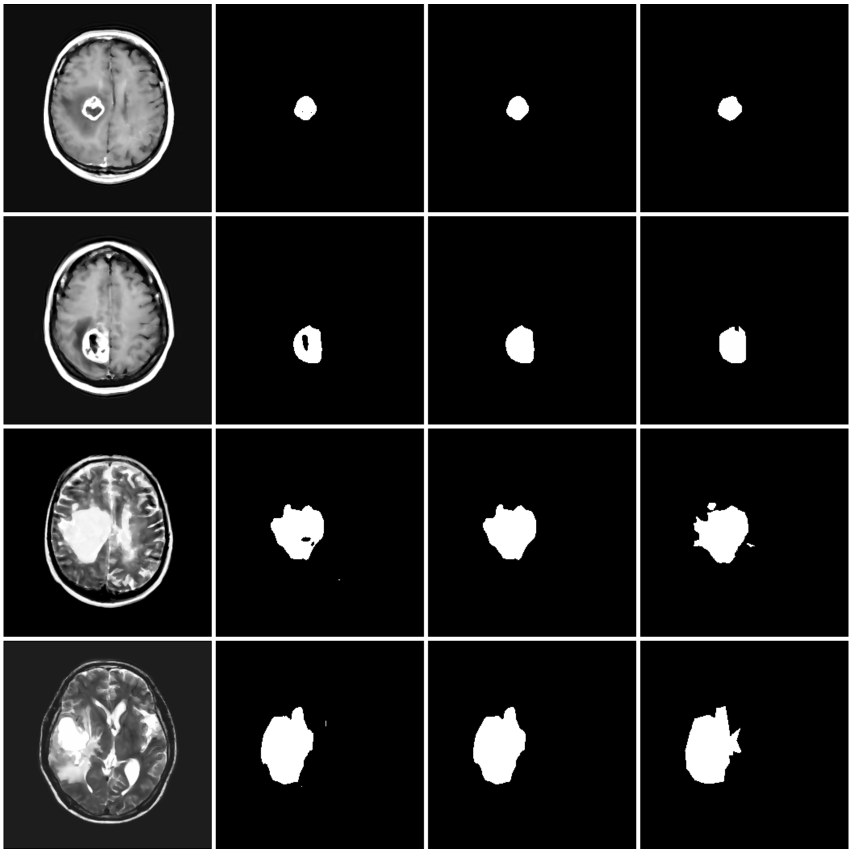


Figure 4.53: Results of label Relaxation. Left to right: Original image, the initial pixel classifications, the pixel classifications after label relaxation, the corresponding manual label. Holes have been filled and regions of small isolated labels have been removed in the relaxed images.

Chapter 5

Results

Assessing the performance of an automatic method for brain segmentation is not trivial. This is partially due to the lack of a standard data set, and the lack of availability of implementations of existing methods. In addition to these challenges, there are also the tasks of defining a performance evaluation measure, obtaining labeled training data, deciding how the training data should be obtained, and selecting a definition of abnormality. Before presenting the experimental results, we will discuss the decision made with respect to these issues.

A major issue that must be addressed in validating an automatic method for brain tumor segmentation is the means through which the segmentation is quantitatively assessed. Many of the approaches in the literature have used pixels or regions of interest that have been manually defined as containing a single tissue type in order to quantitatively assess segmentation quality. However, this represents a simplified task, since regions where the identity of tissues is ambiguous are not evaluated. It would be more desirable to have a method that performs effectively in the 'obviously' normal and abnormal areas, but that also makes similar decisions to a human expert in cases of ambiguity. It was thus decided to test on entire images and assess similarity to a manual segmentation over the entire image. This is more difficult than discriminating selected areas of obvious composition, but provides a more appropriate measure of the performance of the method for practical use.

In order to generate the labels used in training and for validating the testing performance of the system, a simple manual segmentation program was designed. This program allowed users to view each of the different modalities simultaneously, and to draw one or more contours around the abnormal area(s). Pixels within any of the defined contours were labeled as abnormal, while pixels outside of all of the contours were labeled as normal. This produces a binary segmentation for the entire image that can be used in training and to validate the system's performance on unseen test data. These contours were made on the original image data before any processing was performed, and the contours were verified by an expert radiologist.

In order to quantitatively assess the quality of an automatic binary segmentation in comparison to a manual binary segmentation produced by the above method, we chose to use the Jaccard measure for the abnormal class (where M is the set of manually defined tumor pixels, and A is the set pixels classified as tumor by the automatic method):

$$J(A, M) = \frac{A \cap M}{A \cup M} = \frac{tp}{tp + fp + fn} \tag{5.1}$$

In comparison to measuring the number of misclassifications, this measure is less sensitive to the size of the abnormality. And in comparison to measuring both precision and recall, the Jaccard measure provides a single easily interpretable statistic measuring the similarity between the two segmentations. This score will be 1 if the segmentations are identical, while it will approach 0 for completely dissimilar segmentations.

In order to measure significance, we used a Student's t-test of paired examples. In this case the score for each patient is an example, and the scores achieved by two different methods are *paired by patient*. This test measures whether the difference in performance between two methods is statistically significant based on the scores achieved across the patients. To compare two methods (A and B), the Student's t-test of paired examples is defined as follows: (x_{Ai} is the score for patient i with method A, x_{Bi} is the score for patient i with method B, and N is the number of patients) [Press et al., 1998]:

$$\overline{x_A} = \frac{1}{N}\sum_{i=1}^{N} x_{Ai} \tag{5.2}$$

$$\overline{x_B} = \frac{1}{N}\sum_{i=1}^{N} x_{Bi} \tag{5.3}$$

$$Var(x_A) = \frac{1}{N-1}\sum_{i-1}^{N}(x_{Ai} - \overline{x_A})^2 \tag{5.4}$$

$$Var(x_A) = \frac{1}{N-1}\sum_{i-1}^{N}(x_{Bi} - \overline{x_B})^2 \tag{5.5}$$

$$Cov(x_A, x_B) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{Ai} - \overline{x_A})(x_{Bi} - \overline{x_B}) \tag{5.6}$$

$$s_D = \left[\frac{Var(x_A)+Var(x_B)-2Cov(x_A,x_B)}{N}\right]^{1/2} \tag{5.7}$$

$$t = \frac{\overline{x_A}-\overline{x_B}}{s_D} \tag{5.8}$$

The significance of the t value can be determined using the incomplete beta function with $N-1$ degrees of freedom, yielding a value p representing the probability that the size of the t value occurred by chance [Press et al., 1998]. In the analysis that follows, a difference was considered significant if $p < 0.05$.

The source of the training and test data has a major influence on the performance of a supervised method. While some training and testing scenarios can achieve high accuracy with trivial methods, other scenarios require a higher level of generalization for the learned model (the level of generalization will be determined by both the classifier used and the associated feature set). A greater degree of generalization for the learned model is often indicative of a more practically useful method. Already discussed has been the differentiation between patient-specific training and inter-patient training, the latter requiring a significantly higher degree of generalization in order to achieve accurate results. However, there exist subclasses of these two general categories that have been examined in the literature on automatic brain tumor segmentation. Several of the subclasses of patient-specific training that have been examined include:

- Training and Testing data are the same: This case involves training and testing on the same pixels. Although this case may be useful to assess whether it is possible to discriminate normal and abnormal pixels with a classifier and feature set, this 'training' or 're-substitution' case represents a degenerate classification task. For example, high performance in this task could be achieved by ignoring the image data and using a sufficiently large set of randomly generated numbers as features, given a sufficiently expressive classifier.

- Training uses a subset of the pixels within the test slice: This represents a higher degree of generalization, and represents the potentially useful practical application of segmenting full images based on a set of manually selected points. However, feature sets with poor generalization properties (such as coordinates) can perform this task effectively due to the spatial correlation of pixels from the training and test set (unless a very small training set is used).

- Training is performed based on slices that are adjacent to the test slice: This case is similar to training on a subset of the pixels in the test slice, with two important distinctions: This is the first level of generalization that must be able to overcome potential errors or inconsistencies in the manual labels, and this is the first level that must overcome inter-slice intensity variations. Although this represents a more useful practical application, it can still often be addressed effectively using interpolation, a method that ignores the image and simply uses the labels given (from the adjacent slices) and their coordinates.

- Training is performed based on distant slices: This represents the most interesting scenario for patient-specific training. Since tumors can vary in location in cross-sectional views over larger distances, it is the only scenario where coordinate information is not sufficient to achieve a high accuracy (although it can still help). Tumors can also have different visual characteristics in distant slices, and thus the classifier and feature set must have adequate generalization properties to allow for this (in addition to a higher degree of variability in the manual labels). This scenario represents the most practical application of patient-specific training, since its results can significantly reduce the manual time needed to segment each slice in a volume.

We examined the latter level of generalization with respect to patient-specific training, since it must still address many of the challenges associated with inter-patient scenarios, but does not need to address the challenging issues of intensity non-standardization nor large variations in the visual appearance or location of the abnormality. The main motivation for performing experiments with patient-specific training was to evaluate, with a relatively small computational complexity, the performance of different feature sets and classifiers. With respect to *inter-patient* training, the following levels of generalization have been examined in previous systems for automatic brain tumor segmentation in a supervised framework:

- Training and Testing data are the same: This is similar to the patient-specific training scenario where training and testing data are the same, but in this case the training and testing data are sampled from different patients. As in the patient-specific training case, this represents a degenerate classification task that may be useful in evaluating classifiers and feature sets, but has limited practical applicability.

- Training on the same slice and tumor type from different patients: This scenario requires a similar degree of generalization to the highest level of generalization with patient-specific training. With large training sets, coordinate information can still help in this case (since it may be possible to characterize where in the image tumors are likely to occur, and what it expected to be seen at each spatial coordinate). Although this scenario must deal with intensity non-standardization it does not need to account for the large variability between pathologies.

- Training on the same tumor type from different patients: Training on different slices represents a much more challenging task than focusing on a single slice, due to the presence of the larger number of structures with different intensity characteristics that will be observed. This represents a very practical scenario, and larger training sets in this scenario are easier to obtain than in the ‘same slice’ scenario.

In our inter-patient experiments, we examined a more difficult case than those above. Our training data consisted of different slices *and* different tumor types. Furthermore, we used data from patients at different stages of treatment, and our inter-patient experiments used data from two different (1.5 Tesla) MRI scanners.

There exist different interpretations for the definition of the abnormal class in automatic tumor segmentation. Several that have been used in the literature include, in increasing order of difficulty, include:

1. Enhancing Tumor Pixels: Hyper-intense tumor pixels visible in T1-weighted images after the injection of a contrast agent.

2. Enhancing Tumor Area: A contour drawn around the hyper-intense enhancing tumor area. This case is more difficult than segmenting enhancing tumor pixels since this case includes pixels that are not hyper-intense present within the enhancing contour (such as necrotic regions), while the contour boundary may also extrapolate over areas that are not hyper-intense.

3. Homogeneous Tumor Area: Pixels associated with a tumor that has a relatively homogeneous intensity distribution. This case represents a more difficult task than enhancing tumor segmentation due to the potential lack of the discriminating information provided by the contrast agent.

4. Tumor and Edema Area: Pixels representing the edema area associated with a tumor. Since the edema area includes both the tumor and excess fluid, it is likely to be heterogeneous and will have areas with similar intensity properties to normal CSF. Furthermore, the exact extent of the edema boundary can be ambiguous in the image data.

5. Heterogeneous Tumor Area: Pixels associated with a heterogeneous tumor. This case is more difficult than edema segmentation since it further involves the discrimination between pixels that represent edema and those that represent tumor (a task that is inherently ambiguous in the modalities examined).

6. Gross Tumor Area: The definition of the gross tumor depends on the pathology type. Thus, although it will fall under one of the above definitions, this can vary across different tumor types. This can make it more difficult than the cases above for inter-patient scenarios since different images will have different abnormality definitions. However, if a specific tumor type is used that has a consistent definition of abnormality, this case would be equivalent to one of the cases above.

For our patient-specific training experiments, the tasks that were examined were the Enhancing Tumor Area (2), the Tumor and Edema Area (4), and the Gross Tumor Area (6). For the inter-patient training experiments, the Tumor and Edema Area (4) was the focus. This was due to the variability in the definition of the Gross Tumor Area, and the fact that many of the patients used for experimentation did not have an enhancing area.

The data set used for experimentation consisted of image data from 11 adults with primary brain tumors. The types of tumors examined were grade 2 astrocytomas, anaplastic astrocytomas, glioblastoma multiformes, and oligodendrogliomas. The latter tumor type represents a very challenging segmentation task in commonly used MR images, and we are not aware of any previous works that have attempted to segment this type of tumor automatically. For 3 of the patients, multiple segmented timepoints were available. Each of these times was used in the patient-specific training experiments, while only the most recent images were used in the inter-patient training experiments. The 11 patients were selected randomly from a database of patients with primary brain tumors. The selection criteria involved two elements: (i) the patient needed to have at least one fully segmented timepoint, and (ii) the patient had a T1-weighted pre- and post-contrast injection image and a T2-weighted image. This selection criteria led to patients with glioblastoma multiformes (the most common type of primary brain tumor) being the most common type of pathology in the experimental data (5 out of the 11 patients, and 10 out of the 17 segmented timepoints). This selection criterion has resulted in several very challenging cases being included in the experimentation, that represent cases where automatic methods would be expected to have significant difficulty.

5.1 Patient-Specific Training Experiments

This section will be organized as a series of related experiments designed to learn more about the problem in this training scenario. Each experiment will be motivated by a question, and this will be followed by the experimental outline, the predicted results, the experimental results, and a discussion of the results.

5.1.1 Patient-Specific Training Experiment 1

Question: Extensive effort has been expended in order to make the intensities more suitable for classification. A first natural question to ask is, therefore: *Is the intensity data more suitable for* *classification after Noise Reduction?*

Experimental Outline: To answer this question, several classifiers were trained and evaluated on data that underwent Noise Reduction using only the intensities as features. The same classifiers were trained and evaluated on the corresponding original raw data, also using only the intensities as features. The classification methods used represent four popular and computationally efficient classification techniques. We used the implementations in WEKA [Witten and Frank, 2000], with default parameters. The four classification models examined were:

- NB: A Naive Bayes classifier: This classifier computes a Maximum Likelihood model, assuming that the features are statistically independent given the classes. Although this assumption is clearly false in the task of tumor segmentation, it is possible that this highly efficient classifier may still achieve accurate results. Since the features are continuous, a quantization of the features was used, and thus the classifier learned was non-linear. This classifier was used in our experiments to evaluate the results of a classifier that only naively attempts to combine the different features.

- LOGIT: A Logistic Regression classifier: This classifier also uses a Maximum Likelihood model. Logistic Regression involves learning a linear function that minimizes the residual error assuming a Logistic model of the features. In addition to its simplicity, computational efficiency, and well-known statistical properties, this classifier was included since it represents an alternative linear classifier to Support Vector Machines.

- C4.5: The C4.5 Decision Tree learning algorithm: A popular Decision Tree Learning method. In constructing a tree, this method uses an entropy based measurement to evaluate potential new tree nodes during learning. This classifier would be expected to outperform the others if there are non-informative features, or if a small subset of the features can achieve a set of highly accurate decision rules.

- SVM: A linear Support Vector Machine: This classification method was outlined in the previous chapter. The linear kernel was used for this experiment since it is computationally more efficient to solve than in non-linear cases.

These classifers were evaluated over the following two feature sets:

- Raw: Raw intensities (T1-weighted, T2-weighted, and contrast agent difference image).

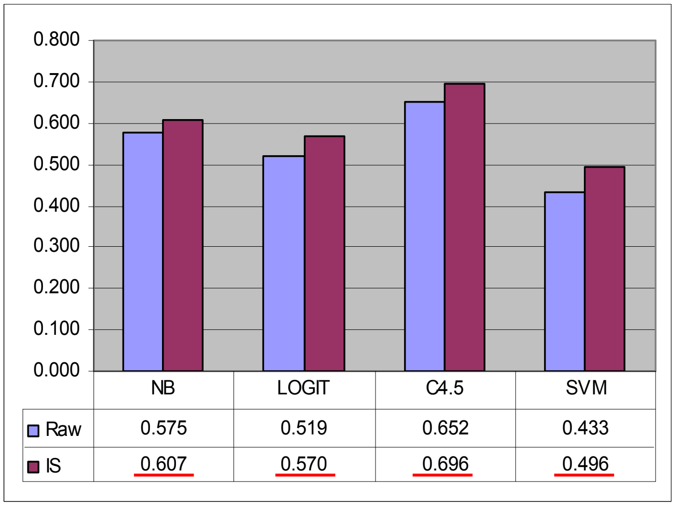
- IS: The same 3 intensities after Noise Reduction.

Predicted Results: The Noise Reduction should reduce the effects of local noise, inter-slice intensity variations, and intra-volume intensity inhomogeneity. This should increase standardization of the intensities within the volume, and it is therefore expected that higher classification accuracies should be observed in the Noise Reduced data for each of the classifiers.

Experimental Results: The average Jaccard scores for the tumor class over the 17 cases for the two feature sets, four classifiers, and three different definitions of abnormality are shown in Figures 5.1, 5.2, and 5.3.

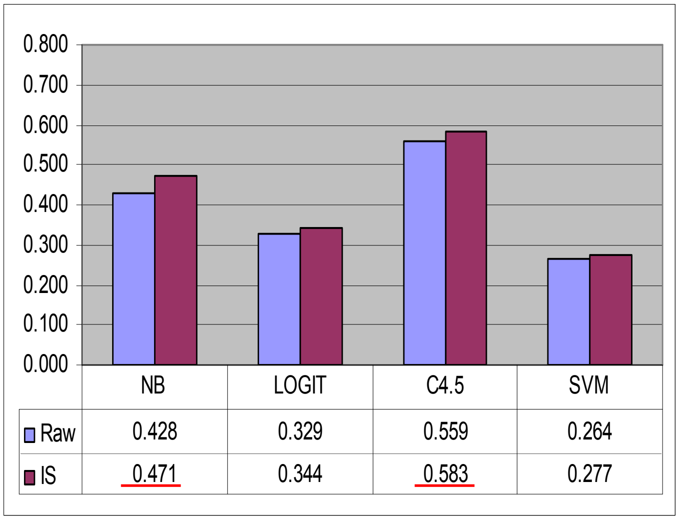
Discussion: A significantly higher score was achieved with the Noise Reduced data than with the raw data in all but 3 cases. This indicates that the intensity data is likely more suitable for classification after Noise Reduction.

In this experiment, the Decision Tree Learner with the IS feature set significantly outperformed the other classifiers and features sets for all three tasks. The two linear classifiers (LOGIT and SVM) had the worst performance in each task, indicating that a linear discriminant is likely not appropriate for classification of the intensity data. These results also support the hypothesis that segmenting the Gross Tumor or the Tumor and Edema are more challenging tasks than segmenting the Enhancing Tumor area. Furthermore, these results indicate that an intensity-based classifier with patient-specific training can perform a fairly effective Enhancing Tumor area segmentation, but that the other two tasks cannot be addressed as effectively by this strategy.



	NB	LOGIT	C4.5	SVM
■ Raw	0.575	0.519	0.652	0.433
■ IS	0.607	0.570	0.696	0.496

Figure 5.1: Average Enhancing Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where the *IS* feature set significantly outperformed the *Raw* feature set using the same classifier.



	NB	LOGIT	C4.5	SVM
■ Raw	0.428	0.329	0.559	0.264
■ IS	0.471	0.344	0.583	0.277

Figure 5.2: Average Tumor and Edema scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where the *IS* feature set significantly outperformed the *Raw* feature set using the same classifier.

5.1.2 Patient-Specific Training Experiment 2

Question: *Can the addition of distribution-based, coordinate-based and/or registration-based features (ie. additional pixel-level features) improve upon an intensity-based classification?*

Experimental Outline: This experiment used the same general format as the previous one. However, the different feature sets evaluated in this experiment explored the use of the noise reduced intensities combined with other pixel-level features. The features evaluated to augment an intensity-based classification were (the numbers in parenthesis indicate the total number of features):

- IS (3): Standardized intensities, no additional features.
- Dis (6): Intensities and the multi-spectral Euclidean distance from the pixel's intensities to the

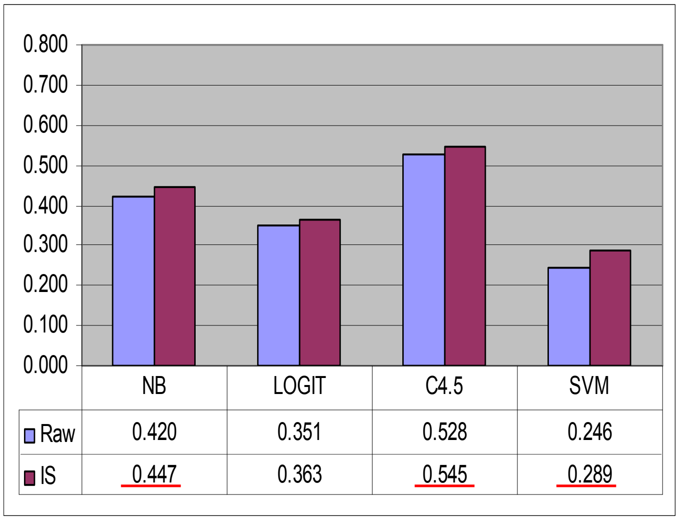


Figure 5.3: Average Gross Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where the *IS* feature set significantly outperformed the *Raw* feature set using the same classifier.

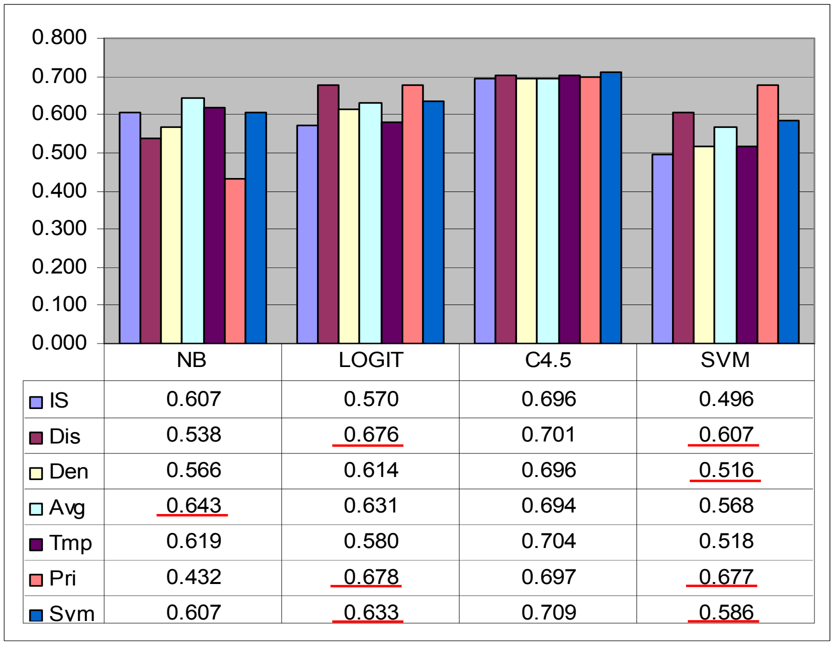
mean intensity in the template of gray matter, white matter, and CSF.

- Den (4): Intensities and the number of pixels in the pixel's multi-dimensional histogram bin.
- Avg (5): Intensities and the average intensities of a large number of individuals registered into the same coordinate system.
- Tmp (5): Intensities and the intensities of the template at the corresponding pixel location.
- Pri (7): Intensities and the prior probabilities for gray matter, white matter, CSF, and the brain area.
- Sym (6): The bi-lateral symmetry features.

Predicted Results: It is expected that several of these additional features could improve performance, while others may not. Based on the existing literature, we expect the priors to significantly improve the results. We also expect that symmetry and the expected intensities should also result in a performance improvement, since symmetry is often a very discriminating feature and the average intensities may be able to remove false positives based on spatial location. It is unclear whether the distances or density should improve results, since the distances assume an effective (and potentially non-linear) intensity standardization, while the densities are highly dependent on the histogram distribution (that varies from slice to slice). It is unclear whether the template intensities measured at the pixel-level will improve results, since these may not align exactly with the image data at a pixel level, and thus this could potentially represent a misleading feature.

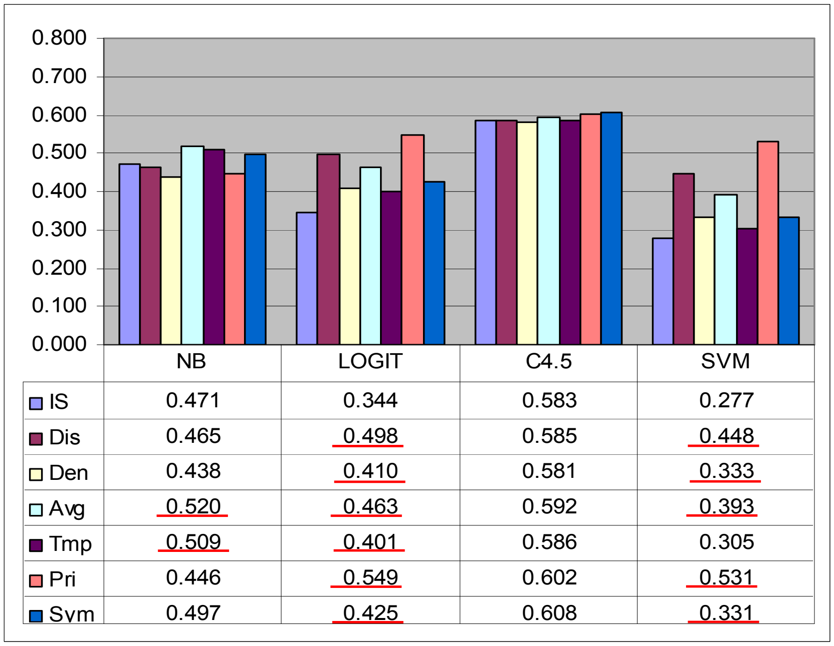
Experimental Results: The average scores are shown in Figures 5.4, 5.5, and 5.6.

Discussion: The use at least one of the types of additional pixel-level features offered a significant advantage of the intensity-based model for each classifier except C4.5. In particular, Logistic Regression and the SVM had the largest performance increases associated the additional features, narrowing the gap between these methods and the other two. The Naive Bayes method was most improved through the addition of the expected intensities (significant for all 3 tasks); This is interesting since it indicates that the expected intensities within the coordinate system may provide important prior knowledge about potential tumor locations, independent of the intensities. The addition of the spatial priors resulted in the largest increase in score for the Logistic Regression and SVM model, supporting the hypothesis that combining these values with the intensity data will improve



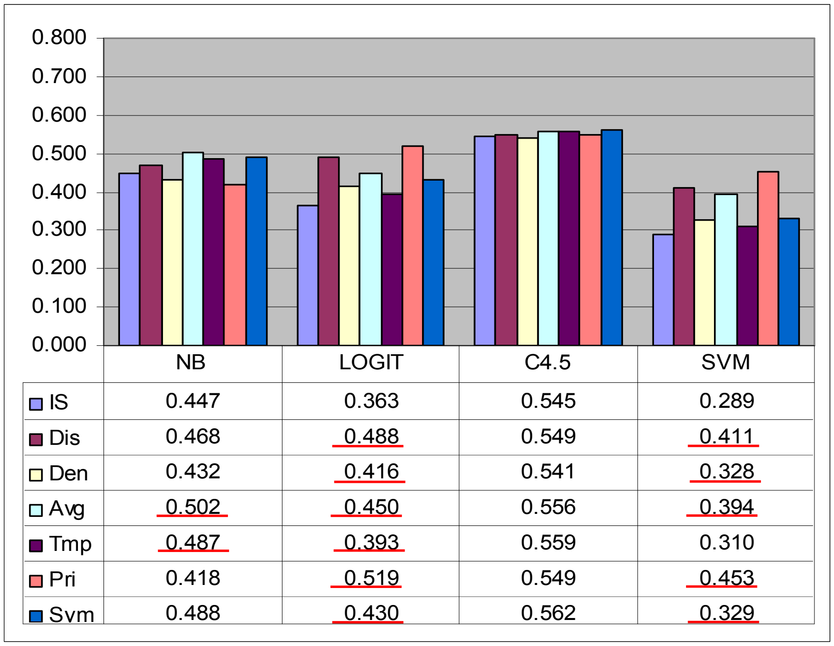
	NB	LOGIT	C4.5	SVM
▨ IS	0.607	0.570	0.696	0.496
■ Dis	0.538	0.676	0.701	0.607
□ Den	0.566	0.614	0.696	0.516
□ Avg	0.643	0.631	0.694	0.568
■ Tmp	0.619	0.580	0.704	0.518
■ Pri	0.432	0.678	0.697	0.677
■ Svm	0.607	0.633	0.709	0.586

Figure 5.4: Average Enhancing Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the *IS* feature set using the same classifier.



	NB	LOGIT	C4.5	SVM
▨ IS	0.471	0.344	0.583	0.277
■ Dis	0.465	0.498	0.585	0.448
□ Den	0.438	0.410	0.581	0.333
□ Avg	0.520	0.463	0.592	0.393
■ Tmp	0.509	0.401	0.586	0.305
■ Pri	0.446	0.549	0.602	0.531
■ Svm	0.497	0.425	0.608	0.331

Figure 5.5: Average Tumor and Edema scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the *IS* feature set using the same classifier.



	NB	LOGIT	C4.5	SVM
IS	0.447	0.363	0.545	0.289
Dis	0.468	0.488	0.549	0.411
Den	0.432	0.416	0.541	0.328
Avg	0.502	0.450	0.556	0.394
Tmp	0.487	0.393	0.559	0.310
Pri	0.418	0.519	0.549	0.453
Svm	0.488	0.430	0.562	0.329

Figure 5.6: Average Gross Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the *IS* feature set using the same classifier.

performance as observed for Multiple Sclerosis lesion segmentation in [Zijdenbos et al., 2002]. The highest classification accuracy in each task was achieved with the C4.5 classifier with the addition of the symmetry features, although in none of the cases was the difference between this classifier and feature set significant over the next closest classifier and feature set.

As a follow-up to this experiment, we explored whether the combination of several types of the additional pixel-level could result in higher scores. We thus evaluated an additional feature set:

- RB (17): Standardized intensities, distances to mean template intensities of normal tissue types, expected intensities, template intensities, tissue priors, the brain area prior, and bi-lateral symmetry.

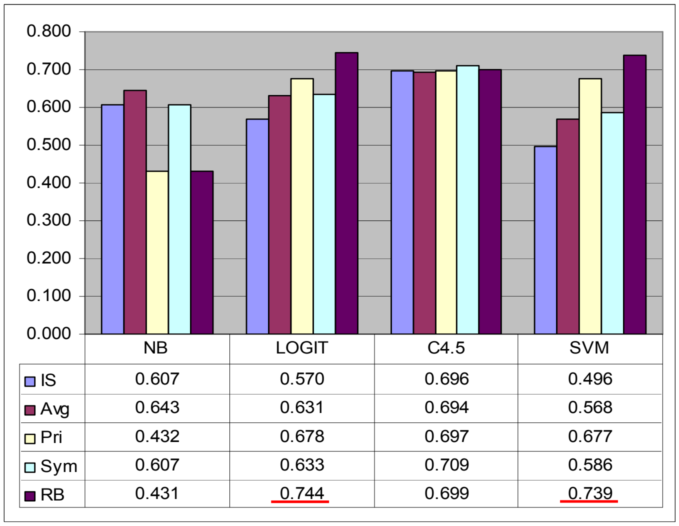
The results of this follow-up are shown in Figures 5.7, 5.8, and 5.9. The intensity-based feature set and the best feature set for each classifier from the previous experiment were included for comparison. This follow-up experiment indicates that, depending on the classification method used, a combination of these features can further improve classification results. Although the Naive Bayes and C4.5 method did not benefit from this combination, Logistic Regression and the SVM showed significant performance improvements. In particular, the Logistic Regression method with this feature set slightly outperformed the C4.5 method that used the intensities and bi-lateral symmetry across all 3 tasks (the best classifier and feature set combination from the original experiment).

5.1.3 Patient-Specific Training Experiment 3

Question: *Can image-based regional features improve an intensity-based classification?*

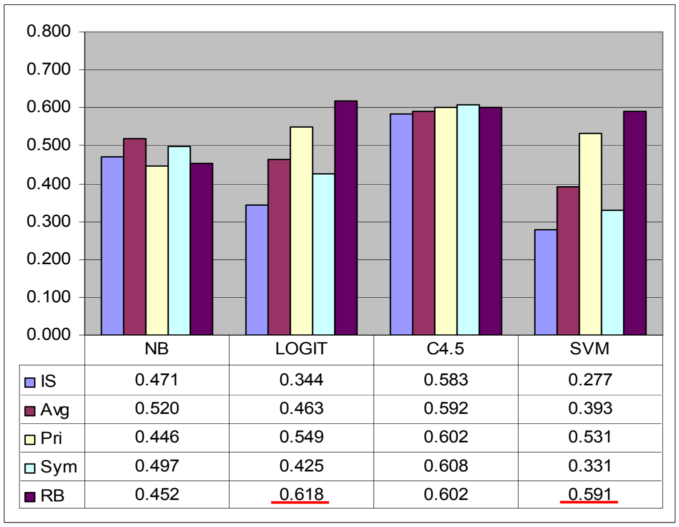
Experimental Outline: We explored this question in the same way that the previous question was explored. However, instead of using additional pixel-level features, additional region-based features were used to augment the intensity-based classification. The feature sets used were as follows:

- IS (3): Standardized intensities, no additional features.



	NB	LOGIT	C4.5	SVM
■ IS	0.607	0.570	0.696	0.496
■ Avg	0.643	0.631	0.694	0.568
□ Pri	0.432	0.678	0.697	0.677
□ Sym	0.607	0.633	0.709	0.586
■ RB	0.431	0.744	0.699	0.739

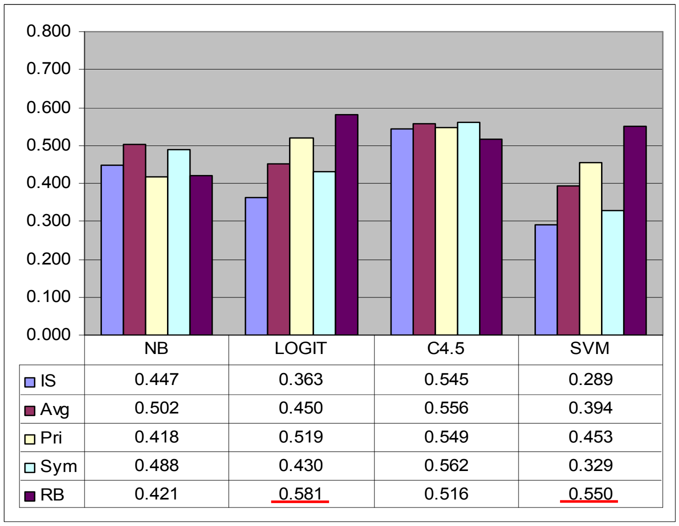
Figure 5.7: Average Enhancing Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where the *RB* feature set significantly outperformed each of the other feature sets using the same classifier.



	NB	LOGIT	C4.5	SVM
■ IS	0.471	0.344	0.583	0.277
■ Avg	0.520	0.463	0.592	0.393
□ Pri	0.446	0.549	0.602	0.531
□ Sym	0.497	0.425	0.608	0.331
■ RB	0.452	0.618	0.602	0.591

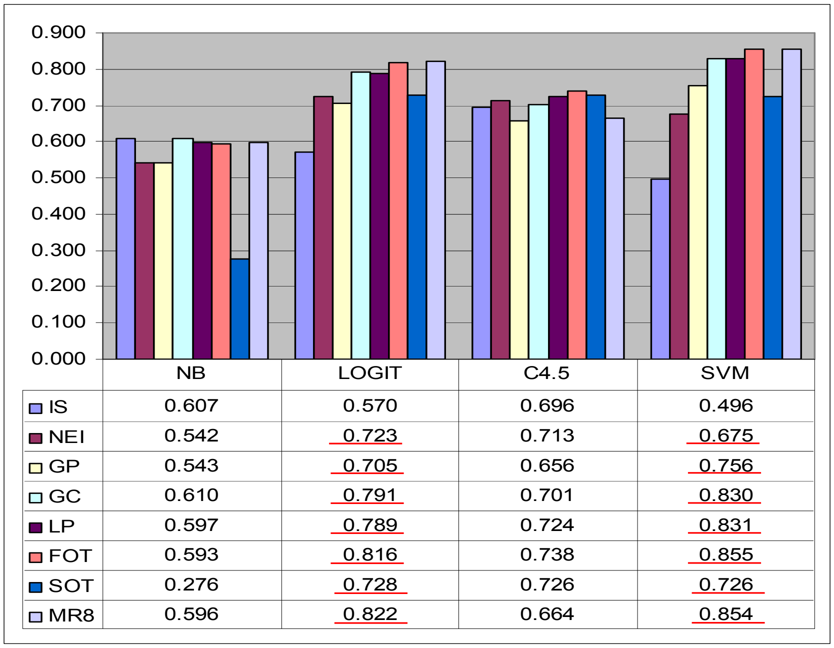
Figure 5.8: Average Tumor and Edema scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where the *RB* feature set significantly outperformed each of the other feature sets using the same classifier.

- NEI (75): Standardized intensities and the intensities of neighboring pixels (from a 5 by 5 region in each modality).
- GP (21): A Gaussian Pyramid of each of the image modalities (seven scales).
- GC (21): A Gaussian Cube of each of the image modalities (seven scales).
- LP (21): A Laplacian Cube of each of the image modalities (seven scales).
- FOT (21): First Order Textures in each of the image modalities.
- SOT (27): Second Order Textures in each of the image modalities.



	NB	LOGIT	C4.5	SVM
IS	0.447	0.363	0.545	0.289
Avg	0.502	0.450	0.556	0.394
Pri	0.418	0.519	0.549	0.453
Sym	0.488	0.430	0.562	0.329
RB	0.421	0.581	0.516	0.550

Figure 5.9: Average Gross Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where the *RB* feature set significantly outperformed each of the other feature sets using the same classifier.



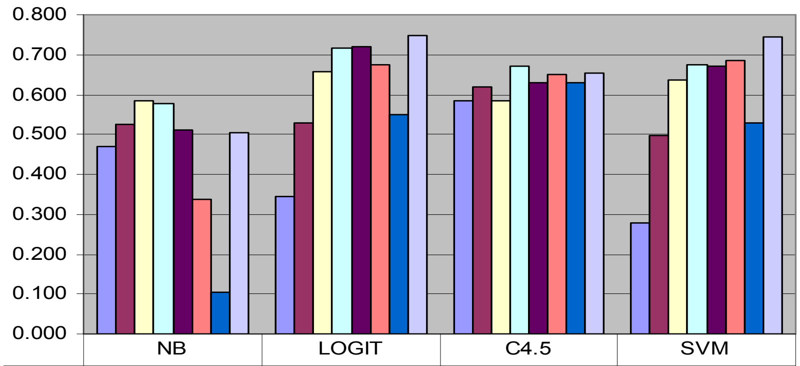
	NB	LOGIT	C4.5	SVM
IS	0.607	0.570	0.696	0.496
NEI	0.542	0.723	0.713	0.675
GP	0.543	0.705	0.656	0.756
GC	0.610	0.791	0.701	0.830
LP	0.597	0.789	0.724	0.831
FOT	0.593	0.816	0.738	0.855
SOT	0.276	0.728	0.726	0.726
MR8	0.596	0.822	0.664	0.854

Figure 5.10: Average Enhancing Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the *IS* feature set using the same classifier.

- MR8 (27): The MR8 Texture features in each of the image modalities (three scales).

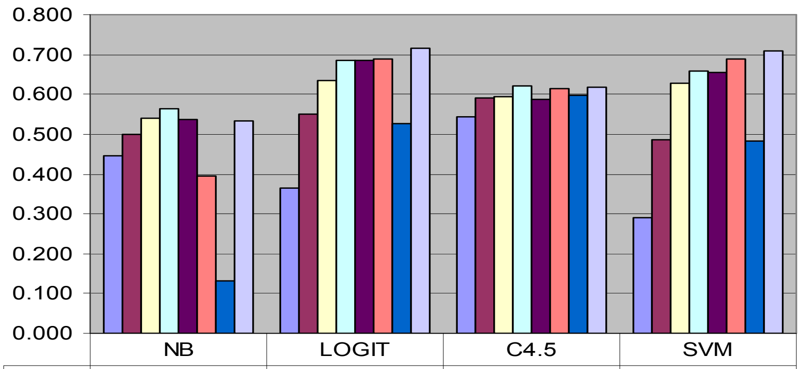
Predicted Results: It is expected that each of these additional feature sets will improve classification results. This is expected since they should confer resistance to mis-classifications made due to noise and partial volume effects. Furthermore, it is expected than regional properties should help disambiguate normal and abnormal regions that have similar intensities.

Experimental Results: The average scores are shown in Figures 5.10, 5.11, and 5.12.



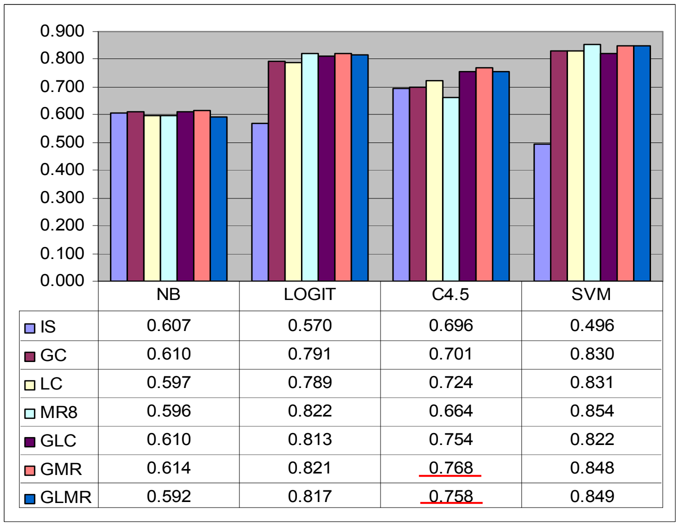
	NB	LOGIT	C4.5	SVM
IS	0.471	0.344	0.583	0.277
NEI	0.524	0.530	0.619	0.496
GP	0.583	0.656	0.586	0.636
GC	0.578	0.715	0.671	0.674
LP	0.510	0.720	0.628	0.672
FOT	0.336	0.675	0.651	0.686
SOT	0.105	0.549	0.629	0.530
MR8	0.505	0.749	0.655	0.746

Figure 5.11: Average Tumor and Edema scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the *IS* feature set using the same classifier.



	NB	LOGIT	C4.5	SVM
IS	0.447	0.363	0.545	0.289
NEI	0.499	0.552	0.592	0.486
GP	0.538	0.634	0.594	0.626
GC	0.563	0.687	0.622	0.657
LP	0.535	0.687	0.586	0.656
FOT	0.396	0.688	0.613	0.690
SOT	0.131	0.527	0.597	0.483
MR8	0.534	0.716	0.617	0.710

Figure 5.12: Average Gross Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the *IS* feature set using the same classifier.



	NB	LOGIT	C4.5	SVM
IS	0.607	0.570	0.696	0.496
GC	0.610	0.791	0.701	0.830
LC	0.597	0.789	0.724	0.831
MR8	0.596	0.822	0.664	0.854
GLC	0.610	0.813	0.754	0.822
GMR	0.614	0.821	0.768	0.848
GLMR	0.592	0.817	0.758	0.849

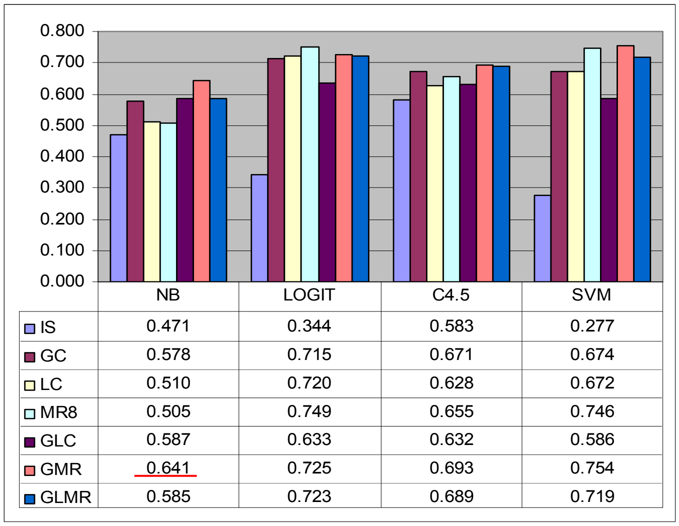
Figure 5.13: Average Enhancing Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the $\{IS,GC,LC,MR8\}$ feature sets using the same classifier.

Discussion: With the exception of segmenting the Enhancing Tumor area with Naive Bayes or C4.5, the performance of each classifier on each task could be significantly improved through the use of region-based features. Although each type of region-based feature did not strictly improve the results for each classifier and task, in many cases the region based features offered a significant (and often very large) improvement. The most significant improvements were again seen in the two linear classifiers (that notably learn to simultaneously combine all of the features), in some cases more than doubling the score of the intensity model. These two classifiers achieved the highest scores with the MR8 textures, while performing almost as effectively with the other linear filtering methods (the Gaussian and Laplacian Cubes), and the first-order textures. Second-order textures, neighboring intensities, and the Gaussian Pyramid also produced significant performance gains with the linear classifiers, but these gains were not as large. The Naive Bayes classifier seemed to benefit the most from the three methods based on linear filtering, the Gaussian Pyramid, and the intensities of neighboring pixels. The performance of the C4.5 classifier was often improved with region-based features (especially in the two harder tasks), and seemed to benefit in particular across the three tasks from the Gaussian Cube and the first-order textures.

Since each of the classifiers generally tended to benefit from the linear filtering features (the Gaussian and Laplacian Cubes, and the MR8 textures), a follow-up experiment evaluated whether combining these features would further improve results. This experiment evaluated the following feature sets:

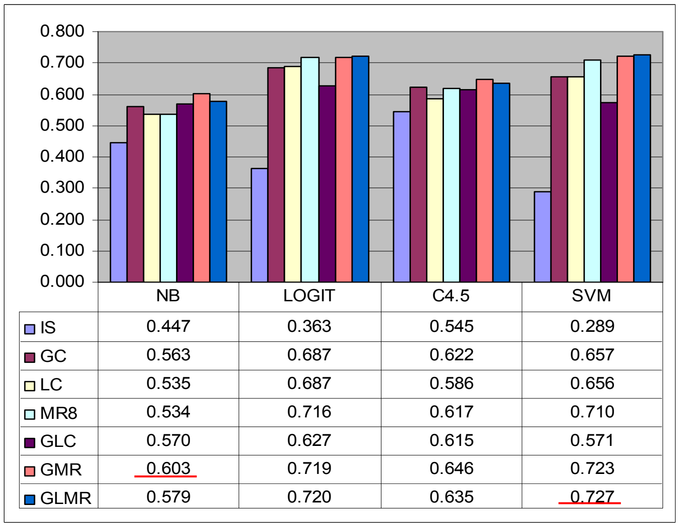
- GLC (18) A Gaussian and Laplacian Cube representation (3 scales)
- GMR (21) A Gaussian Cube (3 scales) and a subset of the MR8 Gabor responses (2 scales).
- GLMR (30) A Gaussian and Laplacian Cube representation (3 scales), and a subset of the MR8 Gabor responses (2 scales).

The results from this experiment are shown in Figures 5.13, 5.14, and 5.15. Based on these results, it is not clear whether this combination (at fewer scales) would be generally beneficial compared to the individual techniques at a larger number of scales.



	NB	LOGIT	C4.5	SVM
IS	0.471	0.344	0.583	0.277
GC	0.578	0.715	0.671	0.674
LC	0.510	0.720	0.628	0.672
MR8	0.505	0.749	0.655	0.746
GLC	0.587	0.633	0.632	0.586
GMR	0.641	0.725	0.693	0.754
GLMR	0.585	0.723	0.689	0.719

Figure 5.14: Average Tumor and Edema scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the {*IS*,*GC*,*LC*,*MR8*} feature sets using the same classifier.



	NB	LOGIT	C4.5	SVM
IS	0.447	0.363	0.545	0.289
GC	0.563	0.687	0.622	0.657
LC	0.535	0.687	0.586	0.656
MR8	0.534	0.716	0.617	0.710
GLC	0.570	0.627	0.615	0.571
GMR	0.603	0.719	0.646	0.723
GLMR	0.579	0.720	0.635	0.727

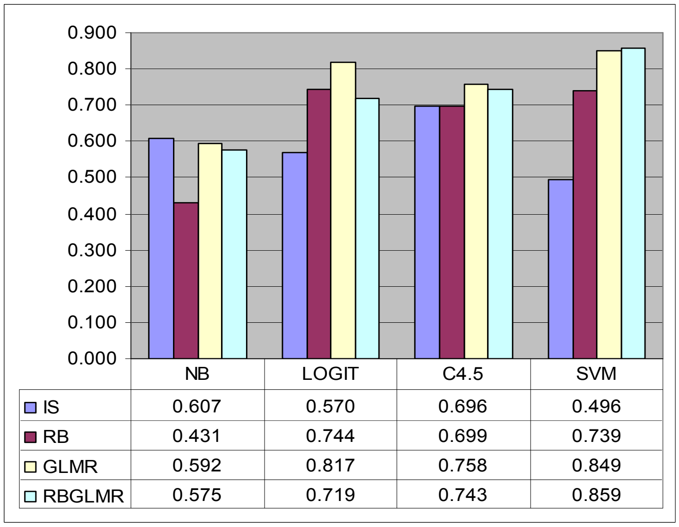
Figure 5.15: Average Gross Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed the {*IS*,*GC*,*LC*,*MR8*} feature sets using the same classifier.

5.1.4 Patient-Specific Training Experiment 4

Question: *Can additional pixel-based features and region-based features be combined to yield higher classification accuracy than either would individually?*

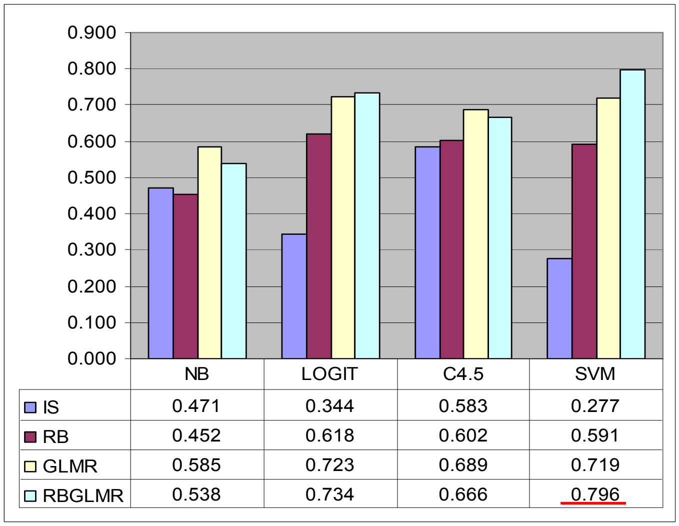
Experimental Outline: To attempt to answer this question, we compared the following feature sets:

- IS (3): Standardized intensities.
- RB (17): 17 pixel-level (image-based, coordinate-based, and registration-based) features from the second experiment.



	NB	LOGIT	C4.5	SVM
IS	0.607	0.570	0.696	0.496
RB	0.431	0.744	0.699	0.739
GLMR	0.592	0.817	0.758	0.849
RBGLMR	0.575	0.719	0.743	0.859

Figure 5.16: Average Enhancing Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed each of the other feature sets using the same classifier.



	NB	LOGIT	C4.5	SVM
IS	0.471	0.344	0.583	0.277
RB	0.452	0.618	0.602	0.591
GLMR	0.585	0.723	0.689	0.719
RBGLMR	0.538	0.734	0.666	0.796

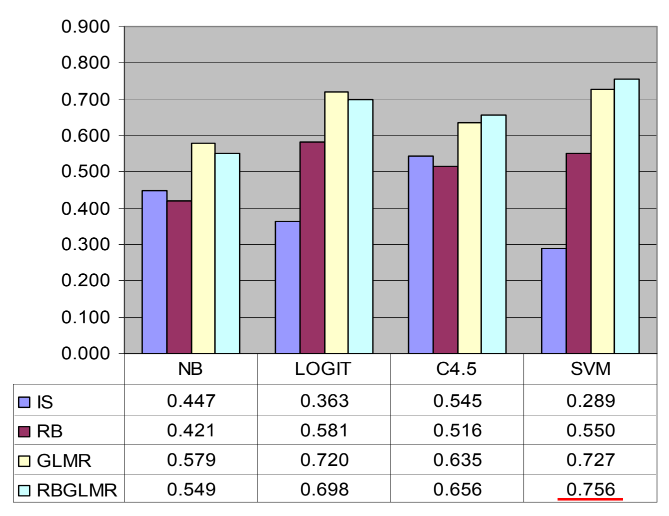
Figure 5.17: Average Tumor and Edema scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed each of the other feature sets using the same classifier.

- GLMR (30): 30 region-level image-based features from the previous experiment.
- RBGLMR (60): The GLMR feature set, in addition to the RB feature set measured at three scales with a Gaussian Cube representation.

Predicted Results: We anticipate that measuring region-level properties of our diverse pixel-level features will achieve higher scores than the region-level properties or pixel-level features would achieve individually. Thus, we expect that RBGLMR will outperform both the RB and the GLMR feature sets.

Experimental Results: The average scores are shown in Figures 5.16, 5.17, and 5.18.

Discussion: The Naive Bayes classifier did not benefit from this combination, since the combined feature set was outperformed by the purely image-based GLMR feature set. The Logistic Re-



	NB	LOGIT	C4.5	SVM
IS	0.447	0.363	0.545	0.289
RB	0.421	0.581	0.516	0.550
GLMR	0.579	0.720	0.635	0.727
RBGLMR	0.549	0.698	0.656	0.756

Figure 5.18: Average Gross Tumor scores for different classifiers and feature sets with patient-specific training (over 17 images). Values that are underlined are cases where a feature set significantly outperformed each of the other feature sets using the same classifier.

gression and C4.5 classifier each achieved a slightly higher score with the combination for one task, but the combination hurt performance in the other two tasks. The combined feature set improved the score of the SVM model for each of three tasks. The improvement was significant in the two harder tasks. Based on these experiments, it is not clear that the combination of diverse pixel-based features and region-based features necessarily provides a major benefit over region-based features. However, it is noteworthy that the SVM model using both the diverse pixel-level features and the region-level features achieved the highest accuracy among all the feature sets and classifiers examined thus far for each of the three tasks, and that this difference was significant for the two harder tasks.

5.1.5 Patient-Specific Training Experiment 5

Question: *Can classifiers that take into account more complex dependencies in the features improve results?*

Experimental Outline: This experiment examined the results obtained with different classification models for the RBGLMR feature set introduced in the previous experiment. This was motivated by the idea that more complex and computationally intensive classification models may be able to achieve higher accuracy than the four simple and efficient classifiers used in the previous experiments. Four additional classifiers were evaluated that were analogous to the four classifiers used in the previous experiment. These additional classifiers were:

- TANB: A Tree-Augmented Naive Bayes model: This classifier augments a Naive Bayes model with additional tree nodes. This allows the classifier to take into account dependencies between the features, and therefore relaxes the statistical independence assumptions made in the Naive Bayes model.

- ANN: An Artificial Neural Network: This classifier augments a Logistic Regression model with 'hidden layers' that allow non-linear dependencies in the features to be taken into account. The number of hidden layers was reduced from the default value to 1, which produced the best performance (in agreement with [Dickson and Thomas, 1997]).

- B-C4.5: A Boosted C4.5 Decision Tree Learner: Boosting is a technique that improves upon individual classification models by learning a series of classifiers. Each classifier puts increased weight on the examples misclassified by the previous classifier. This series of classifiers is weighted and combined to make a final classification, and can potentially achieve a higher classification rate than the individual classifiers.

- SVM-PK: A Support Vector Machine with the Polynomial Kernel: The ‘kernel trick’ was introduced in the previous chapter. This technique allows the classifier to take into account non-linear dependencies in the features by embedding the features in a high dimensional space where the classes are more likely to be linearly separable. A degree of 2 was used.

The implementations from [Witten and Frank, 2000] were used for three of these classifiers (using default parameters with the exception of the number of hidden layers in the ANN), while the implementation of [Joachims, 1999] was used for the SVM model. In addition, three special classifiers were evaluated. These classifiers were:

- AVG: Naive Model Averaging: This meta-classifier classifies pixels based on the prediction made by the majority of the classifiers among the set {TANB,LOGIT,ANN,C4.5,B-C4.5,SVM,SVM-PK}. Similar to Boosting, this simple method could potentially achieve higher accuracies than the individual techniques.

- MAX: Model Selection based on test scores: This meta-classifier classifies pixels based on the method that achieves the highest test score for the patient among the set {NB,TANB,LOGIT,ANN,C4.5,B-C4.5,SVM,SVM-PK}. This classifier was included to assess how far the best classifier on average diverges from the best classifier in each individual case. Note that the use of test scores means that this method ‘cheats’ (the only method to do so), since it uses the test slice labels.

- INTERP: Classification by (trilinear) interpolation: This classifier interpolates between the training slices to make classifications. Since this patient-specific task can be interpreted as being an interpolation problem, this classifier was included to evaluate the accuracy of using interpolation.

Predicted Results: We anticipate that each extended classifier will perform at a similar level or better than each of the corresponding techniques from the previous experiment. Due to the similar levels of accuracies between the different classifiers in the previous experiment, we anticipate that the simple model averaging technique will have a level of performance that is similar to the best classifiers. The MAX classifier will by definition outperform the individual methods, but it will be interesting to see how far the best classifiers diverge from score. We anticipate that the interpolation method may outperform the weaker classifiers, but it is unlikely to achieve the high degree of accuracy observed with the best classifiers, such as the SVM from the previous experiment.

Experimental Results: The average scores are shown in Figures 5.19, 5.20, and 5.21.

Discussion: Each of the more advanced classifiers improved the accuracy of the simpler analogous classifier, although this improvement was not always significant. Specifically, the Tree-Augmented Naive Bayes model significantly outperformed Naive Bayes for the Enhancing Tumor and Gross Tumor cases, the Artificial Neural Network significantly outperformed Logistic Regression for the task of Enhancing tumor task, Boosted C4.5 significantly outperformed C4.5 for all 3 tasks, and finally the SVM with Polynomial Kernel did NOT significantly outperform the linear SVM for any task. Notably, combining the C4.5 method with Boosting yields one of the most effective individual classifiers (only the SVM models performed better). Although the SVM with the Polynomial Kernel did not significantly outperform the linear SVM (nor Boosting in the Tumor and Edema case), it had the highest score among the individual classifiers across the 3 tasks. The Model Averaging technique significantly outperformed the Polynomial Kernel for only the Gross Tumor task. Both Model Averaging and the SVM with the Polynomial Kernel achieved an average score that was relatively close to the MAX classifier, this indicates that they consistently achieve among the best classification results. However, the MAX classifier achieved a significantly higher score

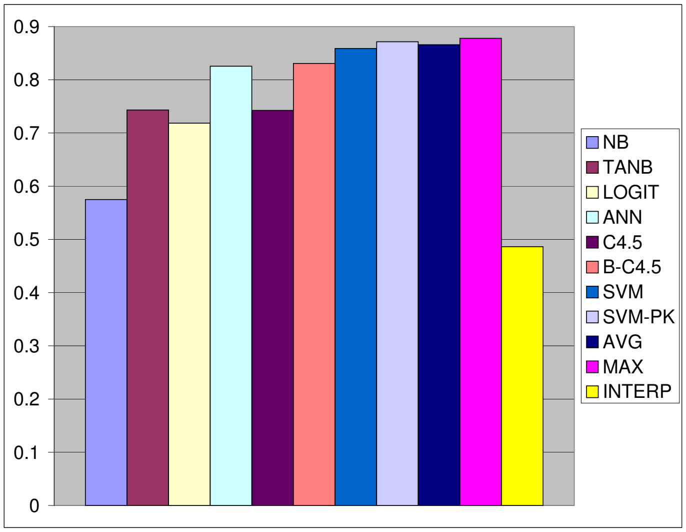


Figure 5.19: Average Enhancing Tumor scores for different classifiers with the RBGLMR feature set and patient-specific training (over 17 images).

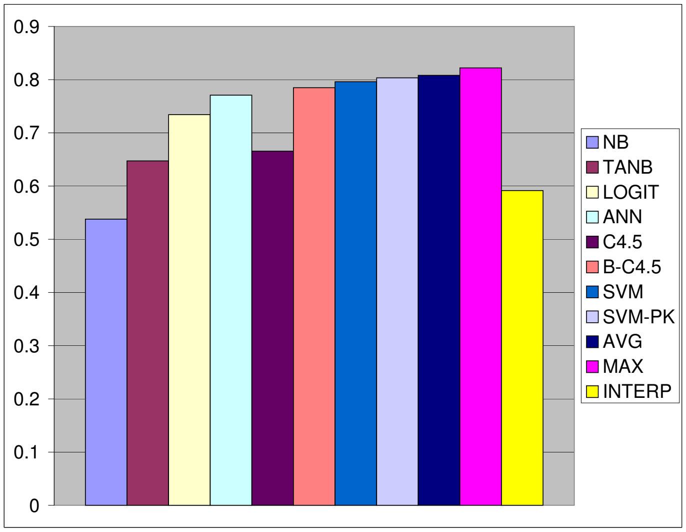


Figure 5.20: Average Tumor and Edema scores for different classifiers with the RBGLMR feature set and patient-specific training (over 17 images).

than these classifiers in all but the Gross Tumor case. With three exceptions (Naive Bayes, Tree-Augmented Naive Bayes, and C4.5), the classification methods significantly outperform the spatial interpolation method for all 3 tasks.

5.1.6 Patient-Specific Training Experiment 6

Question: *Does relaxation of the classification results improve performance?*

Experimental Outline: This experiment compared the scores achieved by the classifiers in the previous experiment to their scores after relaxation with the morphological operations described in the previous Chapter.

Since the relaxed classifications represent the final output of the system, an additional 'special' classifier was included in this experiment:

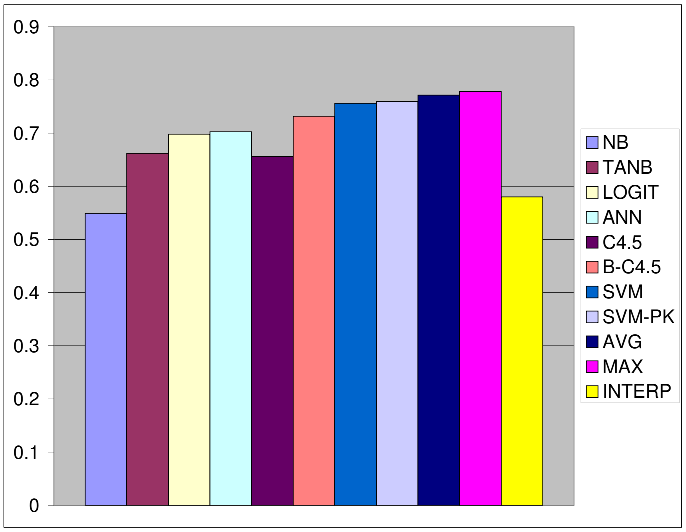


Figure 5.21: Average Gross Tumor scores for different classifiers with the RBGLMR feature set and patient-specific training (over 17 images).

- Manual: An additional manual segmentation, generated by drawing the abnormal region on the image *after viewing the labels of the training data*. This represents an approximation upper bound on the potential performance of the system.

Predicted Results: It is expected that the relaxation step should in general improve performance. We anticipate that the classifiers with the lowest accuracy will benefit the most from a relaxation of the labels, while more accurate classifiers will not benefit from this additional processing.

The performance of the Manual method provides further support for the order of difficulty between the three tasks. The Manual classifier had the best performance on the Enhancing Tumor task, and the worst performance on the Gross Tumor task. In the Enhancing Tumor task, the Manual method significantly outperformed most of the individual classifiers. However, the Manual method did not significantly outperform the Relaxed SVM models (nor the Relaxed Model Averaging strategy). In the Tumor and Edema task, the Manual method significantly outperformed all other methods. Finally, in the Gross Tumor task, the Manual method outperformed all methods except the Relaxed Model Averaging classifier (and the oracle-based MAX classifier).

Experimental Results: The average scores are shown in Figure Figures 5.22, 5.23, and 5.24.

Discussion: In all but one case, relaxation of the classification results improved the score by at least a small amount. This improvement was significant in all cases, with 5 exceptions. For the Tumor and Edema task, the scores of Logistic Regression, SVMs with the Polynomial Kernel, and Model Averaging were not significantly improved after relaxation. And for the Gross Tumor task, the gains from relaxation of the Linear SVM and Model Averaging were not significant. The improvement did not seem to be highly dependent on the initial classification accuracy.

The difference between the Manual metric and both SVM classifiers was not significant for the Enhancing Tumor case. For the Tumor and Edema case, the difference between the Manual score and the individual classifiers was significant (only in the case of MAX was the difference between the Manual score and others not significant). For the Gross Tumor case, the Manual score was significantly higher than all individual classifiers, but was not significantly higher than the Model Averaging score. The closer similarity between the automatic method and the Manual method in the Gross Tumor case compared to the Tumor and Edema case can be explained by the higher degree of ambiguity in the Gross Tumor case that leads to a much lower Manual performance.

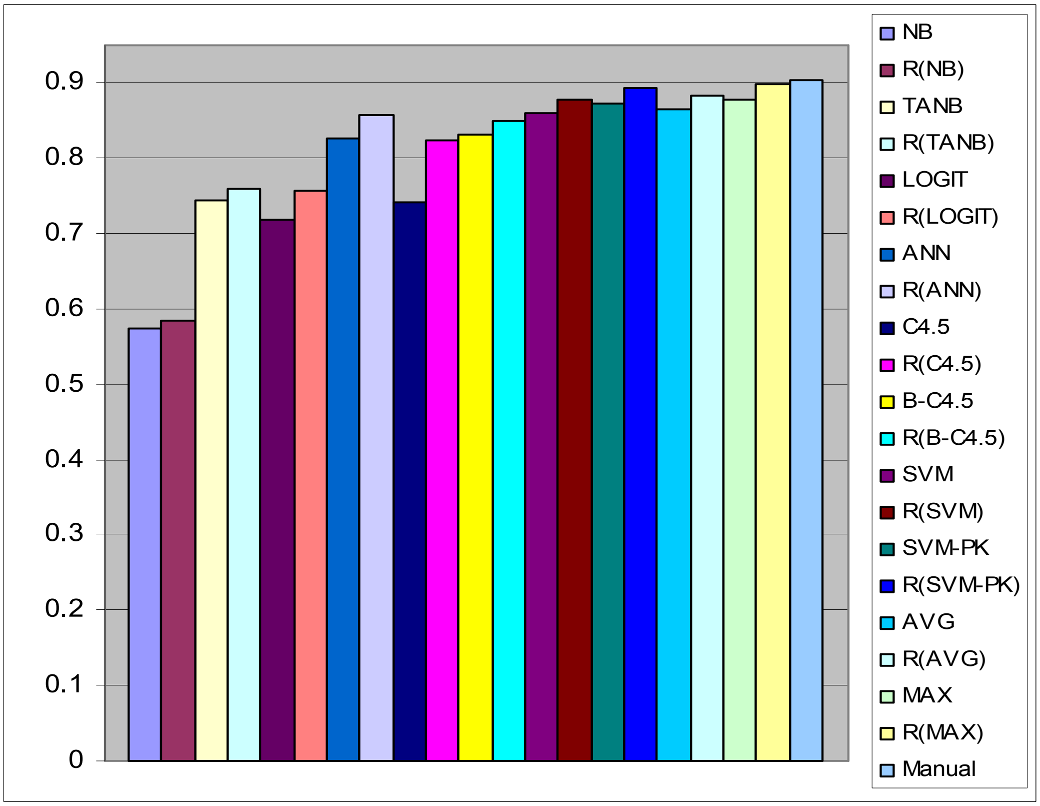


Figure 5.22: Average Enhancing Tumor scores for different classifiers with the RBGLMR feature set and patient-specific training before and after relaxation (over 17 images).

5.1.7 Patient-Specific Training Experiment 7

Question: *How accurate is the final system in each test case, and are the errors primarily made near boundaries where the class ambiguity and manual errors in the labels will be the highest?*

Experimental Outline: To explore this question, we examined the results of each test case under the score previously used, but also under a score that did not penalize for false positives and negatives located at boundary pixels. We defined a boundary pixel as a pixel that has both normal and abnormal pixels in the manual labels of its 5 by 5 neighborhood. The SVM classifier was used in this experiment with the Polynomial Kernel and the RBGLMR feature set (that includes regional image-based, coordinate-based, and registration-based features), since this classifier proves to be one of the most accurate, while being computationally much faster to train perform inference with than the slightly more accurate Model Averaging classifier. We examined the results after label relaxation, since this represents the final output of the system.

Predicted Results: By definition, the scores that do not include false positives and negatives at boundary pixels will be higher than the normal scores. However, since many of the segmentations are very similar to the manual segmentations, we expect to see very high scores in many of the cases under this metric.

Experimental Results: The individual scores for the three tasks are shown in Tables 5.1, 5.2, and 5.3.

Discussion: With respect to the task of Enhancing Tumor area segmentation, these results indicate that the final system is performing nearly perfect segmentations, and is only misclassifying a virtually negligible number of pixels near the enhancing boundary. A slightly diminished level of performance is observed for the segmentation of the Tumor and Edema area. However, with only a few exceptions, these segmentations were typically still very accurate (sometimes achieving perfect

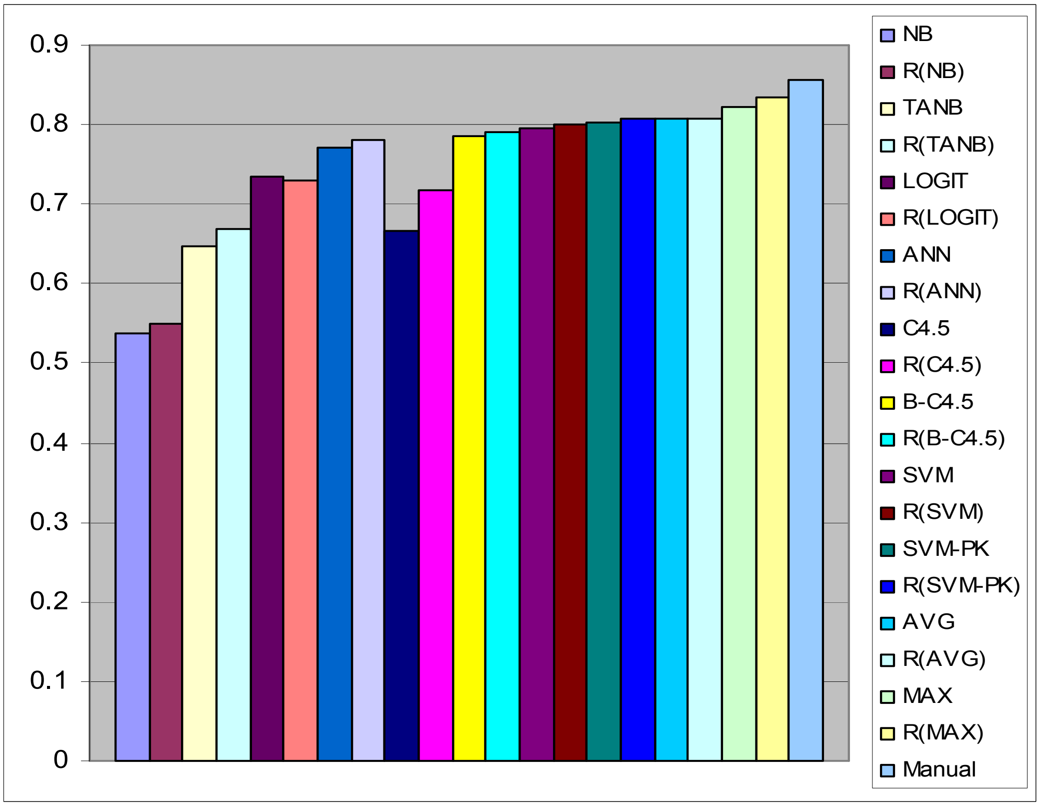


Figure 5.23: Average Tumor and Edema scores for different classifiers with the RBGLMR feature set and patient-specific training before and after relaxation (over 17 images).

Table 5.1: Enhancing Tumor scores for the SVM classifier with the RBGLMR feature set and patient-specific training (including and excluding boundary pixels).

Patient	Timepoint	Score (including Boundary)	Score (excluding Boundary)
1	1	0.77	0.97
2	1	0.91	1.00
3	1	0.90	1.00
3	2	0.93	1.00
5	1	0.92	1.00
5	2	0.88	1.00
5	3	0.90	1.00
5	4	0.94	1.00
9	1	0.90	1.00
10	1	0.89	0.99
11	1	0.93	1.00
11	2	0.86	1.00
11	3	0.88	0.99
Average		0.89	1.00

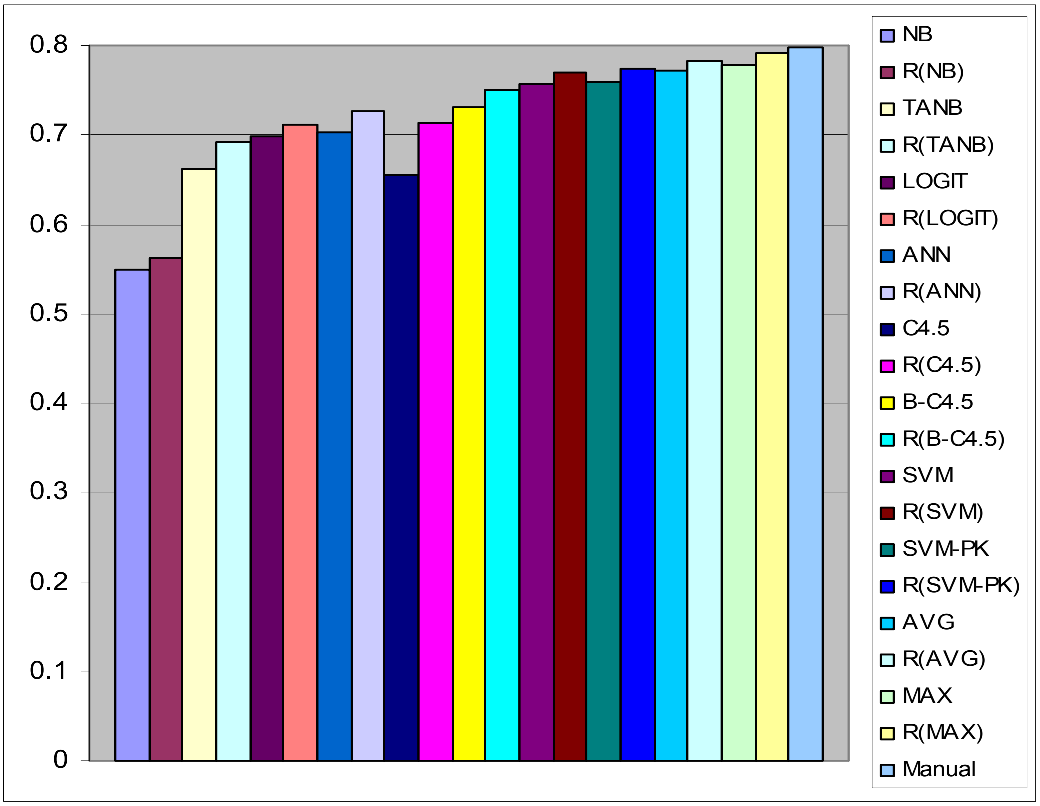


Figure 5.24: Average Gross Tumor scores for different classifiers with the RBGLMR feature set and patient-specific training before and after relaxation (over 17 images).

Table 5.2: Tumor and Edema area scores for the SVM classifier with the RBGLMR feature set and patient-specific training (including and excluding boundary pixels).

Patient	Timepoint	Score (including Boundary)	Score (excluding Boundary)
1	1	0.62	0.74
2	1	0.91	1.00
3	1	0.78	0.90
3	2	0.88	0.95
4	1	0.89	1.00
5	1	0.73	0.87
5	2	0.61	0.75
5	3	0.77	0.86
5	4	0.91	0.98
6	1	0.82	0.94
7	1	0.76	0.88
8	1	0.57	0.79
9	1	0.87	0.98
10	1	0.90	0.98
11	1	0.95	1.00
11	2	0.86	1.00
11	3	0.88	0.98
Average		0.81	0.92

Table 5.3: Gross Tumor scores for the SVM classifier with the RBGLMR feature set and patient-specific training (including and excluding boundary pixels).

Patient	Timepoint	Score (including Boundary)	Score (excluding Boundary)
1	1	0.79	0.95
2	1	0.80	0.93
3	1	0.90	0.98
3	2	0.85	0.93
4	1	0.88	0.99
5	1	0.69	0.82
5	2	0.76	0.86
5	3	0.81	0.91
5	4	0.85	0.93
6	1	0.68	0.82
7	1	0.59	0.70
8	1	0.36	0.46
9	1	0.89	0.98
10	1	0.87	0.98
11	1	0.83	0.95
11	2	0.80	0.99
11	3	0.82	0.95
Average		0.77	0.89

segmentations when ignoring mistakes made near boundaries). Although there was one case where the system performed poorly for Gross Tumor segmentation, the system tended to produce highly accurate segmentations of the Gross Tumor areas in the remaining cases.

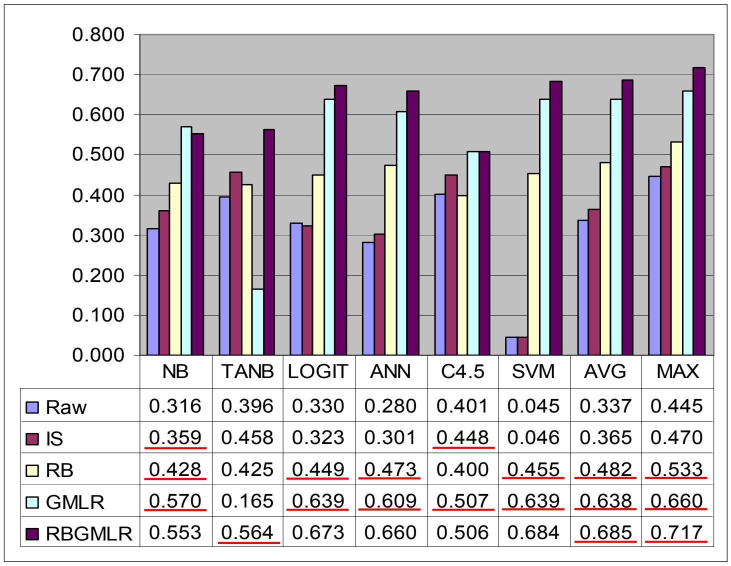
5.2 Inter-Patient Training Experiments

5.2.1 Inter-Patient Training Experiment 1

Question: *Are the trends observed in the patient-specific training scenario also observed in inter-patient scenarios*? Specifically, there are several trends that we would like to explore in the inter-patient experiments:

- *Does Noise Reduction consistently improve results*?
- *Can a set of pixel-level features representing diverse sources of information improve an intensity-based model*?
- *Can a set of region-level image-based feature improve an intensity-based model*?
- *Does combining a diverse set of pixel-level features with region-based features still not significantly outperform the use of either type of feature individually*?
- *Will the classifiers follow approximately the same order in terms of accuracy*?
- *Does relaxation of the predicted labels still increase accuracy*?

Experimental Outline: This experiment repeated a previous experiment examining different feature sets, but used inter-patient training rather than patient-specific training. The inter-patient training scenario involved training on 10 of the 11 patients, and testing on the remaining patient. This was repeated was 11 times in order to classify each patient. In order to reduce computational time, the normal class was sub-sampled by 50% in training, and pixels with a zero valued spatial



	NB	TANB	LOGIT	ANN	C4.5	SVM	AVG	MAX
■ Raw	0.316	0.396	0.330	0.280	0.401	0.045	0.337	0.445
■ IS	0.359	0.458	0.323	0.301	0.448	0.046	0.365	0.470
□ RB	0.428	0.425	0.449	0.473	0.400	0.455	0.482	0.533
□ GMLR	0.570	0.165	0.639	0.609	0.507	0.639	0.638	0.660
■ RBGMLR	0.553	0.564	0.673	0.660	0.506	0.684	0.685	0.717

Figure 5.25: Average Tumor and Edema scores for different classifiers and feature sets with inter-patient training (over 11 images). Values that are underlined are cases where a feature set significantly outperformed each of the other feature sets above it in the column for the same classifier.

probability of being part of the brain were not used in training. Among the classifiers used in previous experiments, several were not evaluated in this scenario. The 'Classification by Interpolation' algorithm was not used since it is no longer applicable for this scenario. The Boosted C4.5 and SVM with the Polynomial Kernel were also not evaluated since the computational time proved to be excessively large for this scenario. Since only the Linear Kernel was evaluated for the SVM, the primal formulation [Shawe-Taylor and Cristianini, 2004] was used instead of the dual formulation. This sped up both the training and testing phases considerably, since a relatively small feature set is used in comparison to the number of training examples (and Support Vectors of the learned model).

Predicted Results: We anticipate that the trends observed in the patient-specific scenario should carry over to the inter-patient scenario:

- Noise Reduction should produce a small increase in classification accuracy for an intensity-based model.
- The effects of the 17 pixel-level features compared to an intensity-based model will vary from hurting performance to providing a significant increase, depending on the classifier used.
- Region-based features should offer a major advantage over purely intensity-based models.
- Combining the pixel-level features with region-based features may result in a small increase in classification accuracy for some classifiers.
- Support Vector Machines are expected to again be the most effective individual classifier, while it is expected that Model Averaging could again provide similar or better results.
- We anticipate that performing relaxation of the labels will again improve the segmentation scores.

Experimental Results: The average scores for different feature sets and classifiers are shown in Figure 5.25. The results for different classifiers before and after relaxation with the RBGLMR feature set are shown in Figure 5.26.

Discussion: Many of the trends observed in the patient-specific training experiments clearly apply in the inter-patient training experiments. Intensity-based classification methods still performed

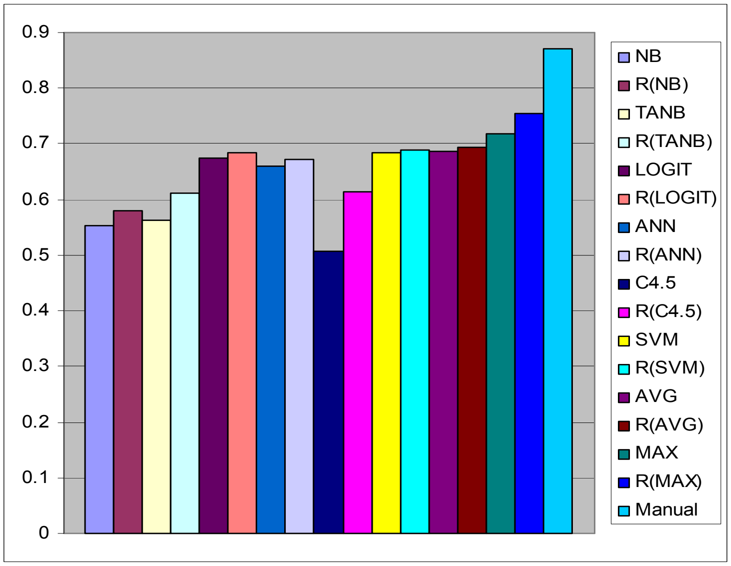


Figure 5.26: Average Tumor and Edema scores for different classifiers and feature sets with inter-patient training before and after relaxation (over 11 images).

similarly or slightly better after Noise Reduction. However, as opposed to the patient-case, the effect of Noise Reduction did not seem to be as significant. The 17 pixel-level features provided a significant advantage over an intensity-based model for all but the TANB and C4.5 classifiers. This is consistent with the patient-specific training experiments, although it is surprising that the Naive Bayes model benefited from these 17 features. The use of region-based features again resulted in (often large) significant performance increases, with the exception of the TANB model.

The combination of the pixel-level features and the region-based features resulted in performance improvements for all methods (although not always significant), except the Naive Bayes and C4.5 models, where this feature set resulted in similar scores to the image-based GLMR feature set. It was not clear from the patient-specific experiments that this would be the case. A possible explanation for why the additional features offered a more general advantage in the inter-patient case is that the intensities are inherently less reliable as features. This could indicate that coordinate-based and registration-based features (that do not rely solely on the intensities) are more important in inter-patient classification. It is noteworthy that since less data was used in the inter-patient experiment, it is likely that these differences would be significant for a slightly larger data set.

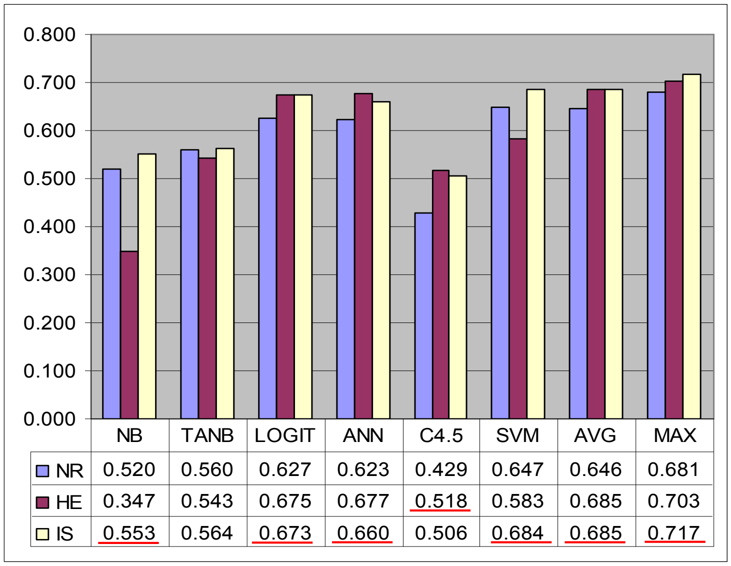
The SVM classifier outperformed each of the individual classifiers (although this was not significant in the case of Logistic Regression and the ANN), and there was no significant difference between the SVM and the Model Averaging technique. In this experiment, relaxation of the labels resulted in a significant score increase for every classifier except the SVM (the consistent score increases are likely a result of the overall decrease in classification accuracy).

For this inter-patient scenario, the Manual method significantly outperformed the other classifiers by a large margin. However, it should be noted that this was the Manual method from the patient-specific experiments. Thus, the manual method was guided by patient-specific labeled data which is likely producing an overestimation of the actual upper bound for this inter-patient scenario.

5.2.2 Inter-Patient Training Experiment 2

Question: *Has the intensity standardization method improved the classification accuracy*?

Experimental Outline: To explore this question, we compared the results obtained from data that underwent Intensity Standardized to data that was not Intensity Standardized. In addition, we included a purely histogram-based method of intensity standardization for comparison (the Histogram



	NB	TANB	LOGIT	ANN	C4.5	SVM	AVG	MAX
🟦 NR	0.520	0.560	0.627	0.623	0.429	0.647	0.646	0.681
🟪 HE	0.347	0.543	0.675	0.677	0.518	0.583	0.685	0.703
🟨 IS	0.553	0.564	0.673	0.660	0.506	0.684	0.685	0.717

Figure 5.27: Average Tumor and Edema scores for different classifiers and feature sets with inter-specific training (over 11 images).

Equalization method included with Matlab [MATLAB, Online]). All three data sets went through Noise Reduction, and the RBGLMR feature set was computed and used for classification.

Predicted Results: Due to the large intensity differences between images, it is expected that higher classification scores will be achieved with either Intensity Standardization method than without an Intensity Standardization step. In addition, we expect that our Intensity Standardization step will outperform Histogram Equalization, since Histogram Equalization will be sensitive to the presence and size of the abnormality.

Experimental Results: The average scores are shown in Figure 5.27.

Discussion: The Intensity Standardized data significantly outperformed the non-standardized data for each classifier, except the TANB and C4.5 models. The classification scores of the data that was histogram equalized was improved over standardized data in some cases, but in several cases the performance decreased and the difference was only significant for the C4.5 classifier. With the exception of the Naive Bayes classifier where Intensity Standardization was significantly better than Histogram Equalization, the Histogram Equalized and Intensity Standardized data sets were not significantly different. This might be explained by the fact that Histogram Equalization is a non-linear method, while our Intensity Standardization method assumes a linear mapping. Thus, sensitivity to the presence of abnormal tissues may be compensated for by a more expressive standardization of the data.

5.2.3 Inter-Patient Training Experiment 3

Question: *How accurate is the final system in each test case, and are the errors primarily made at boundaries*?

Experimental Outline and Predicted Results: This question was evaluated in the same way that it was evaluated in the patient-specific case, and we expected to see similar levels of improvement.

Experimental Results: The individual scores are shown in Table 5.4.

Discussion: The scores achieved when excluding boundary mistakes were much more varied in the inter-patient case. While for some patients the system performed poorly, in many cases the system was nearly as accurate as in the patient-specific training scenario.

Table 5.4: Tumor and Edema scores for the SVM classifier with the RBGLMR feature set and inter-specific training (including and excluding boundary pixels).

Patient	Timepoint	Score (including Boundary)	Score (excluding Boundary)
1	1	0.55	0.61
2	1	0.78	0.93
3	2	0.58	0.65
4	1	0.48	0.57
5	4	0.88	0.96
6	1	0.77	0.91
7	1	0.70	0.83
8	1	0.58	0.71
9	1	0.56	0.61
10	1	0.85	0.95
11	3	0.84	0.96
Average		0.69	0.79

Chapter 6

Conclusion

The previous chapters have introduced the problem of automatic brain tumor and edema segmentation in MR images. The study of this problem is practically motivated, but has properties that make it an interesting and challenging Machine Learning and Pattern Recognition task. Chapter 3 introduced a framework that combined ideas from the prior work into a general method to perform this task automatically, notably expanding on the ideas of coordinate-based and registration-based features. Chapter 4 then introduced a specific implementation of these ideas, which used existing state-of-the-art algorithms and several new methods. Chapter 5 evaluated the performance of this system in the simplified task of segmentation with patient-specific training, and in the extremely challenging task of performing segmentation with inter-patient training. This chapter will begin by discussing potential future directions of research that directly follow from this work. This will be followed by a summary of the innovations presented this work, and finally a summary of the contributions made in this thesis.

6.1 Future Directions

With respect to automatic brain tumor segmentation, there are several obvious future directions to explore. Improving the methods used in each step of the framework could improve the final results. Potential improvements for each step of the system were suggested in Chapter 4. Based on the experimental results, it is likely that significant performance improvements would be expected if improvements were made to the Intensity Standardization and the Relaxation steps. Even the simple Relaxation phase used in this work produced noticeable improvements, while the inter-patient results indicate that a non-linear Intensity Standardization step may be more appropriate. The experimental results also suggest that an automated feature selection algorithm might result in performance gains, since it is clear that combining feature sets does not always improve results. It is also likely that a more effective non-linear Spatial Registration method could improve results. This is based on visual analysis of the non-linear warping results, where it is clear that a slightly more effective registration could increase the utility of the registration-based features. Incorporating further prior knowledge through additional coordinate-based features, or registration-based features that register more than a single template also represent promising future directions.

In addition to improving individual steps in the implementation, modifications to the general framework could be explored. This could include the addition or removal of steps, or changes in the specific order. As an example, performing coregistration earlier in the framework would allow the use of multiple modalities in the Noise Reduction steps (although multiple modalities could also be used to enhance other steps such as template registration or feature extration). Combining steps is also a promising future direction. Obvious candidates include combining Classification and Relaxation, Template Registration and Intensity Standardization, Inhomogeneity Reduction and Intensity Standardization, or Inhomogeneity Reduction and Inter-Slice Intensity Variations Reduction.

With respect to inter-patient training, it is likely that larger training sets would improve results. Although the results are impressive given that 4 different tumor types were present in the 10 patient training set, it is likely that a larger training set (or a training set of the same size for a single tumor type) would result in higher classification accuracies. Another modification to the data used that could be explored is the use of other modalities, as in the techniques that use modalities where normal and abnormal tissue are more easily discriminated as discussed in Chapter 2 (including FLAIR and MR Spectroscopy images). Since relatively few assumptions are made about the underlying modalities, additional modalities or completely different sets of modalities can be used, provided that they satisfy the following two criteria:

- The modalities can be coregistered.
- There is a template in at least one of the modalities.

These two items are relatively easy to satisfy, given that entropy based measures such as Mutual Information allow the coregistration of a large variety of modalities, while templates in standard coordinate systems exist for several different modalities.

Related to the use of other modalities, is the potential to apply this system to different tasks. In Chapter 5, the system learned to perform 3 different types of segmentation tasks, simply by changing the labels of the training data. It is clear that the system could be used in related tasks such as segmenting Enhancing Tumor Pixels, Homogeneous Tumor Areas, or Heterogeneous Tumor Areas. But it is possible that the system could also be applied for tasks that are not directly related to tumors. This could include the segmentation of Multiple Sclerosis lesions, areas affected by Stroke, or other types of brain damage or lesions. The system could also be used to segment normal structures (where coordinate-based and registration-based priors would aid significantly), or to simultaneously segment pixels into more than two classes. As a final note, templates and standard coordinate systems exist (or can be created) for other areas of the body, and thus the system could be adapted for segmentation tasks in other areas of the body.

6.2 Innovations

The most notable innovations presented in this dissertation are as follows:

- We presented a new method to reduce inter-slice intensity variations. This method does not depend on a tissue model nor on a segmentation, and is still effective if the volume has anisotropic voxels.
- We presented a new template-based method to reduce inter-volume intensity variations. This was an extension of the inter-slice intensity variation method, and therefore does not depend on a tissue model nor on a segmentation. This method incorporates symmetry to confer robustness to abnormalities, and incorporates a ‘brain’ weighting to focus the estimation on areas that are part of the brain.
- We presented the most extensive preprocessing ‘pipeline’ for tumor segmentation to date. Notably, no previous system has incorporated both inter-slice intensity variation reduction and inter-volume intensity standardization.
- We presented the first comparative evaluation of different types of coordinate-based and registration-based features. We also presented several new types of coordinate-based and registration-based features. Expected spatial intensities and a characterization of bi-lateral symmetry were the most notable of the new types evaluated.
- We presented the first system that incorporates multiple types of coordinate-based and registration-based features. We showed that using multiple types of coordinate-based and/or registration-based features with an appropriate classifier can offer a significant performance improvement over using a single type of coordinate-based or registration-based feature.

- We presented the first system that uses Machine Learning to combine coordinate-based and registration-based features with more traditional textural features. We showed that this combination can be complimentary and offers a performance improvement over using textural features or using the combination of coordinate-based and registration-based features.

6.3 Contribution

This thesis has presented a framework and an implementation of this framework for automatic brain tumor and edema segmentation in MR images. This framework incorporates many of the important ideas proposed in the literature. This includes Local Noise Reduction, Inter-Slice Intensity Variation Reduction, Intensity Inhomogeneity Reduction, Inter-Modality Coregistration, Linear and Non-Linear Template Registration, Intensity Standardization, Textural Features, Registration-Based and Coordinate-Based Features, a Supervised Classification model, and finally Label Relaxation. For each of these steps, a state-of-the-art method was incorporated into the system to perform the task. In addition, a new method was introduced for inter-slice intensity variation reduction that does not rely on a tissue model. A related method was introduced for Intensity Standardization, which allows robust template-based estimation in the presence of large abnormalities.

This work showed that the *spatial prior probabilities* and *distance transform* introduced in previous works are part of the more general classes of *coordinate-based* and *registration-based* features introduced in this work, respectively. We explored the use of several more of these types of features, including a characterization of bi-lateral symmetry, utilizing the template intensity information, employing an expected intensity map, and incorporating a spatial probability for the brain area. This latter feature allowed the error-prone *skull-stripping* step used in many recent systems to be completely circumvented. In addition to the implemented features, we presented several other potential coordinate-based and registration-based features, including anatomic variability maps, features derived from the non-linear deformation field, and features derived from the registration of multiple templates. Finally, we showed that these coordinate-based and registration-based features could not only be combined with traditional image-based features, but region-based measures such as textural features could also be computed based on the registration-based and coordinate-based features.

In our patient-specific training experiments where training was performed on distant slices within the volume to be segmented, we presented nearly perfect results for the segmentation of the Enhancing Tumor area. In addition, the patient-specific training experiments indicate that the system also performs very accurate segmentation of the Tumor and Edema area, and the Gross Tumor area. The results presented for these 3 task could potentially lead to immediate practical use, since segmentation with patient-specific training would greatly reduce the time needed to segment full three-dimensional volumes. We evaluated the system in the challenging task of segmenting Tumor and Edema areas based on inter-patient training from patients with different types of tumors. These results were competitive with the state-of-the-art system presented in [Prastawa et al., 2004], although our results were validated on a significantly larger data set, that contained more types of tumors and came from patients at different stages of treatment.

We presented a variety of directions for future development within this framework. The exploration of these directions, or even of larger training sets, would likely increase the accuracy of the system. Finally, the presented framework was purposely designed to be able to easily incorporate more advanced imaging modalities and to easily adapt to related tasks.

Glossary

- **Active Contour**: A boundary that adaptively adjusts itself based on the image data.
- **ADF**: Anisotropic Diffusion Filter, a smoothing technique that smoothes images inversely proportional to the local image gradient, resulting in edge-preserving smoothing.
- **Anisotropic Pixels**: Pixels that do not have the same thickness in all dimensions.
- **ANN**: Artificial Neural Network, a learning method consisting of a set of nodes, where each node may apply a mathematical operation to its inputs, and there exists (adjustable) weights along the connections between nodes.
- **Axial**: Orientation of an MRI scan, an axial slice will be parallel to the feet of the patient with the left hand side of the scan being the right hand side of the patient.
- **Bayes Theorum**: A method to optimally calculate conditional probabilites, given prior probabilities.
- β**-Spline**: A piecewise polynomial function that can be recursively defined.
- **Bi-Lateral Symmetry**: ‘Left to right’ symmetry.
- **Bias Field**: See *Inhomogeneity Field*.
- **Boosting**: A model averaging method that combines a series of classifiers to potentially achieve higher accuracy than each classifier would individually.
- **Brain Masking**: The segmentation of the brain from surrounding tissues.
- **C4.5**: A popular method for learning decision tree classifiers.
- **Classification**: The task of assigning a class (from a finite set) to examples based on a set of measured features.
- **Colin27**: An average image of a single individual imaged 27 times and registered to the same coordinate system.
- **Coordinate-Based Features**: Features based on a standard coordinate system, including aggregations over multiple individuals registered to the coordinate system.
- **Conditional Probability**: The probability of an event, given the known information.
- **Conditional Random Field**: A formulation of Markov Random Fields that employs conditional probabilities.
- **Contrast Agent**: A substance used to enhance visualization of structures with specific anatomic properties.
- **Contrast Agent Difference Image**: The difference in intensities between an image before and after the addition of a contrast agent.
- **Coregistration**: The alignment of volumes of different modalities of the same individual.
- **Coronal**: Orientation of an MRI scan, a coronal slice will be perpendicular to the feet and parallel to the shoulder line of the patient with the left hand side of the scan being the right hand side of the patient.
- **CSF**: Cerebrospinal Fluid, normal fluid present within the brian.
- **CT**: Computer Tomography, an X-Ray based method for producing three-dimensional volumes.
- **Decision Tree**: A rooted graph where each node contains a decision. Classification can be done using Decision Trees by proceeding from the root node to a leaf node that will contain a class label.

- **Deformation Field**: A field that measures how far each pixel was warped from its original position during non-linear warping.
- **Dilation**: A morphological operation that grows binary structures.
- **Dimensionality Reduction**: The processing of features in order to find a lower-dimensional representation that encodes the same information.
- **Distance Transform**: The calculation of, for each pixel in an image, its distance to a binary structure present within the image.
- **Edema**: Swelling (excess water).
- **Edge Detection**: Processing of an image to highlight/detect edges.
- **EM**: Expectation Maximization, a general approach to learning with hidden variables.
- **Enhancing**: Regions that appear hyper-intense in an image (typically this refers to regions that are hyper-intense after the injection of a contrast agent).
- **Ensemble Methods**: Learning methods that employ multiple classifiers.
- **Entropy**: An information content measure that considers the likelihoods of individual events occuring.
- **Erosion**: A morphological operation that shrinks binary structures.
- **Extrinsic Markers**: Explicit physical markers used during imaging to allow straightforward image registration.
- **Feature Selection**: The process of re-weighting or selecting features such that only the most relevant subset of a feature set is used.
- **Fiducial Markers**: See *extrinsic markers*.
- **Filter Bank**: A set of (linear) filters whose individual outputs can be used as features describing image texture.
- **First-Order Textures**: See *statistical moments*.
- **FLAIR**: Fluid Attenuated Inversion Recovery, an MR imaging technique that produces images similar to T2-weighted images, but with free water (ie. normal CSF) suppressed.
- **fMRI**: Functional Magnetic Resonance Imaging, a technique for assesssing activation of different parts of the brain.
- **Gabor Filter**: Linear filters used for assessing oriented textural informaiton.
- **Gain Field**: See *Inhomogeneity Field*.
- **Gaussian Distribution**: Synonym for ‘normal distribution’, a parametric distribution characterized by its mean and standard deviations.
- **Gaussian Cube**: A multi-scale image representation obtained by filtering images with Gaussian Filters having different standard deviations.
- **Gaussian Filter**: A linear filter whose values form a Gaussian distribution. Typically, the sum of the values is constrained to be 1 and the distribution is centered at the center of the filter.
- **Gaussian Pyramid**: A multi-scale image representation obtained by recursive Gaussian filtering, and reductions in image size between filterings.
- **GTV**: Gross Tumor Volume, the abnormal tumor region visible in the image.
- **Haralick Features**: See *spatial coocurrence features*.
- **ICBM**: International Consortium for Brain Mapping A project whose goal is ‘the continuing development of a probabilistic reference system for the human brain’. http://www.loni.ucla.edu/ICBM/
- **ICBM152**: A data set of 152 spatially registered normal individuals used in the construction of templates and prior probabilities.
- **ICM**: Iterated Condition Modes, a method of inference in Random Fields.
- **Image-Based Features**: Features that take into account properties of the image being examined.
- **Inhomogeneity Field**: A field that varies spatially over an image and that describes the deviation at each pixel from its uncorrupted value.
- **INSECT**: Intensity Normalized Stereotaxic Environment for the Classification of Tissue, an image processing pipeline for Multiple Sclerosis lesion segmentation.
- **Intensity Inhomogeneity Reduction**: Processing of an image to reduce the intensity vari-

ance between pixels representing the same tissue type, while preserving differences in the intensities of pixels representing different tissue types.

- **Intensity Standardization**: Processing of images in order to standardize their intensities between images. The reduces the variability in the intensities of similar tissue types between images of different individuals.
- **Inter-Slice Intensity Variation Reduction**: Processing of images in order to reduce the effects of intensity offsets in individual slices.
- **Inter-Patient Training**: Utilizing training data from multiple patients, with the goal of applying the learned model to new patients.
- **Isotropic Pixels**: Pixels that have the same thickness along each dimension.
- **Jaccard Measure**: A similarity measurement between two sets.
- **Knowledge-Based Approaches**: A segmentation strategy that consists of manually engineering rules and/or processing steps that will lead to a segmentation.
- **KNN**: K-Nearest Neighbors, a classification method that assigns instances to the class label of the k closest training instances in the feature space.
- **Least Squares**: A regression method that minimizes the sum of the squared distance from the model to the training data.
- **Level Set**: An active contour method that is convenient for modeling three-dimensional objects.
- **Laplacian**: A rotation invariant approximation of the local image second derivative.
- **Laplacian Cube**: A multi-scale feature representation obtained using Laplacian of Gaussian filtering. The finest scale is the raw image data.
- **Logistic Regression**: A modification of linear regression to output values in the range [0,1], and to enforce that the values over the different classes sum to 1.
- **Markov Random Field**: A statistical model that takes into account dependencies in the labels of neighboring instances, in addition to dependencies between features and labels.
- **MAP**: *Maximum a posteriori*, MAP estimation involves calculating the parameters that maximize the likelihood of the data occuring, given the model chosen and a prior probability over the model parameters.
- **Mean Field Approximation**: A method of inference in Markov Random Fields.
- **Meta-Learning**: See *ensemble methods*.
- **Mid-Saggital Plane**: The two-dimensional axis of bi-lateral symmetry.
- **Mixture Model**: A distribution constructured from multiple (often Gaussian) distributions.
- **Model Averaging**: A learning strategy where multiple learned models are averaged to potentially produce more accurate results.
- **ML**: Maximum Likelihood, ML estimation involves calculating the parameters that maximize the likelihood of the data occuring, given the model chosen.
- **MNI**: Montreal Neurological Institute and Hospital, an institute at McGill University studying the nervous system.
- **MNI305**: A data set of 305 spatially registered normal individuals used in the construction of templates and prior probabilities.
- **Modality**: An imaging medium. For example, T1-weighted MR images, or CT images.
- **Morphological Operation**: Simple operations applied to binary images, based on comparing pixel's values to the values of their neighbors.
- **MR**: Magnetic Resonance, the physical property being measured in MRI.
- **MR8**: A (relatively) rotation invariant version of the multi-scale and (relatively) illumination invariant Maximum Response filter bank, consisting of 8 features.
- **MRI**: Magnetic Resonance Imageing, a technique to visualize parts of the human body based on water/fat content.
- **MRS**: Magnetic Resonance Spectroscopy, an advanced MR technique that allows the identification of different chemical compounds.
- **Mutual Information**: An entropy-based measurement of the combined information content

of two sources relative to their individual information content.

- **Naive Bayes**: A classification method that determines the optimal (Maximum Likelihood) parameters of a model that assumes independence of the features, given the class label.
- **N3**: Nonparametric Nonuniform intensity Normalization, an effective technique to perform intensity inhomogeneity reduction without a tissue model or a segmentation.
- **Necrotic**: Region of dead cells, typically observed at the center of highly aggressive tumors.
- **Parametric Distribution**: A probability distribution described by a finite (typically small) number of parameters.
- **Partial Volume Averaging**: The averaging effect observed at pixels representing more than one type of tissue.
- **Patient-Specific Training**: Utilizing training data from a single patient, with the goal of applying the learned model only to that patient.
- **PET**: Positron Emission Tomography, a technique to visualize the uptake of a radioactive agent.
- **Quadratic Programming**: The minimization/maximization of an expression (subject to linear constraints) that can contain quadratic, linear, and constant terms.
- **Registration-Based Features**: Features computed based on properties of one or more aligned template images.
- **Regularization**: The use of smoothness constraints, or prior knowledge about expected parameter values, during paramter estimation.
- **Relaxation**: The use of neighboring pixel labels (and potentially the image information) to refine the labels of individual pixels.
- ρ**-weighted**: An MRI modality.
- **Sagittal**: Orientation of an MRI scan, a sagittal slice will be perpendicular to the feet and perpendicular to the shoulder line of the patient.
- **Second-Order Textures**: See *spatial coocurrence features*.
- **Segmentation**: The division of an object into multiple segments.
- **Series**: A set of slices taken using the same MRI protocol during the same acquisition study (ie. a set of T1-weighted axial slices), often changing position along one axis.
- **Skull Stripping**: See *brain masking*.
- **Slice**: An orthogonal view of the body part being visualized by the MRI.
- **Spatial Coocurrence Features**: Features that are computed based on a spatial coocurrence matrix, which measures the likelihood of pixels with intensities i and j being observed at distance d and angle θ.
- **Spatial Interpolation**: The estimation of unknown values between known values, based on the surrounding known values.
- **Spatial Prior Probabilities**: A likelihood, measured for each pixel in a coordinate system, that the pixel belongs to a specific class. Typically generated by registering multiple individuals to a standard coordinate system and labeling relevant classes, or by smoothing the labels of an individual image.
- **Spatial Registration**: The spatial alignment of one or more images.
- **SPM**: Statistical Parametric Mapping, a collection of Matlab scripts developed at the University College of London for the purpose of statistical analyis of fMRI scans.
- **SPM2**: The most recent version of the SPM software.
- **Statistical Moment Features**: Statistics that characterize properties of the local intensity distribtuion.
- **Study**: A collection of MRI series of a patient taken at the same time.
- **Supervised Learning**: A framework that employs a set of measured features and labeled training examples to learn a model that maps from the values of the features to the labels. This model is often then used to assign labels to unlabeled examples.
- **SUSAN**: Smallest Univalue Segment Assimilating Nucleus, a method used to assess relatedness between pixels in an image.

- **SVM**: Support Vector Machine, an approach to classification that seeks to maximize the margin between two classes.
- **T1-weighted**: An MR image that highlights fat locations.
- **T2-weighted**: An MR image that highlights water locations.
- **TANB**: Tree-Augmented Naive Bayes, a technique that augments a Naive Bayes model to allow dependencies between the features.
- **Template Registration**: The spatial alignment of an image with a template image.
- **Transverse**: A synonym for axial.
- **Voxel**: Volume element, a three-dimensional analog of a picture element (pixel).
- **Weighted Least Squares**: A formulation of the Least Squares method that weights the error associated with individual training instances.

Bibliography

[Alirezaie et al., 1997] Alirezaie, J., Jernigan, M. E., and Nahmias, C. (1997). Neural network based segmentation of magnetic resonance images of the brain. *IEEE Transactions on Nuclear Science*, 44(2).

[Arnold et al., 2001] Arnold, J., Liows, J., Schaper, K., Stern, J., Sled, J., Shattuck, D., Worth, A., Cohen, M., Leahy, R., Mazziotta, J., and Rottenberg, D. (2001). Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. *Neuroimage*, 13(5):931–943.

[Ashburner, 2002] Ashburner, J. (2002). Another mri bias correction approach. In *8th International Conference on Functional Mapping of the Human Brain, Sendai, Japan*.

[Ashburner and Friston, 1999] Ashburner, J. and Friston, K. (1999). Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4):254–266.

[Ashburner and Friston, 2003a] Ashburner, J. and Friston, K. (2003a). Morphometry. In Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., and Penny, W., editors, *Human Brain Function*, chapter 6. Academic Press, 2nd edition.

[Ashburner and Friston, 2003b] Ashburner, J. and Friston, K. (2003b). Rigid body registration. In Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., and Penny, W., editors, *Human Brain Function*, chapter 2. Academic Press, 2nd edition.

[Ashburner et al., 1997] Ashburner, J., Neelin, P., Collins, D., Evans, A., and Friston, K. (1997). Incorporating prior knowledge into image registration. *NeuroImage*, 6(4):344–352.

[Barla et al., 2002] Barla, A., Odone, F., and Verri, A. (2002). Hausdorff kernel for 3d object acquisition and detection. In *European Conference on Computer Vision*.

[BrainWeb, Online] BrainWeb (Online). Brainweb: a www interface to a simulated brain database (sbd) and custom mri simulations, http://www.bic.mni.mcgill.ca/brainweb/.

[Brown and Semeka, 2003] Brown, M. and Semeka, R. (2003). *MRI: Basic Principles and Applications*. John Wiley and Sons, Inc., 3rd edition.

[Busch, 1997] Busch, C. (1997). Wavelet based texture segmentation of multi-modal tomographic images. *Computer and Graphics*, 21(3):347–358.

[Capelle et al., 2000] Capelle, A., Alata, O., Fernandez, C., Lefevre, S., and Ferrie, J. (2000). Unsupervised segmentation for automatic detection of brain tumors in mri. In *IEEE International Conference on Image Processing*, pages 613–616.

[Capelle et al., 2004] Capelle, A., Colot, O., and Fernandez-Maloigne, C. (2004). Evidential segmentation scheme of multi-echo MR images for the detection of brain tumors using neighborhood information. *Information Fusion*, 5(3):203–216.

[Carr et al., 2001] Carr, J., Beatson, R., Cherrie, J., Mitchell, T., Fright, W., McCallum, B., and Evans, T. (2001). Reconstruction and representation of 3d objects with radial basis functions. In *ACM SIGGRAPH*, pages 67–76.

[Carr et al., 1997] Carr, J., Fright, W., and Beatson, R. (1997). Surface interpolation with radial basis functions for medical imaging. *IEEE Transactions on Medical Imaging*, 16(1):96–107.

[Choi et al., 1991] Choi, H., Haynor, D., and Kim, Y. (1991). Partial volume tissue classification of multichannel magnetic resonance images-a mixel model. *IEEE Transactions on Medical Imaging*, 10(3):395–407.

[Christenson, 2003] Christenson, J. (2003). Normalization of brain magnetic resonance images using histogram even-order derivative analysis. *Magn Reson Imaging*, 21(7):817–820.

[Clark et al., 1998] Clark, M., Hall, L., Goldgof, D., Velthuizen, R., Murtagh, F., and Silbiger, M.

(1998). Automatic tumor segmentation using knowledge- based techniques. *IEEE Transactions on Medical Imaging*, 17:238–251.

[Clarke, 1991] Clarke, L. (1991). Mr image segmentation using mlm and artificial neural nets. *Medical Physics*, 18(3):673.

[Clatz et al., 2004] Clatz, O., Bondiau, P.-Y., Delingette, H., Sermesant, M., Warfield, S., Malandain, G., and Ayache, N. (2004). Brain tumor growth simulation. Technical report, INRIA.

[Cocosco et al., 1997] Cocosco, C., Kollokian, V., Kwan, R.-S., and Evans, A. (1997). Brainweb: Online interface to a 3d mri simulated brain database. *NeuroImage*, 5(S425).

[Collignon, 1998] Collignon, A. (1998). *Multi-modality medical image registration by maximization of mutual information*. PhD thesis, Catholic Univ. Leuven.

[Collignon et al., 1995] Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Sutens, P., and Marchal, G. (1995). Automated multimodality image registration using information theory. In Bizais, Y. and Barillot, C., editors, *Information Processing in Medical Imaging*, pages 263–274. Kluwer Academic Publishers, Dordrecht.

[Collins et al., 1994] Collins, D., Neelin, P., Peters, T., and Evans, A. (1994). Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. *J Comput Assist Tomogr*, 18(2):192–205.

[Collins et al., 1998] Collins, D., Zijdenbos, A., Kollokian, V., Sled, J., Kabani, N., Holmes, C., and Evans, A. (1998). Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3):463–468.

[Collowet et al., 2004] Collowet, G., Strzelecki, M., and Mariette, F. (2004). Influence of mri acquisition protocols and image intensity normalization methods on texture classification. *Magn Reson Imaging*, 22(1):81–91.

[Dickson and Thomas, 1997] Dickson, S. and Thomas, B. (1997). Using neural networks to automatically detect brain tumours in MR images. *International Journal of Neural Systems*, 4(1):91–99.

[Evans and Collins, 1993] Evans, A. and Collins, D. (1993). A 305-member mri-based stereotactic atlas for cbf activation studies. In *40th Annual Meeting of the Soceity for Nuclear Medicine*.

[Evans et al., 1992a] Evans, A., Collins, D., and Milner, B. (1992a). An mri-based stereotactic atlas from 250 young normal subjects. In *Society for Neuroscience Abstracts*, volume 18. Abstract no. 179.4, page 408.

[Evans et al., 1992b] Evans, A., Marrett, S., Neelin, P., Collins, L., Worsley, K., Dai, W., Milot, S., Meyer, E., and Bub, D. (1992b). Anatomical mapping of functional activation in stereotactic coordinate space. *Neuroimage*, 1(1):43–53.

[Fletcher-Heath et al., 2001] Fletcher-Heath, L., Hall, L., Goldgof, D., and Murtagh, F. R. (2001). Automatic segmentation of non-enhancing brain tumors in magnetic resonance images. *Artificial Intelligence in Medicine*, 21:43–63.

[Forsyth and Ponce, 2003] Forsyth, D. and Ponce, J. (2003). *Computer Vision: A Modern Approach*. Prentice-Hall.

[Friston et al., 1995] Friston, K., Ashburner, J., Fristh, C., Poline, J., Heather, J., and Frackowiak, R. (1995). Spatial registration and normalization of images. *Human Brain Mapping*, 2:165–189.

[Garcia and Moreno, 2004] Garcia, C. and Moreno, J. (2004). Kernel based method for segmentation and modeling of magnetic resonance images. *Lecture Notes in Computer Science*, 3315:636–645.

[Gerig et al., 1992] Gerig, G., Kubler, O., Kikinis, R., and Jolesz, F. (1992). Nonlinear anisotropic filtering of mri data. *IEEE Transactions on Medical Imaging*, 11(2):221–232.

[Gering, 2003a] Gering, D. (2003a). Diagonalized nearest neighbor pattern matching for brain tumor segmentation. *R.E. Ellis, T.M. Peters (eds), Medical Image Computing and Computer-Assisted Intervention*.

[Gering, 2003b] Gering, D. (2003b). *Recognizing Deviations from Normalcy for Brain Tumor Segmentation*. PhD thesis, MIT.

[Gibbs et al., 1996] Gibbs, P., Buckley, D., Blackb, S., and Horsman, A. (1996). Tumour volume determination from MR images by morphological segmentation. *Physics in Medicine and Biology*, 41:2437–2446.

[Gispert et al., 2004] Gispert, J., Reig, S., Pascau, J., Vaquero, J., Garcia-Barreno, P., and Desco,

M. (2004). Method for bias field correction of brain t1-weighted magnetic resonance images minimizing segmentation error. *Hum Brain Mapp.*, 22(2):133–144.

[Gispert et al., 2003] Gispert, J., Reig, S., Pascau, J., Vaquero, M., and Desco, M. (2003). Inhomogeneity correction of magnetic resonance images by minimization of intensity overlapping. In *International Conference on Image Processing*, volume 2, pages 14–17.

[Gosche et al., 1999] Gosche, K., Velthuizen, R., Murtagh, F., Arrington, J., Gross, W., Mortimer, J., and Clarke, L. (1999). Automated quantification of brain magnetic resonance image hyperintensisites using hybrid clustering and knowledge-based methods. *International Journal of Imaging Systems and Technology*, 10(3):287–293.

[Haralick et al., 1973] Haralick, R., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classificaiton. *IEEE Trans. on Systems Man and Cybern.*, SMC-3(6):610–621.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning: data mining, inference and prediction*. Springer-Verlag.

[Hayman et al., 2004] Hayman, E., Caputo, B., Fritz, M., and Eklundh, J.-O. (2004). On the significance of real-world conditions for material classificaiton. In *8th ECCV*.

[Hellier et al., 2002] Hellier, P., Ashburner, J., Corouge, I., Barillot, C., and Friston, K. (2002). Inter subject registration of functional and anatomical data using spm. In *Medical Image Computing and Computer Assisted Intervention*, volume 587-590.

[Hellier et al., 2001] Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Goualher, G. L., Collins, L., Evans, A., Malandin, G., and Ayache, N. (2001). Retrospective evaluation of inter-subject brain registration. In Viergever, M.-A., Dohi, T., and Vannier, M., editors, *Medical Image Computing and Computer-Assisted Intervention*, volume 2208, pages 258–265.

[Herbrich et al., 2001] Herbrich, R., Graepel, T., and Campbell, C. (2001). Bayes point machines. *Journal of Machine Learning Research*, 1:245–279.

[Herlidou-Meme et al., 2003] Herlidou-Meme, S., Constans, J., Carsin, B., Olivie, D., Eliat, P., Nadal-Desbarats, L., Gondry, C., Rumeur, E. L., Idy-Peretti, I., and de Certaines, J. (2003). Mri texture analysis on texture test objects, normal brain and intracranial tumors. *Magnetic Resonance Imaging*, 21(9):989–993.

[Ho et al., 2002] Ho, S., Bullitt, E., and Gerig, G. (2002). Level set evolution with region competition: automatic 3D segmentation of brain tumors. *In 16th International Conference on Pattern Recognition*, pages 532–535.

[Holmes et al., 1998] Holmes, C., Hoge, R., Collins, L., Woods, R., Toga, A., and Evans, A. (1998). Enhancement of mr images using registration for signal averaging. *J Comput Assist Tomogr*, 22(2):324–333.

[ICBM View, Online] ICBM View (Online). Icbm view: an interactive web visualization tool for stereotaxic data from the icbm and other projects, http://www.bic.mni.mcgill.ca/icbmview/.

[Joachims, 1999] Joachims, T. (1999). Making large-scale svm learning practical. In Scholkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

[Joshi et al., 2003] Joshi, S., Lorenzen, P., Gerig, G., and Bullitt, E. (2003). Structural and radiometric asymmetry in brain images. *Med Image Anal.*, 7(2):155–70.

[Just and Thelen, 1988] Just, M. and Thelen, M. (1988). Tissue characterization with T1, T2 and proton density values: Results in 160 patients with brain tumors. *Radiology*, 169:779–785.

[Kaus et al., 2001] Kaus, M., Warfield, S., Nabavi, A., Black, P., Jolesz, F., and Kikinis, R. (2001). Automated segmentation of MR images of brain tumors. *Radiology*, 218:586–591.

[Kjaer et al., 1995] Kjaer, L., Ring, P., Thomsen, C., and Henriksen, O. (1995). Texture analysis in quantitative mr imaging. tissue characterisation of normal brain and intracranial tumours at 1.5 t. *Acta Radiol*, 36(2):127–135.

[Kumar and Hebert, 2003] Kumar, S. and Hebert, M. (2003). Discriminative random fields: A discriminative framework for contextual interaction in classification. In *IEEE Conf. on Computer Vision and Pattern Recognition*.

[Kwan et al., 1996] Kwan, R.-S., Evans, A., and Pike, G. (1996). An extensible mri simulator for post-processing evaluation. *Lecture Notes in Computer Science*, 1131(11):135–140.

[Kwan et al., 1999] Kwan, R.-S., Evans, A., and Pike, G. (1999). Mri simulation-based evaluation of image-processing and classification methods. *IEEE Transactions on Medical Imaging*, 18(11):1085–1097.

[Lafferty et al., 2001] Lafferty, J., Pereira, F., and McCallum, A. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.

[Lee et al., 1997] Lee, S., Wolberg, G., and Shin, S. (1997). Scattered data interpolation with multilevel β-splines. *IEEE Transactions on Visualization and Computer Graphics*, 3(3):228–244.

[Leemput et al., 1999a] Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999a). Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908.

[Leemput et al., 1999b] Leemput, K., Mase, F., Vandermeulen, D., and Suentens, P. (1999b). Automated model-based bias field correction of mr images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896.

[Leung and Malik, 2001] Leung, T. and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *Interational Journal of Computer Vision*, 43(1):29–44.

[Leung et al., 2001] Leung, W., Bones, P., and Lane, R. (2001). Statistical interpolation of sampled images. *Optical Engineering*, 40(4):547–553.

[Likar et al., 2001] Likar, B., Viergever, M., and Pernus, F. (2001). Retrospective correction of mr intensity inhomogeneity by information minimization. *IEEE Transactions on Medical Imaging*, 20(12):1398–1410.

[Ling and Bovik, 2002] Ling, H. and Bovik, A. (2002). Smoothing low-snr molecular images via anisotropic median-diffusion. *IEEE Transactions on Medical Imaging*, 21(4):377–384.

[Liu et al., 2001] Liu, Y., Collins, R., and Rothfus, W. (2001). Robust midsaggital plane extraction from normal and pathological 3d neuroradiology images. *IEEE Transactions on Medical Imaging*, 20(3):175–192.

[Madabhushi and Udupa, 2002] Madabhushi, A. and Udupa, J. (2002). Evaluating intensity standardization and inhomogeneity correction in magnetic resonance images. In *IEEE 28th Annual Northeast Bioengineering Conference*, pages 137–138.

[Mahmoud-Ghoenim et al., 2003] Mahmoud-Ghoenim, D., Toussaint, G., Constans, J.-M., and de Certaines, J. (2003). Three dimensional texture analysis in mri: a preliminary evaluation in gliomas. *Magnetic Resonance Imaging*, 21(9):983–987.

[Maintz and Viergever, 1998] Maintz, J. and Viergever, M. (1998). An overview of medical image registration methods. *Medical Image Analysis*, 2:1–36.

[Mangin, 2000] Mangin, J.-F. (2000). Entropy minimization for automatic correction of intensity nonuniformity. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 162–169.

[Materka and Strzelecki, 1998] Materka, A. and Strzelecki, M. (1998). Texture analysis methods - a review. Technical report, COST B11 Technical Reports, Brussels.

[MATLAB, Online] MATLAB (Online). Matlab - the language of technical computing, http://www.mathworks.com/products/matlab/.

[Mazzara et al., 2004] Mazzara, G., Velthuizen, R., Pearlman, J., Greenberg, H., and Wagner, H. (2004). Brain tumor target volume determination for radiation treatment planning through automated mri segmentation. *International Journal of Radiation Oncology*Biology*Physics*, 59(1):300–312.

[Mazziotta et al., 2001] Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Goualher, G. L., Boomsma, D., Cannon, T., Kawashima, R., and Mazoyer, B. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). *Philosophical Transactions: Biological Sciences*, 356(1412):1293–1322.

[McClain et al., 1995] McClain, K., Zhu, Y., and Hazle, J. (1995). Selection of MR images for automated segmentation. *Journal of Magnetic Resonanse Imaging*, 5(5):485–492.

[Meijering, 2002] Meijering, E. (2002). A chronology of interpolation: From ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, 90(3):319–342.

[Miller et al., 1981] Miller, A., Hoogstraten, B., Staquet, M., and Winkler, M. (1981). Reporting results of cancer treatment. *Cancer*, 47:207–214.

[MIPAV, Online] MIPAV (Online). Medical image processing, analysis and visualization, http://mipav.cit.nih.gov/.

[Moler, 2002] Moler, C. (2002). Numerical computing with matlab. http://www.mathworks.com/moler/.

[Moon et al., 2002] Moon, N., Bullitt, E., Leemput, K., and Gerig, G. (2002). *Automatic Brain and Tumor Segmentation*, pages 372–379. T. Dohi, R. Kikinis, eds. Medical Image Computing and Computer-Assisted Intervention. Springer, Tokyo, Japan.

[Murray, 2003] Murray, J. (2003). *Mathematical Biology II: Spatial Models and Biomedical Applications*. Springer-Verlag, 3rd edition.

[Muzzolini et al., 1998] Muzzolini, R., Yang, Y., and Pierson, R. (1998). Classifier design with incomplete knowledge. *Pattern Recognition*, 31(4):345–369.

[Nyul and Udupa, 1999] Nyul, L. and Udupa, J. (1999). On standardizing the mr image intensity scale. *Magn Reson Med*, 42(6):1072–1081.

[Nyul et al., 2000] Nyul, L., Udupa, J., and Zhang, X. (2000). New variants of a method of mri scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150.

[O'Donnell, 2001] O'Donnell, L. (2001). Semi-automatic medical image segmentation. Master's thesis, MIT.

[Otsu, 1979] Otsu, N. (1979). A threshold selection method from gray-level historgrams. *IEEE Trans. Systems, Man adn Cybernetics*, 9(1):62–66.

[Ozkan et al., 1993] Ozkan, M., Dawant, B., and Maciunas, R. (1993). Neural-network-based segmentation of multi-modal medical images: a comparative and prospective study. *IEEE Transactions on Medical Imaging*, 12(3):534–544.

[Peck et al., 2001] Peck, D., Hearshen, D., Soltanian-Zadeh, H., Scarpace, L., Dodge, C., and Mikkelsen, T. (2001). Segmentation of brain tumor boundaries using pattern recognition of magnetic resonance spectroscopy. In *RSNA*.

[Perona and Malik, 1990] Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639.

[Pirzkall et al., 2001] Pirzkall, A., McKnight, T., Graves, E., Carol, M., Sneed, P., Wara, W., Nelson, S., Verhey, L., and Larson, D. (2001). Mr-spectroscopy guided target delineation for high-grade gliomas. *International Journal of Radiation Oncology*Biology*Physics*, 50(4):915–928.

[Platt, 1999] Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In Scholkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 185–208. MIT Press.

[Platt, 2000] Platt, J. (2000). *Probabilistic outputs for support vector machines and comparison to regularized likelihood methods*. MIT Press, Cambridge, MA.

[Pluim et al., 2003] Pluim, J., Maintz, J., and Viergever, M. (2003). Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004.

[Prastawa et al., 2004] Prastawa, M., Bullitt, E., Ho, S., and Gerig, G. (2004). A brain tumor segmentation framework based on outlier detection. *Medical Image Analysis*, 8(3):275–283.

[Prastawa et al., 2003] Prastawa, M., Bullitt, E., Moon, N., Leemput, K., and Gerig, G. (2003). Automatic brain tumor segmentation by subject specific modification of atlas priors. *Academic Radiology*, 10(12):1341–1348.

[Press et al., 1998] Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1998). *Numerical Recipes in C: the art of scientific computing*. Cambridge University Press.

[Price et al., 2004] Price, S., Pena, A., Burnet, N., Jena, R., Green, H., Carpenter, T., Pickard, J., and Gillard, J. (2004). Tissue signature characterization of diffusion tensor abnormalities in cerebral gliomas. In *Workshop on Advances in Experimental and Clinical MR in Cancer Research*.

[Quinlan, 1993] Quinlan, J. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

[Randen and Husoy, 1999] Randen, T. and Husoy, J. (1999). Filtering for texture classification: a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):291–310.

[Russell and Norvig, 2002] Russell, S. and Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition.

[Sammouda et al., 1996] Sammouda, R., Niki, N., and Nishitani, H. (1996). A comparison of hopfield neural network and boltzmann machine in segmenting mr images of the brain. *IEEE Transactions on Nuclear Science*, 43(6):3361–3369.

[Schad et al., 1993] Schad, L., Bluml, S., and Zuna, I. (1993). MR tissue characterization of intracranial tumors by means of texture analysis. *Magnetic Resonance Imaging*, 11(6):889–896.

[Shattuck et al., 2001] Shattuck, D., Sandor-Leahy, S., Schaper, K., Rottenberg, D., and Leahy, R. (2001). Magnetic resonance image tissue classification using a partial volume model. *Neuroimage*, 13(5):856–876.

[Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

[Shen et al., 2003] Shen, S., Sandham, W., and Granat, M. (2003). Preprocessing and segmentation of brain magnetic resonance images. In *4th International IEEE EMBS Specific Topic Conference on Information Technology Applications in Biomedicine*, pages 149–152.

[Simmons et al., 1994] Simmons, A., Tofts, P., Barker, G., and Arridge, S. (1994). Sources of intensity nonuniformity in spin echo images at 1.5 t. *Magn Reson Med*, 32(1):121–128.

[Sled, 1997] Sled, J. (1997). *A nonparametric method for automatic correction of intensity nonuniformity in MRI data*. PhD thesis, McGill University.

[Sled et al., 1999] Sled, J., Zijdenbos, A., and Evans, A. (1999). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17:87–97.

[Smith and Brady, 1997] Smith, S. and Brady, J. (1997). Susan - a new approach to low level image processing. *Int. Jounral of Computer Vision*, 23(1):45–78.

[Soltanian-Zadeh et al., 1998] Soltanian-Zadeh, H., Peck., D., Windham, J., and Mikkelsen, T. (1998). Brain tumor segmentation and characterization by pattern analysis of multispectral nmr images. *NMR Biomed*, 11(4-5):201–208.

[Soltanian-Zadeh et al., 2004] Soltanian-Zadeh, H., Rafiee-Rad, F., and D, S. P.-N. (2004). Comparison of multiwavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms. *Pattern Recognition*, 37(10):1973–1986.

[SPM, Online] SPM (Online). Statistical parametric mapping, http://www.fil.ion.bpmf.ac.uk/spm/.

[Stadlbauer et al., 2004] Stadlbauer, A., Moser, E., Gruber, S., Buslei, R., Nimsky, C., Fahlbusch, R., and Ganslandt, O. (2004). Improved delineation of brain tumors: an automated method for segmentation based on pathologic changes of 1h-mrsi metabolites in gliomas. *Neuroimage*, 23(2):454–461.

[Studholme et al., 2004] Studholme, C., Cardenas, V., Song, E., Ezekiel, F., Maudsley, A., and Wiener, M. (2004). Accurate template-based correction of brain mri intensity distortion with application to dementia and aging. *IEEE Transactions on Medical Imaging*, 23(1):99–110.

[Studholme et al., 1998] Studholme, C., Hawkes, D., and Hill, D. (1998). A normalized entropy measure of 3-d medical image alignment. In *Medical Imaging*, volume 3338, pages 132–143.

[Talairach and Tourneaux, 1988] Talairach, J. and Tourneaux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System - an Approach to Cerebral Imaging*. Thieme Medical Publishers.

[TD, Online] TD (Online). Talairach daemon applet, http://ric.uthscsa.edu/tdapplet/.

[Therasse et al., 2000] Therasse, P., Arbuck, S., Eisenhauer, E., Wanders, J., Kaplan, R., Rubinstein, L., and et al. (2000). New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst*, 92:205–216.

[Toga et al., 2003] Toga, A., Thompson, P., Narr, K., and Sowell, E. (2003). Probabilistic brain atlases or normal and diseased populations. In Koslow, S. and Subramaniam, S., editors, *Databasing the Brain: From Data to Knowledge (Neuroinformatics)*. John Wiley & Sons, Inc.

[Tuceryan and Jain, 1998] Tuceryan, M. and Jain, A. (1998). Texture analysis. In Chen, C., Pau, L., and Wang, P., editors, *The Handbook of Pattern Recognition and Computer Vision*, pages 207–248. World Scientific Publishing Co.

[Tzourio-Mazoyer et al., 2002] Tzourio-Mazoyer, N., Landeau, B., papthanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single subject brain. *Neuroimage*, 15(1):273–289.

[Unser et al., 1991] Unser, M., Aldroubi, A., and Eden, M. (1991). Fast β-spline transforms for continuous image representation and interpolation. *IEEE Trans. Pattern Anal. Machine Intell.*, 13:277–285.

[Varma and Zesserman, 2002] Varma, M. and Zesserman, A. (2002). Classifying images of materials: Achieving viewpoint and illumination independence. *Lecture Notes in Computer Science*, 2352:255–271.

[Varma and Zisserman, 2003] Varma, M. and Zisserman, A. (2003). Texture classification: are filters banks necessary? In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 18–20.

[Velthuizen et al., 1998] Velthuizen, R., Heine, J., Cantor, A., Lin, H., Fletcher, L., and Clarke, L. (1998). Review and evaluation of mri nonuniformity corrections for brain tumor response measurements. *Med Phys*, 25(9):1655–66.

[Vinitski et al., 1997] Vinitski, S., Gonzalez, C., Mohamed, F., Iwanaga, T., Knobler, R., Khalili, K., and Mack, J. (1997). Improved intracranial lesion characterization by tissue segmentation based on a 3D feature map. *Magnetic Resonance in Medicine*, 37:457469.

[Viola, 1995] Viola, P. (1995). *Alignment by Maximization of Mutual Information*. PhD thesis, MIT.

[Vokurka et al., 1999] Vokurka, E., Thacker, N., and Jackson, A. (1999). A fast model independent method for automatic correction of intensity non-uniformity in mri data. *JMRI*, 10(4):550–562.

[Vovk et al., 2004] Vovk, U., Pernus, F., and Likar, B. (2004). Mri intensity inhomogeneity correction by combining intensity and spatial information. *Physics in Medicine and Biology*, 49(17):4119–4133(15).

[Wang et al., 1998] Wang, L., Lai, H., Barker, G., Miller, D., and Tofts, P. (1998). Correction for variations in mri scanner sensitivity in brain studies with histogram matching. *Magn Reson Med*, 39(2):322–327.

[Weisenfeld and Warfield, 2004] Weisenfeld, N. and Warfield, S. (2004). Normalization of joint image-intensity statistics in mri using the kullback-leibler divergence. In *IEEE Interational Symposium on Biomedical Imaging*.

[Wells et al., 1996] Wells, W., Kikinis, R., Grimson, W., and Jolesz, F. (1996). Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*.

[Wells et al., 1995] Wells, W., Viola, P., and Kikinis, R. (1995). Multi-modal volume registration by maximization of mutual information. In *Medical Robotics and Computer Assisted Surgery*, pages 55–62. Wiley.

[West et al., 1997] West, J., Fitzpatrick, J., Wang, M., Dawant, B., Jr., C. M., Kessler, R., Maciunas, R., Barillot, C., Lemoine, D., Collignon, A., Maes, F., Suetens, P., Vandermeulen, D., van den Elsen, P., Napel, S., Sumanaweera, T., Harkness, B., Hemler, P., Hill, D., Hawkes, D., C.Studholme, Maintz, J., Viergever, M., Malandin, G., Pennec, X., Noz, M., Jr., G. M., Pollack, M., Pelizzari, C., Robb, R., Hanson, D., and Woods, R. (1997). Comparison and evaluation of retrospective intermodality image registration techniques. *J. Comput. Assisted Tomogr.*, 21:554–566.

[Witten and Frank, 2000] Witten, I. and Frank, E. (2000). *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann.

[Yoon et al., 1999] Yoon, O.-K., Kwak, D.-M., Kim, D.-W., and Park, K.-H. (1999). Mr brain image segmentation using fuzzy clustering. In *Fuzzy Systems Conference Proceddings, 1999. FUZZ-IEEE '99. 1999 IEEE International*, volume 2, pages 853–857.

[Zhang et al., 2004] Zhang, J., Ma, K., Er, M., and Chong, V. (2004). Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine. *International Workshop on Advanced Image Technology*, pages 207–211.

[Zhu and Yan, 1997] Zhu, Y. and Yan, H. (1997). Computerized tumor boundary detection using a hopfield neural network. *IEEE Transactions on Medical Imaging*, 16:55–67.

[Zijdenbos et al., 1995] Zijdenbos, A., Dawant, B., and Margolin, R. (1995). Intensity correction and its effect on measurement variability in mri. In *Computer Assisted Radiology*.

[Zijdenbos et al., 1998] Zijdenbos, A., Forghani, R., and Evans, A. (1998). Automatic quantification of MS lesions in 3D MRI brain data sets: Validation of INSECT. *Medical Image Computing and Computer-Assisted Intervention*, pages 439–448.

[Zijdenbos et al., 2002] Zijdenbos, A., Forghani, R., and Evans, A. (2002). Automatic "pipeline" analysis of 3-d mri data for clinical trials: application to multiple sclerosis. *IEEE Transactions on Medical Imaging*, 21(10):1280–1291.