Recognizing Deviations from Normalcy for Brain Tumor Segmentation

by

David Thomas Gering

S.M., Massachusetts Institute of Technology, 2000
B.S., University of Wisconsin – Madison, 1994

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2003

Author.....
Department of Electrical Engineering and Computer Science
May 29, 2003

Certified by.....
W. Eric L. Grimson
Bernard M. Gordon Professor of Medical Engineering
Thesis Supervisor

Accepted by.....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Recognizing Deviations from Normalcy for Brain Tumor Segmentation

by

David Thomas Gering

Submitted to the Department of Electrical and Engineering and Computer Science on May 30, 2003 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

A framework is proposed for the segmentation of brain tumors from MRI. Instead of training on pathology, the proposed method trains exclusively on healthy tissue. The algorithm attempts to recognize deviations from normalcy in order to compute a fitness map over the image associated with the presence of pathology. The resulting fitness map may then be used by conventional image segmentation techniques for honing in on boundary delineation. Such an approach is applicable to structures that are too irregular, in both shape and texture, to permit construction of comprehensive training sets.

We develop the method of diagonalized nearest neighbor pattern recognition, and we use it to demonstrate that recognizing deviations from normalcy requires a rich understanding of context. Therefore, we propose a framework for a Contextual Dependency Network (CDN) that incorporates context at multiple levels: voxel intensities, neighborhood coherence, intra-structure properties, inter-structure relationships, and user input. Information flows bi-directionally between the layers via multi-level Markov random fields or iterated Bayesian classification. A simple instantiation of the framework has been implemented to perform preliminary experiments on synthetic and MRI data.

Thesis Supervisor: W. Eric L. Grimson
Title: Bernard Gordon Professor of Medical Engineering

Readers: Tomás Lozano-Pérez
William Freeman
Ron Kikinis

Contents

Chapter 1

Introduction

1.1 Motivations

On Friday, November 8, 1895, German physicist Wilhelm Conrad Roentgen recorded a photograph of his wife's hand with mysterious rays labeled "X" for unknown. Doctors' future dependence on internal imaging was so immediately apparent, that exactly 3 months later, X-rays were first used clinically in the United States.

That dependence has grown dramatically in the subsequent century as technological innovations have increased the value of doctors' "X-ray vision". While the original radiographs revealed only 2D projections, today's Computed Tomography (CT) scanners rotate the imaging apparatus to reconstruct 3D volumetric maps of X-ray attenuation coefficients. Furthermore, instead of producing contrast between only bones and soft tissues, today's Magnetic Resonance Imaging (MRI) scanners can differentiate between various soft tissues. They accomplish this by detecting radio frequency signals emitted by the excited magnetic dipoles of each tissue's constituent molecules. In addition to these modalities for gathering anatomical data, functional information can be acquired by functional MRI (fMRI) or Positron Emission Tomography (PET). fMRI measures the indirect effects of neural activity on blood flow and oxygen consumption. PET can distinguish metabolically active tumors from necrotic areas by detecting the gamma rays emitted by positrons that collide with the brain's electrons. These positrons originate from the breakdown of radioactive tracers that are injected into the circulatory system to concentrate in regions of high blood flow and metabolism.

While the advances in medical imaging have been impressive, the need for scientific progress does not end with the image acquisition process. Post-processing, or computational analysis of the image data, has attracted researchers in artificial

intelligence, pattern recognition, neurobiology, and applied mathematics. Many clinical applications of medical image analysis rely on computers to embody the capability to understand the image data to some degree. This understanding involves comprehension of knowledge of the image content. Hence, the basic component of image understanding is image *segmentation*. Segmentation is the process of labeling a scan's volume elements, or *voxels*, according to the tissue type represented. A subset of the clinical applications dependent on segmentation are outlined below.

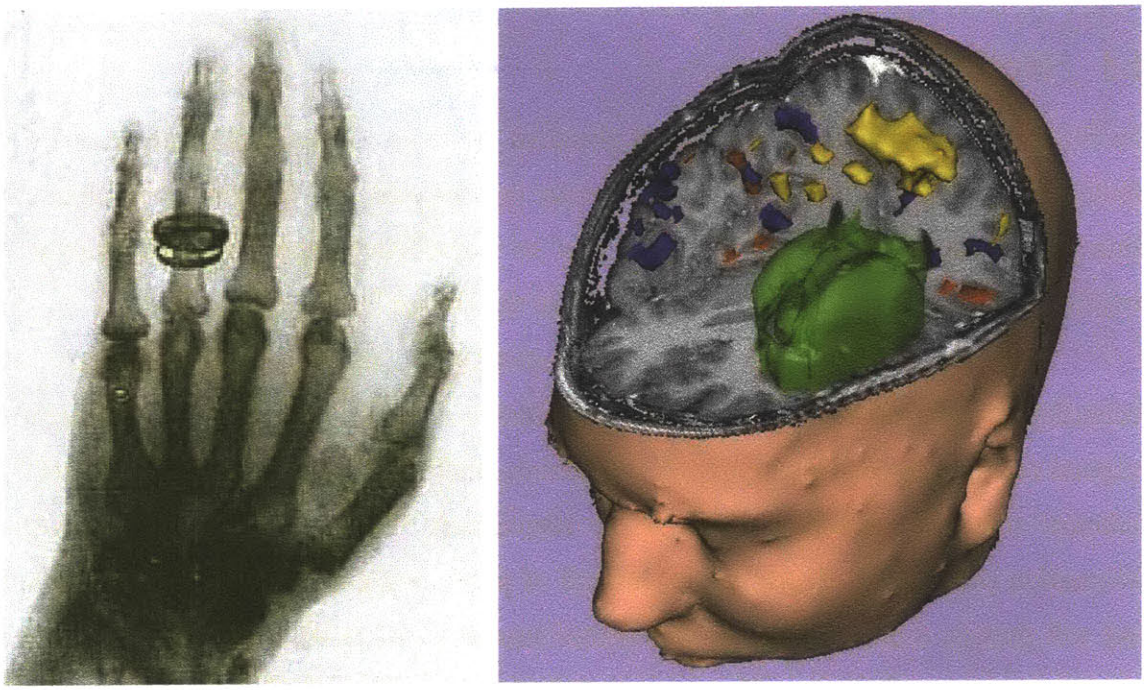

Figure 1.1. Advances in Internal Medical Imaging (Left:) In 1895, X-ray vision of Bertha Roentgen's hand and wedding ring fascinated the public and puzzled scientists. (Right:) Today, "augmented X-ray" vision is enabling doctors to optimize patient diagnosis, treatment, and monitoring, as well as improve surgical planning and guidance. In this example, the 3D Slicer [Gering01] is used to fuse anatomical MRI data of a tumor (green) with functional MRI data that localizes visual verb generation (blue), auditory verb generation (red) and the motor cortex (yellow).

1.1.1 Surgical Planning

Many surgeries are delicate operations that require pre-operative planning to ascertain the operability, or identify the optimum approach trajectory. The benefits of planning vary widely with the circumstances encompassing each case, but planning is most critical in cases where the target tissue is situated either deeply or within fragile surroundings. Consider neurosurgery, where tumors can either infiltrate functional tissue, or push it

aside. A tumor that invades eloquent cortex can be considered inoperable for the sake of preserving the quality of life rather than its longevity. For example, the patient depicted in Figure 1.1 had a tumor in Broca's area where 96% of speech is generally processed. The 3D integrated visualization clearly demonstrated that speech activity had migrated to the right side, proving the operability of this lesion.

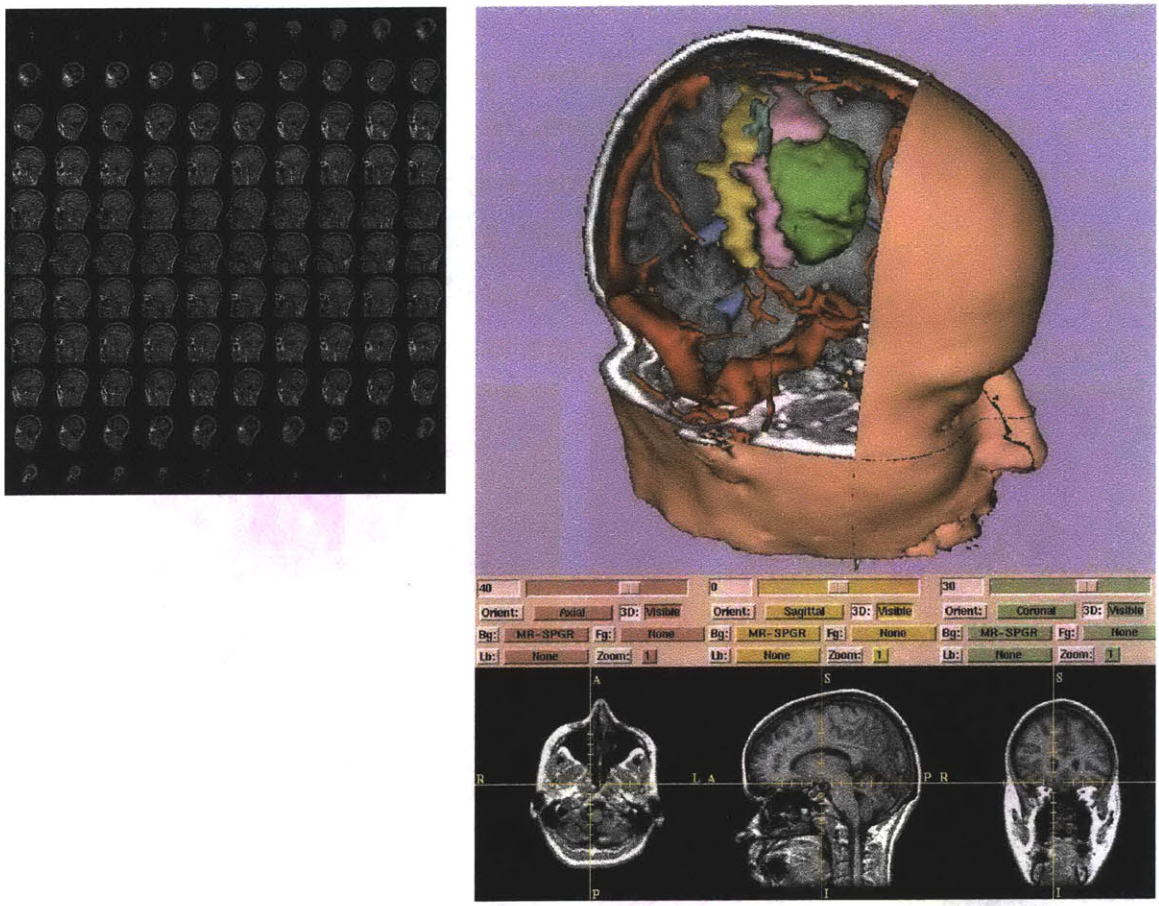

Figure 1.2. Lightbox vs. 3D Graphics (Left:) 3-D data is traditionally viewed by radiologists as a set of consecutive 2-D slices. (Right:) Multiple data sets (MRI, fMRI, MR Angiography) are registered, or aligned, and the surfaces of critical structures are rendered to reveal their spatial relationships: vessels (red), tumor (green), pre-central gyrus (pink), post-central gyrus (yellow), and motor cortex (blue).

Accurate visualization is vital in a variety of other neurosurgical cases. For malignant tumors, the complete resection of diseased tissue is required for prolonged survival. For biopsies and benign tumors, the tolerance for error is significantly lower given that the risks of complications, such as speech impairment, blindness, paresis, or hemorrhaging, threaten to outweigh the benefits of operating. Since the operational

hazards are structures arrayed in 3D space, they lend themselves to 3D explorative viewing from novel trajectories not physically possible. Figure 1.2 illustrates the contrast between the traditional approach of viewing a sequence of slices on a 2D sheet of film, and the 3D visualization made possible by computational analysis [Gering99b].

1.1.2 Surgical Guidance

Surgeons can benefit not only from pre-operative planning, but also online guidance for precise, intra-operative localization [Gering99a], as depicted in Figure 1.4. Patients can benefit from the smaller access holes, shorter hospital stays, and reduced pain made possible by minimally invasive surgery [Jolesz97, Black97]. Therefore, surgical guidance aims to equip the surgeon with an enhanced vision of reality that enables the surgeon to approach the target tissue without inflicting harm to neighboring healthy structures

Figure 1.3. Systems for Surgical Guidance The surgeon stands within the gap of an Intervention MRI suite [Schenk95], monitoring the 3D display screen that presents the results of computational analysis. (Images appeared in [Grimson99]. Used with permission.)

While an unassisted surgeon can see the surfaces of exposed tissues, the internal structures are invisible. Image-guided surgery provides "X-ray" vision of what lies beyond the exposed surfaces, what types of tissue are seen, and what functions the tissues serve. Different types of tissue may be difficult to distinguish with the eye alone, but

appear markedly different on certain medical imaging scans. Similarly, tissues that handle critical functions, such as voluntary movements, speech, or vision, appear identical to companion tissue, but can be highlighted by a functional exam.

Surgical guidance systems, such as Instatrak (GE Nav, Lawrence, MA) and Signa-SP (GE Medical Systems, Waukesha, WI), track surgical instruments for rendering their position relative to anatomical structures within the 3D operating theater, as depicted in Figures 1.3 and 1.4.

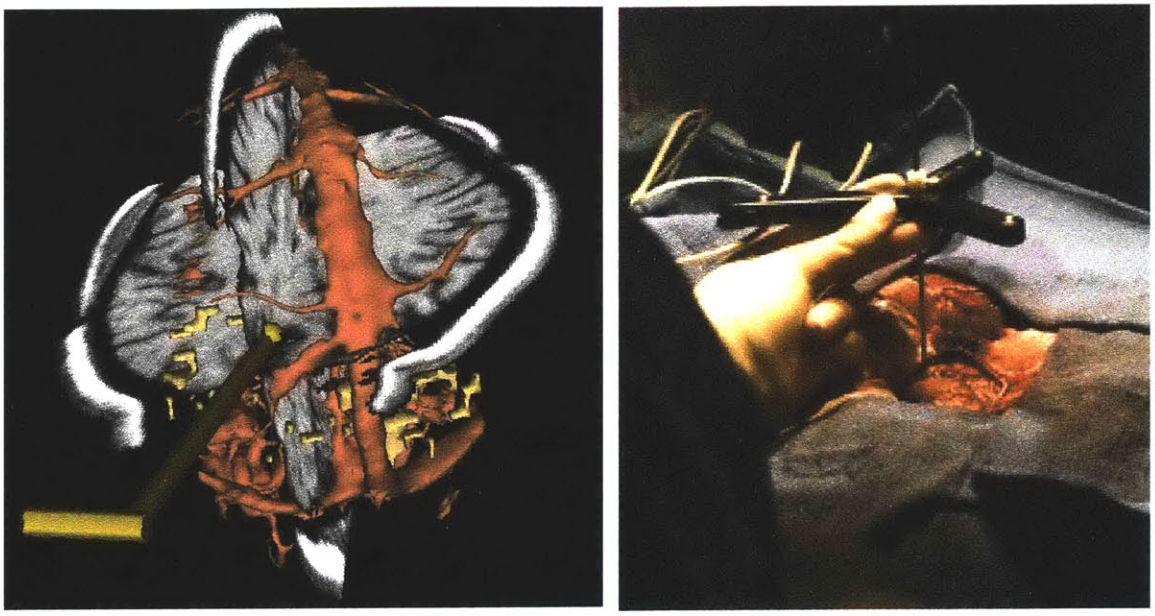

Figure 1.4. Tracking and Rendering Instruments for Surgical Guidance (Left:) The surgeon resects a cavernoma by maneuvering the instrument (yellow wand) to avoid the hazards posed by the vasculature (red) and visual cortex (yellow). (Right:) Photograph of the tracked wand in surgery.

1.1.3 Volumetric Analysis

Quantitative measurements often contribute to disease characterization, treatment planning, and progress assessment. Traditional metrics have been crudely based on 2D geometry. For example, muscle volume was characterized by radius, and joint range-of-motion studies were drawn on X-ray films with rulers and protractors. Computational image analysis allows true volumetric measurements to be performed, as shown in Figure 1.5 in a study of female incontinence [Fielding00, Dumanli00].

Figure 1.5. Volumetric Analysis and Studies of Dynamics 3D models of the female pelvis such as bones (white), bladder/urethra (yellow), vagina (blue), uterus (green), rectum (gray), and the levator ani muscle (pink) can be visualized and quantified in 3D space – independent of the orientation of the slice acquisition. The purple line between two blue markers is measuring the distance of the pubococcygeal line (level of the pelvic floor, and minimum width of the birth canal).

1.1.4 Time Series Analysis

Certain forms of quantitative analysis are not performed at a single snapshot in time, but rather, over a series of many imaging exams covering several days or decades. Example studies include responsivity of pathology to pharmacutical treatments, effects of exercise on certain tissues, and the time course of disease such as schizophrenia and Alzheimer's disease [Guttmann99].

1.1.5 Computer Aided Diagnosis

While the applications listed above have focused on treatment, computational analysis has recently begun to focus on computer-aided diagnosis (CAD) as well. Particular attention has been given to breast and respiratory system lesions, and we refer the reader to [Giger00, Ginneken02] for survey articles pertaining to each of these two applications. Technological trends suggest that the need for CAD will expand beyond such niche applications. CT scanners have recently progressed from scanning not one slice at a time, but 16 slices concurrently. Similarly, commercial MR scanners have progressed from

having two independent receivers to currently featuring eight or more. These advances in data acquisition enable unprecedented applications such as 4-D cardiac exams and non-invasive, rapid, whole-body screening. As a corollary to Moore's law for the growth of semiconductor chip densities, the amount of medical data is growing exponentially despite the fact that the human brain – and therefore a radiologist's capacity – does not adhere to Moore's law. Understanding such massive amounts of data will eventually become too costly and time-consuming, or even impossible, for human radiologists. With the number of US radiologists growing a mere 3% annually [BusinessWeek02], we believe the future of CAD will align less with attempting to perform tasks at which human radiologists excel, and more with performing tasks that humans simply cannot do.

1.2 Brain Tumor Segmentation

All the applications discussed thus far have relied on computers embodying the capability to understand the image data as a result of performing segmentation. Widespread clinical use of segmentation is hindered by two shortcomings: the inordinate amount of a user's time required to generate the segmentations, and the inter- and intra-operator variability. For example, the 3D figures displayed above each required several hours of an operator's time to manually trace the outline of each anatomic structure on every slice – typically 124 per volume. Figure 1.6 details this painstakingly long process. There is a significant amount (~15%) of both inter- and intra-operator variability resulting in an inconsistency between experts, and a lack of repeatability for a single expert. Therefore, automatic and nearly automatic techniques can potentially assist clinicians by greatly reducing the requisite time while increasing the repeatability.

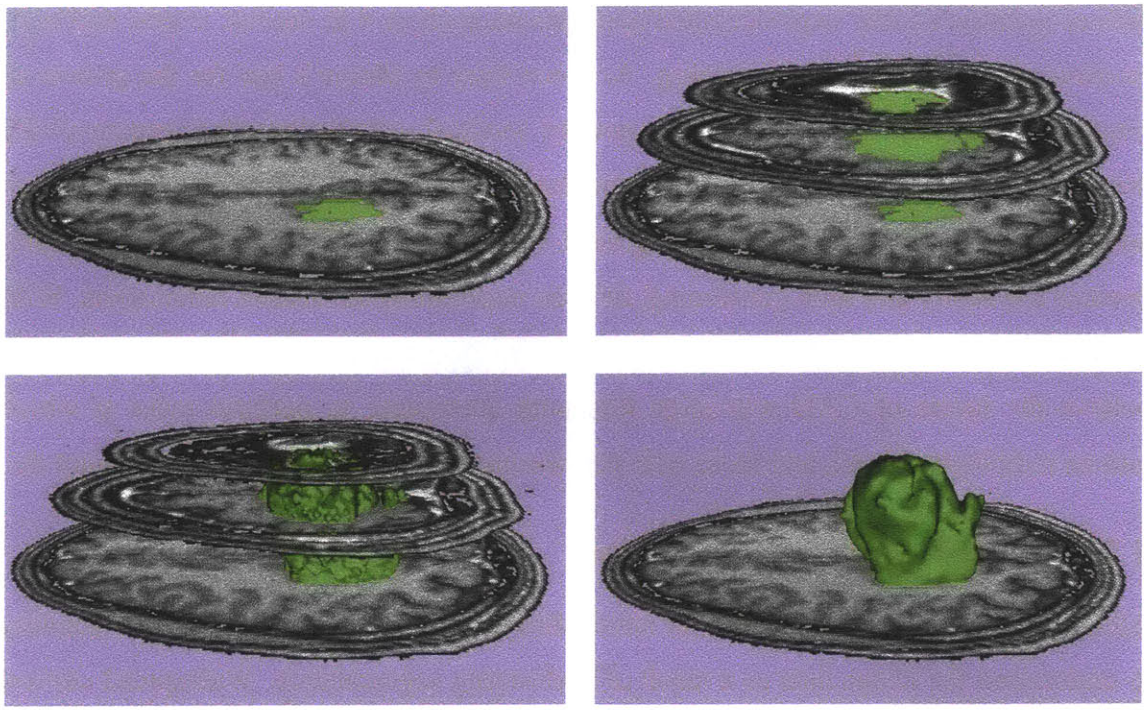

Figure 1.6. Manual Tumor Segmentation Process for 3D Surface Generation (Top Left) The operatator traces the outline of the tumor boundary, (Top Right) and repeats this process on every slice in the volume. (Bottom Left:) A 3D surface is then generated to encompass the segmentation [Lorensen87] (Bottom Right) and smoothed to remove digitazation artifacts [Schroeder92].

1.2.1 Related Work

The literature is rich with techniques for segmenting healthy brains – a task simplified by the predictable appearance, size, and shape of healthy structures. See [Clarke95, Pham00b] for survey articles. Many of these methods fail in the presence of pathology – the very focus of segmentation for image-guided surgery. Furthermore, the techniques that are intended for tumors leave significant room for increased automation and applicability.

Specifically, we consider the task of segmenting large brain tumors such as gliomas, meningiomas, astrocytomas, glioblastoma multiforme, cavernomas, and Arteriovenous Malformations (AVM). In practice, segmentation of this class of tumors continues to rely on a combination of manual tracing and semi-automation using low-level computer vision tools such as thresholds, morphological operations, and connective component analysis. Automatic techniques tend to be either region- or contour-based. (Note that the

term "automatic" has been applied very liberally in the literature. Automatic algorithms greatly reduce, but rarely completely remove, user interaction.)

Region-based methods seek out clusters of voxels that share some measure of similarity. Most methods reduce operator interaction by automating some aspects of applying low-level operations. From early on, these methods were grounded in a statistical modeling of each tissue class, combined with morphological operations such as smoothing and connectivity [Cline87, Cline90]. Threshold selection can be assisted through histogram analysis [Joe99], and logic can be applied to the application of low-level vision techniques through a set of rules to form a knowledge-based system [Clark98]. Another approach is to perform unsupervised clustering with the intention that the tumor voxels will congeal into their own cluster [Capelle00]. Such methods, although fully automatic, only apply to enhancing tumor, that is, tumor that appears markedly hyper-intense on MRI following admission of a contrast agent such as gadolinium. Since statistical classification alone may not allow differentiation between non-enhancing tumor and normal tissue, anatomic information derived from a digital atlas has been used to identify normal anatomic structures. Of these approaches, the most successful has been the iteration of statistical classification and template matching as developed in [Warfield95, Warfield00, Kaus01]. However, there remains a reliance on several minutes of the operator's time for patient-specific training. For good results, the template needs to be closely similar to the patient's anatomy, and the tumors must be homogenous. The use of morphological operations has the drawback of making a very crude assumption about the radius parameter that is both application-dependent (anatomy) and scan-dependent (voxel size). Furthermore, such operations destroy fine details and commit to irreversible decisions at too low of a level to benefit from all the available information – thus violating Marr's principle of least commitment [Marr82].

Contour-based methods evolve a curve based on internal forces (e.g. curvature) and external forces (e.g. image gradients) to delineate the boundary of a tumor. Since they experience similar drawbacks as the region-based approaches, methods that claim to be fully automatic can do so only because they apply to tumors that are easily separable from their surroundings. (See [Zhu97] for an example using a Hopfield neural network to evolve a snaking contour). Level set based curve evolution [Kichenassamy95, Yezzi97]

has the advantage over region-based approaches in that the connectivity constraint is imposed implicitly rather than through morphological operations. However, 3D level-sets find limited use in medical practice due to their reliance on the operator to somehow set the sensitive parameters that govern the evolution's stopping criteria. Furthermore, the more heterogeneous a tumor may be, the more user interaction is required.

Both region- and contour-based segmentation methods have ignored the bias field, or patient-specific, signal inhomogeneity present in MRI. While acceptable for small tumors, an accurate segmentation method cannot overlook the bias. One reason it is overlooked is the difficulty in computing an inhomogeneous field over an inhomogeneous tumor (and the fact that inhomogeneous tumors have been largely overlooked due to their difficulty anyway). Regardless, the bias field is slowly varying, and therefore its computation from the regions of healthy tissue could be extrapolated over tumor tissue to provide some degree of benefit. Methods for segmenting healthy brains have incorporated the EM algorithm [Dempster77] to simultaneously arrive at both a bias field and a segmentation into healthy tissue classes [Wells96b]. There have been several extensions, such as collecting all non-brain tissue into a single class [Guillemaud97], handling salt and pepper noise with Markov random fields [Held97], using a mean-field solution to the Markov random fields [Kapur99], incorporating geometric constraints [Kapur99], using a digital brain atlas as a spatially-varying prior [Leemput99a], automating the determination of the tissue class parameters [Leemput99b], and identifying MS lesions as hyper-intense outliers from white matter [Leemput01a]. Coincident with our work in [Gering02b], [Moon02] also extended EM-based segmentation to apply to brain tumors, but only those that enhance with administration of contrast agents. The technique does not apply to the single-spectrum MRI considered in our study.

1.3 Contributions

The two primary contributions of this thesis are the approach of recognizing deviations from normalcy, and the framework for a contextual dependency network that incorporates context – both immediate and broad.

1.3.1 Recognizing Deviations from Normalcy

In contrast to the aforementioned methods for tumor segmentation, the novel hypothesis underlying this thesis is that we can segment brain tumors by focusing not on what typically represents pathology, but on what typically represents healthy tissue. Therefore all training is performed exclusively on healthy brains, and all other forms of *a priori* knowledge that are embedded into the algorithm represent descriptors of normal anatomy. Our method extends the EM-based segmentation to compute a fitness map over the image to be associated with the probability of pathology. That is, we extend the segmentation algorithms for healthy brains in order to make progress toward solving the recognition problem encountered when segmenting tumors. Indeed, the entire motivation behind the Live Wire semi-automatic approach [Falcao98, Falcao00, O'Donnell01] was an acknowledgement that segmentation tightly couples two processes: recognition and delineation. While computers have been adept at delineation (specifying the precise spatial extent of an object), humans – by nature of their global knowledge – are far better suited for recognition (roughly identifying an object's whereabouts). Rather than leaving that aspect for humans, the goal of this thesis is to improve the computer's capability for recognizing brain tumors, and thereby address the drawbacks to the existing region- and contour-based methods.

1.3.2 Contextual Dependency Networks (CDN)

We designed a framework for Contextual Dependency Networks that incorporate context, both immediate and broad. We extended EM-based segmentation with region-level properties such as shape descriptors, and we derived a novel multi-level MRF approach.

Inherent ambiguity necessitates the incorporation of contextual information into the brain segmentation process. Consider the example of non-enhancing tumor tissue that mimics the intensity of healthy gray matter, but is too thick to be gray matter. An algorithm's low-level computer vision techniques could first classify the tissue as gray matter, and a higher-level stage – through its broader understanding of context – could correct the classifications of the first-pass. This example motivates the introduction of hierarchical context into the segmentation process. A voxel's classification could be considered on several levels: the voxel itself, the voxel's immediate (Markov)

neighborhood, the voxel's region (entire connected structure), the global setting (position of the voxel's structure relative to other structures), and user guidance. Just as a voxel-wise classification must be computed prior to a neighborhood-wise refinement, a voxel's region must be classified before features regarding the size and shape (or other intrinsic properties) of that region can be computed.

Table 1.1. A Contextual Dependency Network is a framework that features no decisions made by certain layers that permanently (and perhaps adversely) affect other layers. Information flows between the layers (bidirectionally depending on implementation details) while converging toward a solution

#	Layer	Definition	Our Simple Computation
5	User *(oracle)*	Spatially specific points clicked on by the user on the fly as corrective action.	Mouse clicks trigger re-iteration.
4	Inter-structure *(global)*	Relative position of a voxel's structure to other structures.	Distance from other region boundaries.
3	Intra-structure *(region)*	Relative position of a voxel within its own structure.	Distance from own boundary.
2	Neighborhood *(local)*	Classification of a voxel's immediate neighbors.	Mean Field MRF
1	Voxel *(point)*	Classification based on voxel's intensity.	EM, ML or MAP

Figure 1.7 previews the results from Chapter 6 to demonstrate that by recognizing deviations from normalcy, the same algorithm can identify both hyper-intense and hypo-intense tumors.

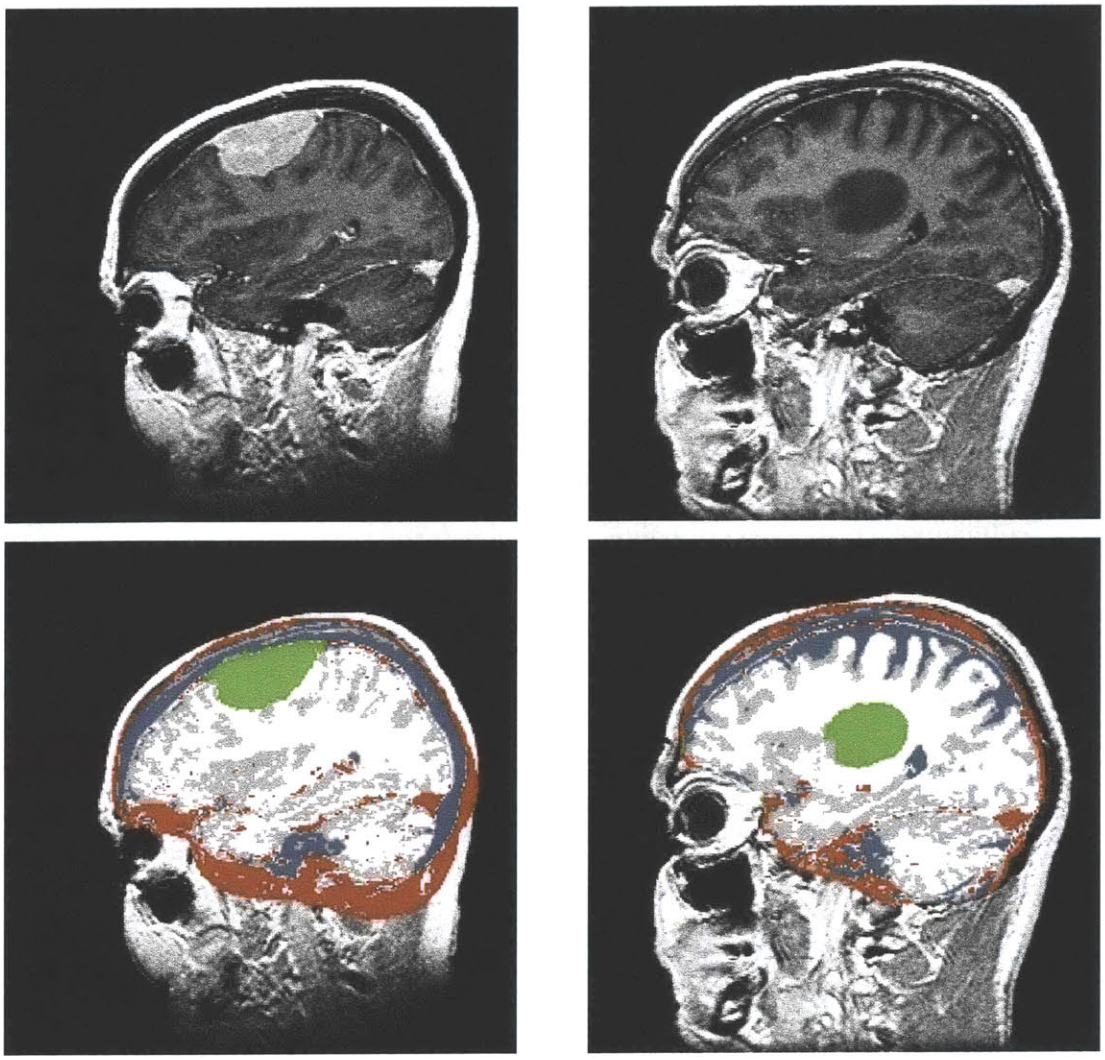

Figure 1.7. Preview of Results. The original input images on top were segmented to produce the results on the bottom. The algorithm has knowledge of the expected properties, with respect to both intensity and shape, of healthy tissues only. Colors represent tumor (green), white matter (white), gray matter (gray), CSF (blue), and vessels (red).

1.4 Roadmap

In the next two chapters, we develop the rationale for our unique approach to tumor segmentation. And in the following three chapters, we present the enabling technology. In all, this thesis exhibits the following organization by chapter:

Chapter 1: Introduction

Motivations, brain tumor segmentation, contributions, roadmap

Chapter 2: Imaging Model

Imaging model, experimental data

Chapter 3: Recognizing Deviations from Normalcy

Feature detection vs. anomaly detection, deviations from normalcy, nearest neighbor pattern matching, contextual dependency networks

Chapter 4: CDN Layer 1: Voxel Classification

Mathematical background, robust bias estimation, spatially-varying priors, computing a probability of pathology, and generative models

Chapter 5: CDN Layer 2: Neighborhood Classification

Markov and Gibbs random fields, MRF design, MRF optimization, factorizing the joint distribution, algorithmic comparisons, recognizing deviations from normalcy

Chapter 6: CDN Layers 3-5: Intra-structure and Inter-structure Classification

The ACME segmenter, multi-layer MRF, correcting misclassified voxels, correcting misclassified structures, user interaction, and results on real data

Chapter 7: Conclusions and Future Work

Summary, future work

Chapter 2

Imaging Model

To set the stage for the experiments ahead, this chapter introduces our imaging model and the data sets used throughout this thesis.

2.1 Imaging Model

Before we begin experimenting, we need to model the image generation process. There are four reasons to construct such a model:

1. The image generation process is incredibly complex, but minor subtleties can be ignored, resulting in much greater simplicity. Constructing a model is our process for discerning which aspects to include, and which to exclude, from our algorithm.
2. The model will support all assumptions that we make while deriving algorithms throughout this thesis.
3. The model will be computer-simulated to generate synthetic data to use in experimentation. Although synthetic data should not be used for final validation of an algorithm designed for real data, it is very useful for the designer to have control over various image aspects in order to better explore both the problem and its solution.
4. Because ground truth is known, model-generated data is useful for validating the correctness of the software implementation of an algorithm.

The image generation process consists of two main components: an object function that describes the spatial extent of the object with perfect resolution, and a mapping function that maps object space to image space. This mapping function is essentially the image acquisition process, taking an object as input, and producing an image as output, as shown in Figure 2.1.

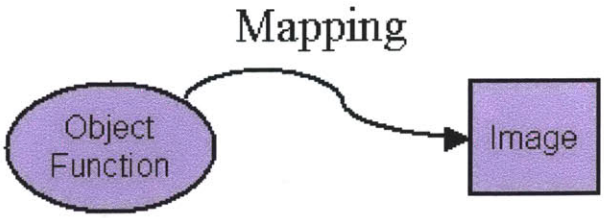

Figure 2.1. Image Acquisition Process. The image acquisition process performs a mapping from the object function to image space.

The mapping function's components are depicted in Figure 2.2, and each will be described in detail below. Recall that this is intended to be our working model, but not a fully accurate description of the real process.

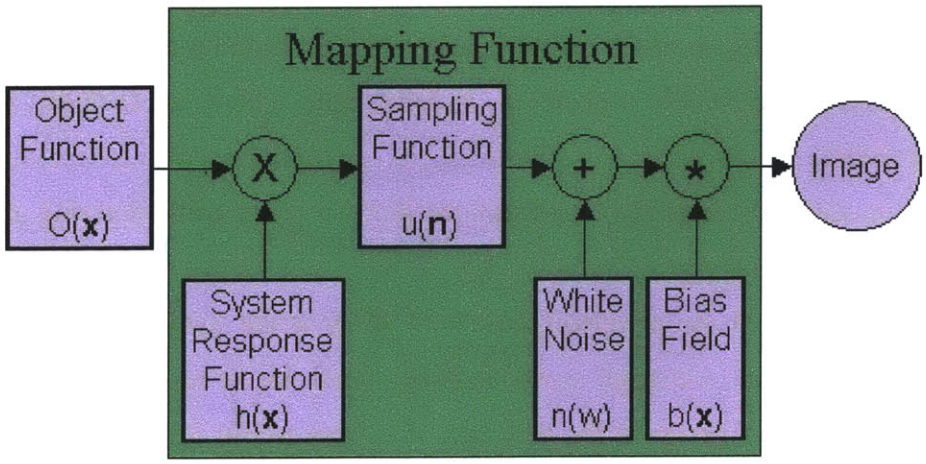

Figure 2.2. Image Generation Process. The image acquisition process combines functions of space (x), tissue type (w), and discretization (n).

The first step in the image acquisition process is to convolve the *object function* $O(\mathbf{x})$ with a *system response function* $h(\mathbf{x})$, which is also referred to as a *point spread*

function. This convolution operation models the system's limited resolution by blurring the object so that sufficiently fine structures become unresolvable. Note that if $h(\mathbf{x})$ were the impulse function, then image voxels would be statistically independent. However, MR scanners physically perform the Fourier transform, so image reconstruction involves applying the inverse transform to recover an image. Finite and discrete computation results in sinc-shaped Gibbs ringing surrounding each voxel's signal. Scanning protocol parameters (including voxel size described in a later stage of this acquisition process) are chosen to minimize the signal's spread over neighboring voxels, but a very small quantity of correlation does exist.

The second step in the imaging process is the sampling that produces a discrete lattice of image voxels. This digitization of a continuous function is responsible for introducing partial volume artifacts, which we will examine in Chapter 6.

The next stage in the process introduces additive white noise with tissue-dependent variances. Noise in MR images has peculiarities caused by rectification during image reconstruction. MR signal detection is performed in quadrature, producing real and imaginary signals. Medical images are produced by taking the magnitude of these signals, which rectifies both the signal and the noise:

$$\text{magnitude image} = \text{SQRT}[(\text{real signal} + \text{real noise})^2 + (\text{imag. signal} + \text{imag. noise})^2]$$

As a result that is elegantly derived in [Henkelman85], the noise in the presence of strong signal has a nearly Gaussian distribution [Simmons96], but noise near low signal, such as in the background, is best modeled with a Raleigh distribution [Haacke99].

The final stage in the pipeline involves combination with a multiplicative *bias field* $b(\mathbf{x})$ to model spatial inhomogeneity. Present in every medical imaging modality, the cause of the bias field varies greatly. For example, the bias field is attributed to dissipation with depth in Ultrasound, Compton scattering in CT, and asymmetric positioning of reception coils, among other effects, in MRI [Simmons94, Sled98].

The above imaging model will form the basis for making a number of assumptions throughout this thesis. The model reveals that the problem of classifying

image voxels is very ill-posed. According to [Tikhonov77], a problem is mathematically ill-posed if its solution does not exist, is not unique, or does not depend continuously on the initial data. In our case, the solution is not unique because the model accounts for five major voxel intensity modifiers, as summarized in Table 2.1. Therefore, additional constraints are needed to guarantee the uniqueness of the solution, and convert this ill-posed problem into a well-posed one. Computer vision algorithms have long relied on regularization to make a problem well-posed, as surveyed in [Poggio85]. The approach taken by this thesis will be to impose the typical smoothness constraints in addition to novel contextual constraints. Observe that an approach of searching for deviations from normalcy renders an ill-posed problem to be even more ill-posed because an extra voxel modifier of *pathology* is effectively added to Table 2.1. Regardless, this approach has the benefit of allowing general tumor recognition, so we will confront the challenge of making the problem well-posed by adding contextual constraints.

Table 2.1. Voxel Intensity Modifiers

Effect	Cause
Tissue heterogeneity	Object Function
Voxel correlation	System Response Function
Nonuniformity	Bias Field
Partial volume artifacts	Sampling Function
Additive noise	Detector noise, and rectification

2.2 Experimental Data

This section introduces the data sets that will be used for experimentation throughout this thesis.

2.2.1 Synthetic Data

Synthetic data will be shown to be useful in the experiments of the subsequent chapters. This is because the ground truth is known, and vast amounts of data can be easily produced. We must be careful to ensure that the synthetic data spans an interesting and important space of possible cases. Therefore, we generated the synthetic data set by simulating each stage of the pipeline developed in Section 2.1.

2.2.1.1 Synthetic Object Function

The object function for 2-D brains was simulated by generating white matter that was shaped as a disc with its radius modulated by a sine wave. The white matter was then surrounded with a layer of cortical gray matter, which was surrounded with a coating of CSF, which was enveloped by a perimeter of scalp. Then, subcortical gray matter, the left ventricle, and the right ventricle were each added as overlapping discs near the brain center. Finally, vessels were added as arcs. With uniform distributions governing the parameters for shape and position, there are 2.5×10^{17} equally probable "healthy brains" from the object function. Figure 2.3 depicts several examples to demonstrate the variability. Furthermore, 5.8×10^{5} different circular tumors can be randomly added.

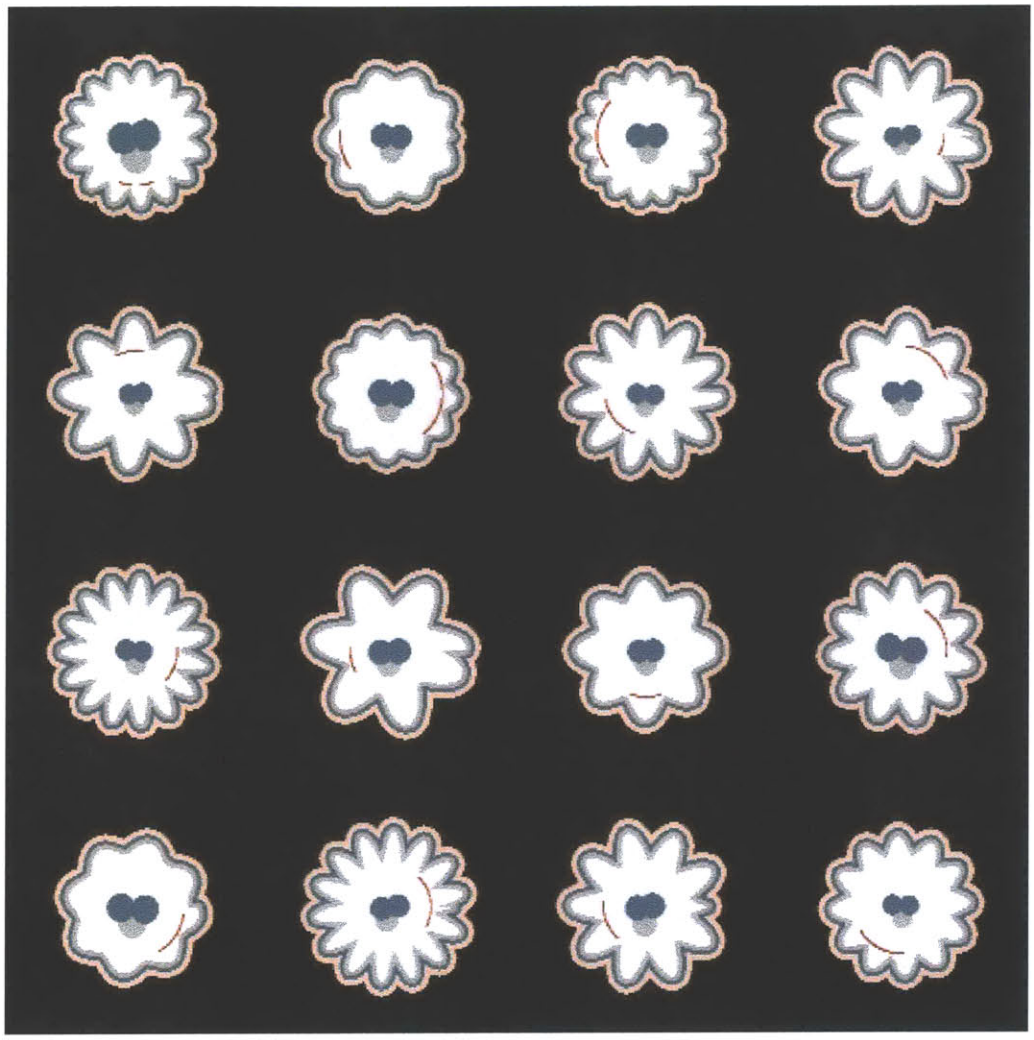

Figure 2.3. Synthetic Object Function. Several examples drawn at random from the simulated Object Function are shown as ground truth segmentations. Color coding: white matter (white), gray matter (gray), CSF (blue), scalp (tan), vessel (red).

2.2.1.2 Synthetic Imaging Function

Given a tissue labeling from the object function, the imaging process is simulated by adding Gaussian-distributed intensities to form an image. Statistical parameters for each tissue class were measured from computing the mean and variance of voxels in one of the scans in the real data set. To prevent partial volume artifacts from corrupting the measurements, the tissue was segmented, and then the segmentation was eroded to remove boundary voxels (Figure 2.4). Table 2.2 lists the resultant measurements both

before and after erosion. We will reference this table again in the discussion of handling partial volume artifacts in Chapter 6.

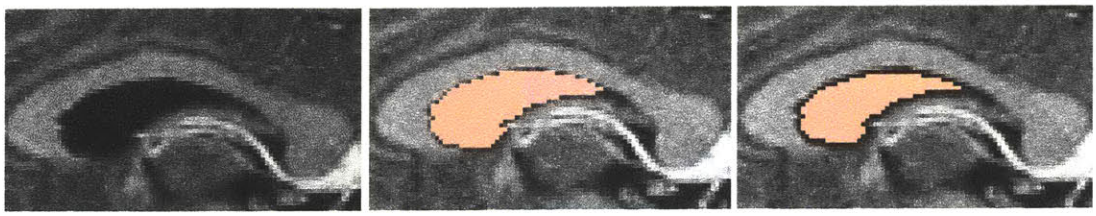

Figure 2.4. Measuring Statistical Parameters. Parameters were measured from a real scan (left) by segmenting a tissue (center) and eroding its boundary (right). Pictures are shown for CSF in the ventricle, and Table 2.2 lists the results for all tissue types.

Table 2.2. Statistical Measurments for Synthetic Data. The model used the values obtained without partial volume artifacts (PVA) to avoid inaccurately inflated variances.

Tissue Type	With PVA		Without PVA	
	Mean	Variance	Mean	Variance
White matter	117	55	120	33
Gray matter	91	43	90	29
CSF	32	97	28	48
Scalp	198	1919	217	1150
Vessel	179	631	183	200

Using the mean values shown in the right side of Table 2.2, a 512x512, high-resolution, intensity image is produced from the object function's label map. Then, to simulate the system response function, this image is convolved along each dimension with a Gaussian kernel (1,4,6,4,1), and down-sampled to form a 256x256 image. Figure 2.5 reveals that the result accurately depicts the limited resolution and partial volume artifacts of real scanners. Next, additive white noise is simulated by adding random samples drawn from a 0-mean, Gaussian process. (For convenience, we used the same variance of 36 for all tissues, where this value was chosen based on inspection of the right side of Table 2.2.)

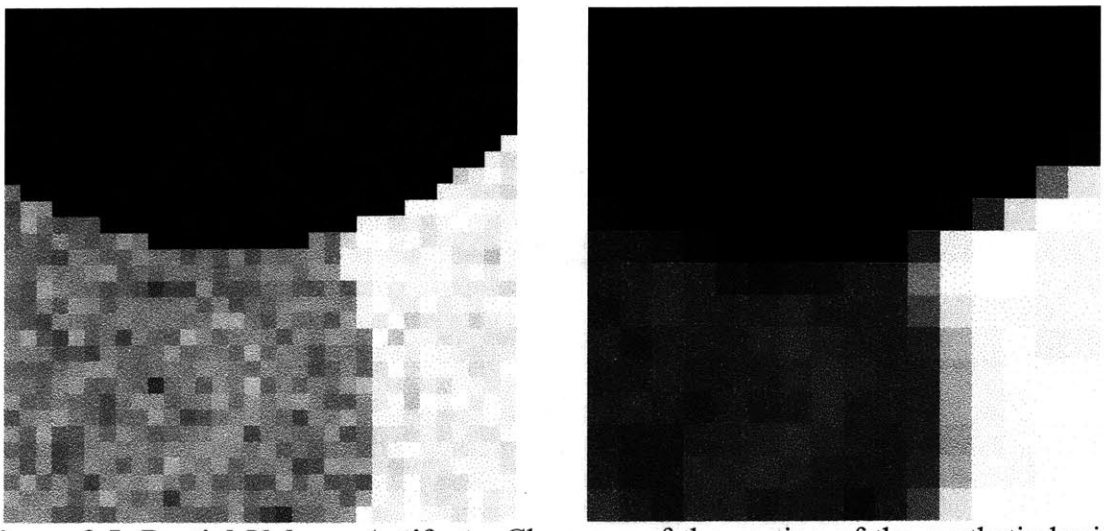

Figure 2.5. Partial Volume Artifacts. Close-ups of the portion of the synthetic brain where ventricle, subcortical gray matter, and white matter converge are shown. An image with PVA (right) is computed as a blurred, down-sampled version of a high-resolution image without PVA (left).

Furthermore, spatially-varying bias fields are included by modulating the image with a smoothly varying function. We experimented with a linear ramp and a low-frequency sinusoidal wave, as pictured in Figure 2.6.

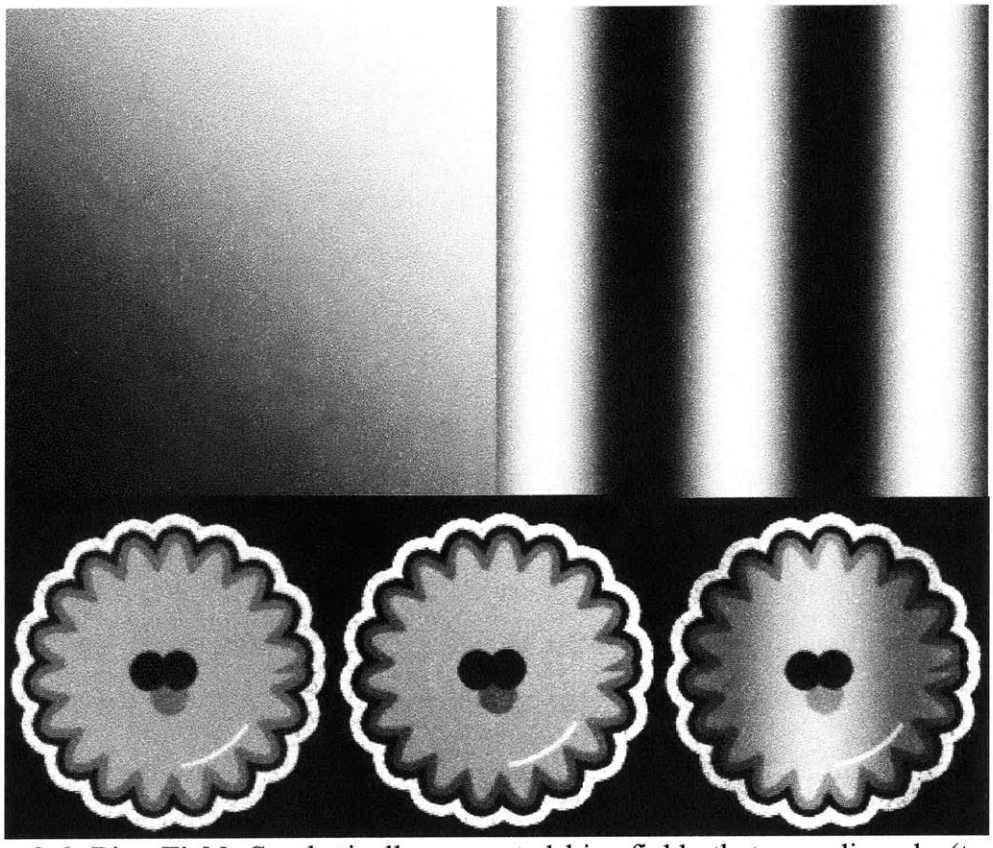

Figure 2.6. Bias Field. Synthetically generated bias fields that vary linearly (top left) and sinusoidally (top right) are applied to an original image (bottom left) to produce the bottom center and right images, respectively.

2.2.2 Real Data

Besides using synthetic data, experiments were performed on a publicly available database of 10 tumor scans [BWHSPL]. To understand this data set, we briefly describe the nature of multi-spectral MRI.

2.2.2.1 MRI

MR imaging is performed by measuring the radio signal emitted by magnetic dipoles (hydrogen nuclei) as they relax back to their equilibrium position following excitation by a momentarily-applied magnetic field. The dipoles cannot merely align themselves with the magnetic field as little bar magnets would, because the laws of quantum physics restrict these dipoles to be in one of two states. They precess like spinning tops, and the "tops" can make one of two angles with the axis of rotation. The applied magnetic field

excites approximately one in a million of the dipoles to flip states, and the total sum of all these miniature magnets is a magnetization that decays once the field ceases to be applied.

This decay has two separate components referred to as T1 and T2 relaxation. T1 relaxation occurs as the dipoles return their orientation to the equilibrium position, and T2 relaxation results from the precession of the dipoles falling out of phase with each other. The rate of T1 and T2 decay varies depending on the molecular chemistry of the tissue inhabited by the hydrogen nuclei. Scanning parameters can be set so that the source of image contrast (light and dark regions) is weighted more toward either the T1 or T2 relaxations.

In many instances, physicians acquire both T1- and T2-weighted MRI. For example, extracting a well-defined tumor boundary from diagnostic images may be hindered by surrounding edema. Edema, or liquid diffused between cells, spreads finger-like into the white matter, while avoiding the gray matter and cortex whose cell packing is too dense to harbor as much fluid [Youmans96]. The extra-cellular fluid of edema and increased intra-cellular fluid of tumors can be confused when ascertaining the tumor/tissue interface. Ambiguity can be diminished by having both T2-weighted MR images and T1-weighted MR images with contrast. A contrast medium (liquid that appears bright on MRI) is administered to the patient, and taken up more by the areas of active tumor tissue. The contrast agent forms a hyperintense region on MRI where the agent leaks out of vasculature into tissue. This occurs where the blood-brain barrier breaks down, and is thus an indication of a high grade, rather than a low grade, glioma (a mass created in the brain by the growth of abnormal cells, or the uncontrolled proliferation of cells).

Brain segmentation techniques have long exploited the increased soft-tissue contrast available from multi-channel MRI [Vannier85]. Standard diagnostic protocols involve collection of proton density, T2-weighted, T1-weighted pre-contrast, and T1-weighted post-contrast images. Therefore, if we can demonstrate our framework to function reasonably well given only noisy, single-channel data, then results will be that much better on better data. The fact remains that humans can easily recognize tumors to a large degree from noisy, single-channel MRI. For example, although edema is

remarkably clear given both T1- and T2-weighted scans, radiologists do tend to identify edema from T1-weighed imagery alone. Our motivation is to progress toward endowing computers this human-like ability.

2.2.2.2 Tumorbase

The tumorbase [BWHSPL] is an especially difficult data set with which to work because it contains only single-channel, post-contrast MRI with poor gray-matter / white-matter contrast. For performing validation, one slice of each scan was segmented by 4 different experts, and the entire volumes were segmented by one expert. Table 2.3 lists the patient characteristics. The acquisition protocol was:

```
SPGR T1 POST GAD
resolution: 256x256x124
pixel size: 0.9375 x 0.9375 mm
slice thickness: 1.5 mm
slice gap: 0.0 mm
acquisition order: LR
```

Table 2.3 Tumorbase

Case #	Tumor Type	Tumor Location	Slice #
1	Meningioma	Left frontal	44
2	Meningioma	Left parasellar	58
3	Meningioma	Right parietal	78
4	Low grade glioma	Left frontal	35
5	Astrocytoma	Right frontal	92
6	Low grade glioma	Right frontal	81
7	Astrocytoma	Right frontal	92
8	Astrocytoma	Left temporal	39
9	Astrocytoma	Left frontotemporal	31
10	Low grade glioma	Left temporal	35

The slices listed in the righthand column of the above table are depicted in Figure 2.7.

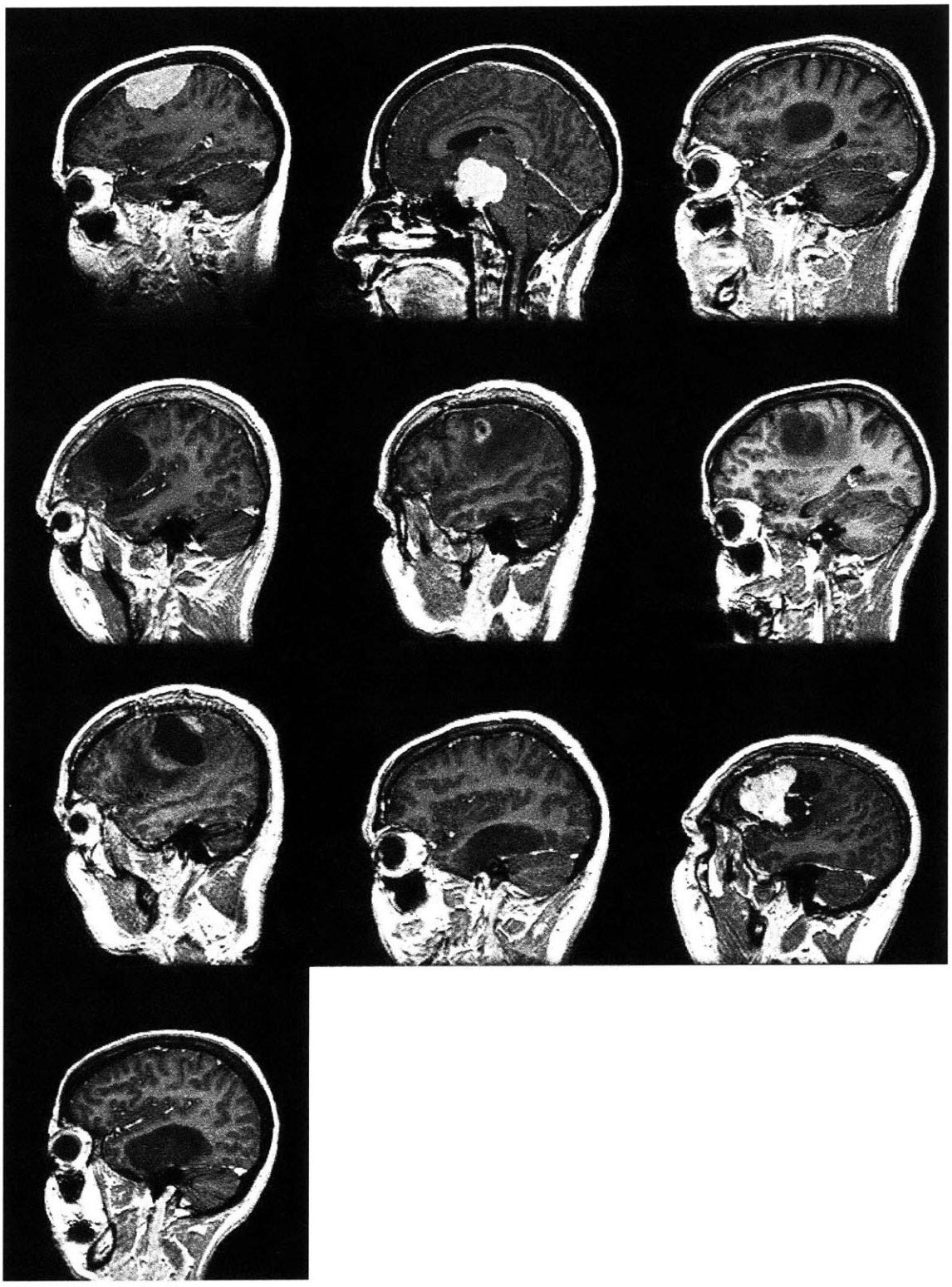

Figure 2.7. Tumorbase The central tumor slice of each of 10 scans

Chapter 3

Recognizing Deviations from Normalcy

In this chapter, we develop the rationale for our unique approach to tumor segmentation. By viewing the problem from a general perspective, we describe tumor recognition as a form of anomaly detection rather than feature detection. By taking this posture, we position ourselves to derive our method for diagonalized nearest neighbor pattern recognition, and also our framework for contextual dependency networks.

3.1 Feature Detection vs. Anomaly Detection

3.1.1 Tumor Segmentation Based on Feature Detection

Much of the related work in tumor segmentation reviewed in Chapter 1 can be classified as signal processing and pattern recognition. Signals, taking the form of imagery, are generally processed through a three-stage pipeline consisting of preprocessing, feature extraction, and classification [Duda01]. Stages are sometimes combined, or applied in iteration, such that intermediate results are fed back into earlier stages for re-processing. Nonetheless, in general, each stage serves to simplify the operations of the subsequent stage.

The first stage, **preprocessing**, simplifies feature extraction by reducing noise or inhomogeneity. Some algorithms perform nonlinear filtering designed to reduce noise while preserving object edges [Gerig92]. We cited several methods in Chapter 1 that correct for the non-uniform bias field present in MRI. Others require scaling images in intensity or extent to match certain templates.

The second stage, **feature extraction**, strives to reduce the amount of data passed on to the classifier. This data reduction is achieved by measuring features, or properties, that characterize the objects to be recognized. The measurements are chosen so that measurement values are similar for objects that share the same class membership, but are quite different for objects belonging to other classes. The goal, then, is to identify features that are both distinguishing, and invariant to irrelevant transformations of the data. Due to their ease of computation, segmentation features are typically intensities and distances.

The third stage, the **classifier**, decides the class membership of each object. While the final segmentation may display the assignment of each object to a single class, the classifier typically solves the more general problem of computing the probability of membership of each object with each class. If the features are ideally chosen to linearly separate the object classes in feature space, then the design of the classifier can be as simple as a threshold. On the other hand, a poorly designed feature extractor requires a more intelligent classifier, as illustrated in Figure 3.1.

Figure 3.1. Features and Classifiers

A common task used in the literature to evaluate a segmentation method is to discern buildings from trees and shrubs. However, consider this photograph from Boston's historic Beacon Hill district. Its sheer complexity suggests a need for an extremely intelligent classifier.

However, if one were to photograph it again in early autumn (after the tree leaves have turned bright yellow while the vine remains deep green), and again in late autumn (after the vine has also lost its leaves), the three images would comprise a *feature vector* of colors. Only a simple classifier would be required to operate on this feature vector because the objects (building, vine, and tree) are easily separable across the new dimension of time.

3.1.2 Tumor Segmentation Based on Anomaly Detection

Existing work in tumor segmentation has tended to reduce the problem to a form of pattern recognition, with a focus on feature extraction. Given this stance, the central question that the algorithm designer seeks to answer is:

1.) "What features will separate tumors from their surroundings?"

Given the answer to this question, the designer subsequently asks:

2.) "What preprocessing is required to facilitate extraction of these features?"
3.) "Which classifier will perform best on this feature set?"

However, the goal of this thesis was to shift the focus from the features to the classifier, and to consider the problem not just as pattern recognition, but within the more general scope of artificial intelligence. Consequently, we replaced the above questions with the following:

1.) "How does a doctor recognize tumors?"

While answers may vary, we believe that a doctor's knowledge of normal anatomy permits recognition of any form of pathology. As before, the answer to the first question leads us to two follow-up questions:

2.) "What is normal?"
3.) "How is abnormality measured?"

These are the two questions on which we will focus in considerable detail as we develop our framework for a tumor segmentation system.

3.2 Deviations from Normalcy

3.2.1 Expressing Abnormality

Given a univariate, normally-distributed, random process, the answers to our two guiding questions are straightforward: normalcy is defined as the population mean, and abnormality is measured as some distance from the mean. The units of measurement for this distance should be standard deviations because a Gaussian process is fully characterized by its mean and standard deviation. For variable x with mean μ and standard deviation σ, expressing distance in this way is commonly known as the Mahalonobis distance:

$$d_1 = \sqrt{\frac{(x-\mu)^2}{\sigma^2}} \tag{3.1}$$

Next, consider a multivariate process of n independent variables. Like a Euclidean distance for Cartesian space, abnormality can be expressed as the square root of the sum of squared Mahalonobis distances for each variable:

$$d_n = \sqrt{\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} + \cdots + \frac{(x_n-\mu_n)^2}{\sigma_n^2}} \tag{3.2}$$

Finally, consider a multivariate process of correlated variables. The expression for abnormality begins as above, but contains additional cross-terms under the radical. Combining the variances and covariances into a convariance matrix $\boldsymbol{\Sigma}$, we have:

$$d_n = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})} \tag{3.3}$$

With medical applications, however, access to all variables is rarely obtainable. For example, physical health could be expressed as a single quantity using the above equation for distance from normalcy. Such a distance could be computed from the set of status and DNA contents of each cell, yet the normalcy of newborn babies is merely expressed with the five non-invasive measurements of the Apgar Score [Sears93]: heart rate, breathing effort, color, muscle tone, and response to stimulation. That is, all the possible axes of variation are reduced to a very small and manageable feature set.

This analogy shares two similarities with MRI. First, we do not have access to the complete condition of the brain; we have only the measurements expressed as the intensities of the image voxels. Brains do not have voxels; images do. Given that the image itself is a non-ideal representation of the brain, it is reasonable to consider further representational abstractions for convenient computation. Second, all the axes of variation can be compressed into a small and manageable set, which we will explore next.

We can regard an MR image as a set of voxels that specify the Cartesian coordinates of a point with respect to a set of axes – one axis per voxel. In this interpretation, each image can be thought of as a point in an abstract space of images. A set of N images represents a cloud of N points in *image space*. We can perform data dimensionality-reduction by deriving a set of degrees of freedom which may be adjusted to reproduce much of the variability observed within a training set. (Informally, imagine creating a small set of knobs which may be turned to generate reconstructions of all the image instances.)

Brains, being similar in overall configuration, will not be randomly distributed throughout a huge image space, and thus can be described by a relatively low dimensional subspace. For example, consider having a stack of brain images that could be ordered in such a way that when viewed in rapid succession, they formed a nearly seamless movie. Whenever this is achievable, then those images lie along a continuous curve through image space. Generating the entire sequence of images can be achieved by altering only one degree of freedom, the curve's parameterization. That is, brain variability is reduced to a one-dimensional curve that is embedded in a high-dimensional image space, where the number of dimensions is equal to the number of voxels per image. By reducing the data dimensionality of normal brains to one, the expression of abnormality becomes simple: the distance from the curve. When one dimension is not sufficient to capture an adequate amount of variability, several may be used, producing not just a curve, but a surface or manifold in image space. We next examine very briefly how to discover such a manifold.

While newborn measurements were chosen partly for convenience, the axes of variability for brain images can be found automatically given a training set. There are several mathematical methods [Chatfield80, Turk91, Bregler95, Hinton95, Basri98,

Tenenbaum00, Roweis00, Cox01] that can discover the underlying structure of brain images (different from that of cardiac images, for example) in order to map a given data set of high-dimensional points into a surrogate low-dimensional space:

$$\mathbf{X} \in \Re^{D} \Rightarrow \mathbf{Y} \in \Re^{d}, \quad d << D \tag{3.4}$$

For example, Principle Component Analysis (PCA) replaces the original variables of a data set with a smaller number of uncorrelated variables called the *principle components*. If the original data set of dimension D contains highly correlated variables, then there is an effective dimensionality $d < D$ that explains most of the data. This representation has two advantages. First, the fact that the new variables are uncorrelated means that equation 3.2 can be used instead of equation 3.3. Second, the presence of only a few components of d results in more efficient computation, and it makes it easier to label each dimension with an intuitive meaning, such as "height". The earliest descriptions of PCA were presented in [Pearson1901] and [Hotelling33], and we refer the reader to [Gering02b] for detailed derivations and comparisons of both linear and non-linear data dimensionality methods.

3.2.2 Partitioning Abnormality

To summarize the discussion thus far, we have concluded that computing the Mahalonobis distance using every MR image voxel would be too cumbersome, and we therefore wish to reduce the data dimensionality. However, we cannot simple run PCA on a vast training set of brain images because we are not seeking to measure the total abnormality of a brain. Rather, we aim to recognize the abnormal tissue within a brain, and label those areas as pathology. Thus, our goal is to partition the space into healthy and diseased regions.

Partitioning can be achieved through concentrating on local image patches. If we divide the brain into a large number of sub-regions, PCA (or a similar variant) could be performed on each local patch. However, this approach faces the two hurdles of somehow reconciling a given brain sample with some appropriately chosen subdivision process, and training on an extensive set of brain imagery. How should the image be subdivided into local patches? We will answer this question during our development of nearest-neighbor pattern matching in the next section.

3.2.3 *Defining Normal using Symmetry*

Throughout the above discussion, answers to the questions of what is normal, and how to measure abnormality, were dependent on possessing a training set of example instances of normal images. In the absence of an extensive training population, a definition for normal can be derived from an exploitation of symmetry. For example, it has been proposed that computer-aided diagnosis algorithms for detecting breast and respiratory lesions could exploit left/right symmetry to define normal as the healthy breast or lung. (See [Giger00] for a survey article). In practice, however, texture from a single healthy breast has been insufficient to capture all the variability, requiring a training set of many scans. We perform experiments here to judge how well normal brain anatomy can be defined as the healthy hemisphere. The problem of recognizing brain tumors may be better suited to exploiting symmetry because the application is for treatment planning rather than screening. Consequently, while breast tumors can appear minutely small on a routine screen, brain tumors tend to not be scanned until their size has grown sufficiently large to become symptomatic.

With symmetry providing examples of normal texture, abnormality can be measured using an appropriate distance metric such as the sum-of-squares distances for a Euclidean space. This leads us naturally to the method of nearest neighbor pattern recognition, developed below.

3.3 Nearest Neighbor Pattern Matching

In this section, we experiment with applying nearest neighbor pattern matching (NNPM) to segmenting brain tumors. This method forms the basis of an initial study for measuring deviations from normalcy in our application. The results represent a baseline against which we can benchmark the more sophisticated methods developed during the remainder of this thesis.

The main idea is to compute a map of the probability of pathology, and then segment this map instead of the original input intensity image. Alternatively, the map could be used as a feature channel in an existing tumor segmentation method, such as [Kaus00]. Figure 3.2 illustrates the concept of segmentation based on an abnormality map computed as the set of Mahalonobis distances.

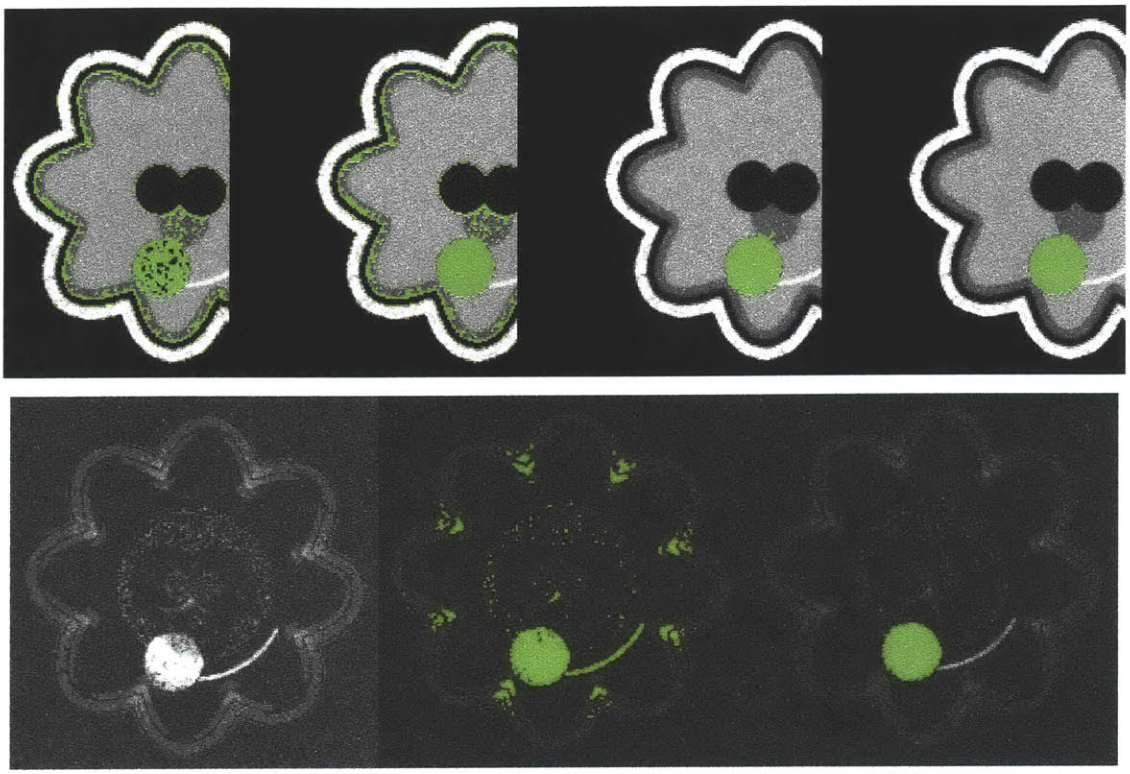

Figure 3.2. Segmenting Abnormality Maps instead of Intensity Images. Top: basic semi-automatic segmentation steps applied in sequence to an intensity image. From left to right: threshold, internal island removal, external island removal, erosion/dilation. Bottom: same sequence of steps applied to the map of abnormality computed using NNPM with a database of 300 normal images.

3.3.1 NNPM Algorithm

As diagrammed in Figure 3.3, a simple pattern matcher can be constructed from two elements: a container and a comparator. The container holds a set of template patterns, and the comparator computes a distance value, according to an appropriate metric, between each template and the sample under study. The template with the smallest distance is the nearest neighbor to the sample. Classification can be accomplished with NNPM by classifying the sample by assigning it the label associated with its nearest neighbor [Duda01]. We will adapt NNPM for use as a means of measuring deviations from normalcy.

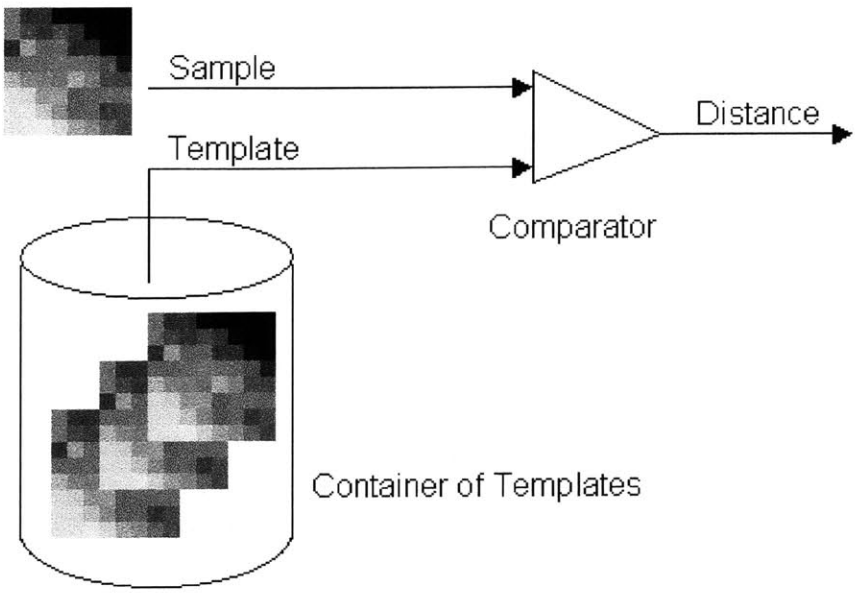

Figure 3.3. NNPM Pattern Matcher

For our application, define a sample to be a small rectangular window surrounding a certain voxel of the patient's image. Let there be a different container C_i of templates T_j for each sample S_i in the patient image. Then perform the following algorithm:

```
For each sample Si in the patient image:
  For each template Tj in container Ci:
    Compute disparity between Si and Tj
  Record the lowest distance as pixel i of the result
```

We next consider how NNPM can be used to answer our two guiding questions of what is normal, and how to measure abnormality.

3.3.2 Measuring Abnormality with NNPM

Let us express the above algorithm mathematically. The method searches for the template with the smallest distance:

$$d_i = \min_{j \in C_i} d_{ij} \tag{3.5}$$

We next need to define d_{ij}: the distance between the i^{TH} sample in the image, and the j^{TH} template in C_i. If we were to treat each variable within a window as independent, we

could adapt equation 3.2. Then, in place of the mean value representing "normal" in equation 3.2, we use the reference value. Instead of normalizing with standard deviations, we normalize with window size W to accommodate comparing the results achieved using various window sizes. These substitutions result in the following equation, which is essentially the root-mean-squared error. Let $S_i[k]$ represent the k^{TH} voxel of the i^{TH} sample, and let $T_j[k]$ represent the corresponding voxel in the j^{TH} template.

$$d_{ij} = \sqrt{\frac{\sum_{k=1..W}(S_i[k]-T_j[k])^2}{W}} \tag{3.6}$$

Combining the above two equations produces a mathematical expression of the algorithm, given our metric for measuring abnormality:

$$d_i = \min_{j \in C_i} \sqrt{\frac{\sum_{k=1..W}(S_i[k]-T_j[k])^2}{W}} \tag{3.7}$$

3.3.3 Defining Normal with NNPM

NNPM defines normal as the set of templates in each container C_i. Each template is an example of normal texture that one would expect to find within the window of W pixels surrounding the i^{TH} voxel of the patient's image. Since no probability distributions are fit to these templates, building collections of them is straightforward. However, enough templates must be gathered into each container to sufficiently span the space of normal variation within a window, and none must be examples of abnormal texture near voxel i. This can be a significant task given that the variation within a window is comprised from variation in both anatomy and the bias field. The next few paragraphs examine how to fill these containers.

Consider the simple case of defining all C_i to identically contain all windows within a reference image of a healthy brain. The algorithm would effectively search an entire reference image for the template window that best fits a given window in the patient image. However, by searching the *entire* reference image, spatial information – the location of voxel i – is ignored. For example, if the reference image contained a dark window anywhere, then the algorithm would consider any dark windows in the patient

image to be permissible. However, it should be considered abnormal to find a dark window where one would expect a light window, so this approach fails as a search for deviations from normalcy.

Therefore, a more plausible choice of C_i would be the window surrounding the one voxel of the reference image that exhibits the best correspondence with voxel i of the patient image. Correspondence would need to be established by defining a mapping from voxels in the patient image to voxels in the reference image. Such a mapping could be computed as a linear or affine transform using rigid registration, or as a polynomial function or vector displacement field using non-rigid registration. Either way, robustness to registration errors could be introduced by expanding C_i to include all windows centered around the small set of neighboring voxels surrounding the one voxel with the best correspondence. The algorithmic time complexity would then be $O(NMW)$, where N is the image size, W is the window size, and M is the neighborhood size, and $M, W < N$.

How well does a single reference image capture the extent of normal variation within a population? The sample on the left of Figure 3.4 looks little like the reference on the right. With this thought in mind, perhaps a better approach to defining C_i would involve not one reference image, but a set of images that have been selected to be representative of the complete population. Call this the *training set* of images, and define C_i to include all templates defined as follows:

- For each image t of the training set:
- For the one voxel j in image t that exhibits the best correspondence with voxel i of the patient's image:
- For each voxel k in the neighborhood $\{j_N\}$ surrounding j:
- Create a template as the window $\{k_W\}$ surrounding voxel k.

The time complexity of this algorithm scales linearly with the training set size: $O(NMWT)$. Figure 3.4 illustrates the difference between using a single reference image, and an extensive training set. Observe that the larger atlas alleviates the need for a larger search neighborhood. No search neighborhood is as good as a more complete atlas, especially for expressing concepts such as the vessels which rarely appear in exactly the same place on any two scans, but always occur in the same general area.

Figure 3.5 presents a measurement of the algorithm's reliance on all the images in the training set. A spatial map was generated by setting each voxel's value to the index of the atlas image (1-300) where the nearest neighbor was found. For example, if all the nearest neighbors had been found in the same atlas image, the spatial map would appear as a constant gray. Instead, the map appears quite speckled. The map on the right is less homogenous than the map on the left because the search space was expanded to include the 9x9 neighborhood the best corresponding pixel of each image in the atlas (instead of just 1x1). Note how the tumor is conspicuous by its homogeneity – testifying to its distance from the cluster of healthy atlas patches.

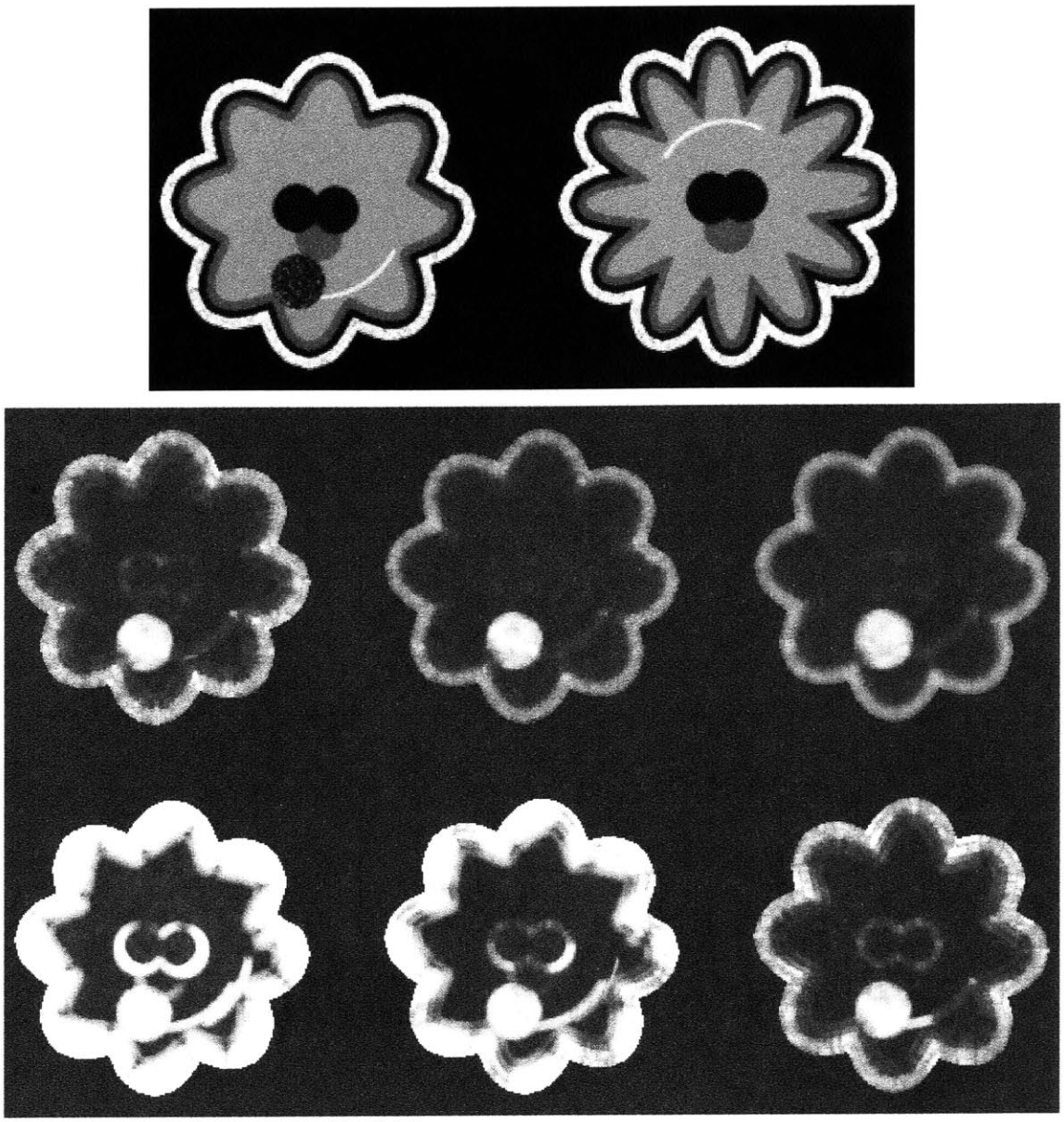

Figure 3.4 Atlas Size and Search Space. (Top:) The "sample" image is on the left, and one "reference" image is on the right. (Middle:) Results of running NNPM on the "sample" using an atlas of 300 scans. From left to right, are the results of searching a square neighborhood around the best corresponding pixel with radius 0, 2, and 16. (Bottom:) Inferior results of NNPM using the single "reference" image instead of an atlas of 300.

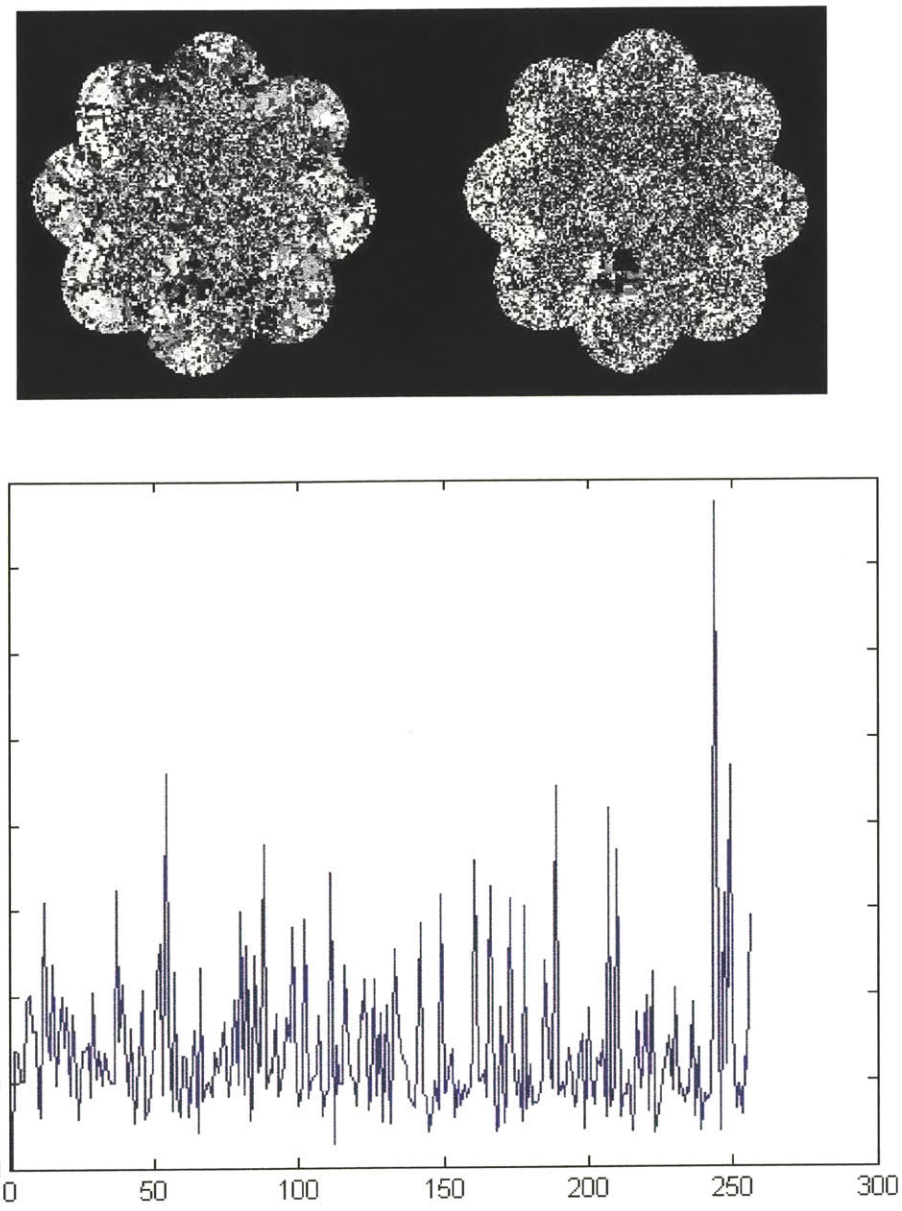

Figure 3.5. Nearest Neighor Distribution within an Atlas of 256 Scans. (Top:) A spatial map was generated by setting each voxel's value to the index of the atlas image (1-300) where the nearest neighbor was found. On the left, is the result of using 1x1 neighborhoods, and the right is the result of searching 9x9 neighborhoods. (Bottom:) Histogram of indices for the top left image demonstrating the breadth of the distribution.

3.3.4 Selecting Window Size

Consider selection of the window size W. For the foregoing discussion, define micro-texture to refer to the normal intensity patterns found over small regions, and macro-texture to refer to the patterns spread over large areas.

The optimal choice of window size is application-dependent, as it varies with the interplay between micro- and macro-textures. Selecting a small window size would be adequate to incorporate the context necessary to recognize normal micro-texture, and run times would also be favorable. Large windows, on the other hand, would have the advantage of capturing macro-texture, but they would situate the micro-texture within the macro-texture. That is, if a certain micro-texture pattern could normally be found anywhere, than enough macro templates would be required to express this fact by exhibiting the certain micro texture in various situations. Thus, the run-time of the algorithm that correctly uses large window sizes would be dramatically lengthened for two reasons: more time is required to process larger windows, and more template windows are required to encode more situations. We will refer to this as the double trouble with large window sizes.

One way to handle this dilemma would be to isolate the searches for micro- and macro-texture. This will be our goal in the next two subsections, as we derive our novel *diagonalized NNPM.*

3.3.5 Multi-scale NNPM

As we seek a means to somehow isolate the searches for micro- and macro-patterns, we acknowledge that there has been much experience within the computer vision community with multi-scale algorithms. We employ such a tactic in Chapter 4, for example, when we automatically align patient images to atlas images by maximizing mutual information [Wells96a]. Our implementation applies the same algorithm to several different resolutions of the input data. The objective of this approach is for greedy algorithms to have greater scope to avoid local minima, as well as faster convergence toward a solution. Coarse solutions can be reached very quickly given an input data size that is merely a small fraction of the original. Then, finer processing can refine the coarser solutions using progressively larger input data sizes.

For our purposes within this chapter, we seek to exploit multi-scale computation not to aid greedy searches or minimize time to convergence, but rather to separate micro- and macro-texture. When the input data set is downsampled to halve the size of each dimension, 3-D computation with the same window size proceeds 8 times more quickly, and incorporates context from a region 8 times larger. More importantly, at progressively smaller image dimensions, micro-textures become blurred out, allowing the computation to concentrate on macro-textures alone. Figure 3.6 displays one of our synthetically-generated brains at multiple resolutions.

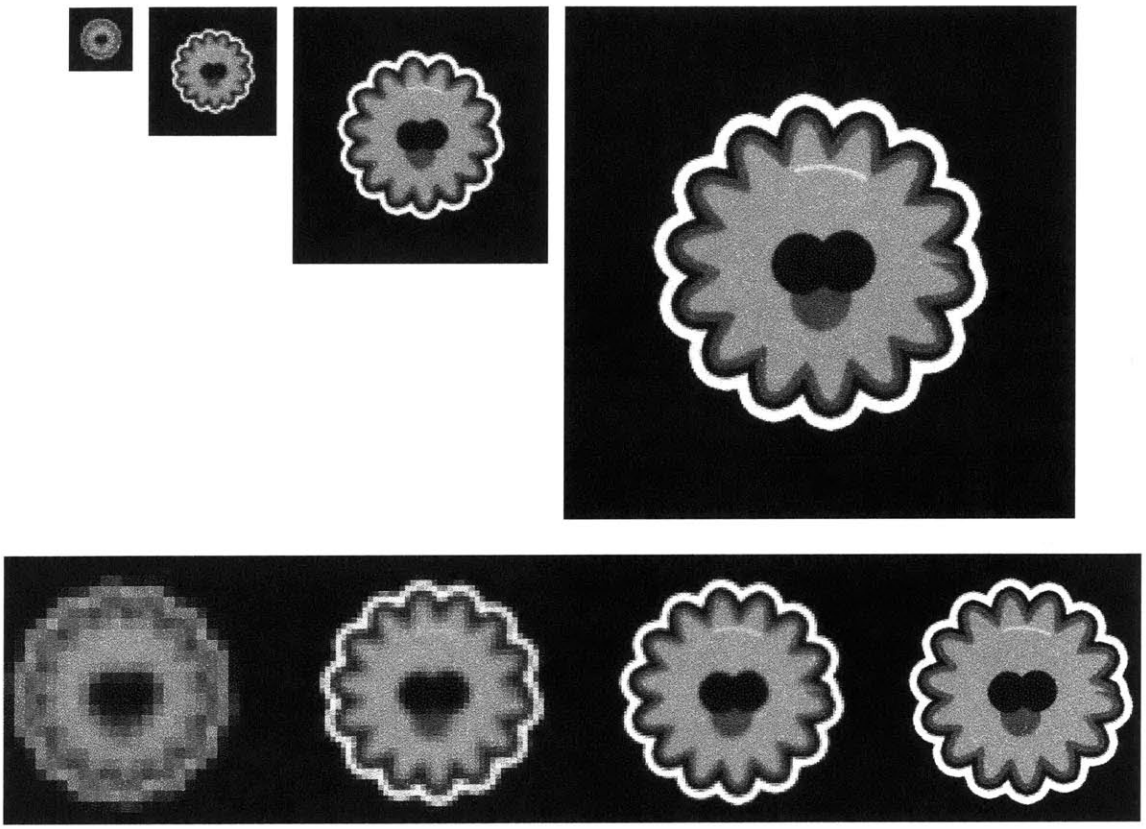

Figure 3.6. Multi-scale Computation. The top row displays each downsampled image at actual size, while the bottom rows displays the same images scaled for equal comparison of detail. At small scale (left), note the disappearance of micro-texture (vessels) and preservation of macro-texture (CSF divides scalp from white matter).

Downsampling must be performed properly to avoid the artificial introduction of spurious features, as shown in Figure 3.7. This is the purpose of scale-space theory, and in particular, the scaling theorem. Multi-scale analysis for extracting features from a continuum of scales was initiated by [Rosenfeld71], and followed by the well-known

work of Ellen Hildreth and David Marr [Marr80]. The scaling theorem arose when [Witkin83] analyzed zero crossings over a range of scales simultaneously by plotting the zero crossings of a Gaussian-smoothed signal over a continuum of scales. The resulting contours form either lines or bowls as the scale progressed from small to large. Thus, the transformation from a fine scale to a coarse scale can be regarded as a simplification. Fine-scale features disappear monotonically with increasing scale such that no new artificial structures are created at coarser scales. Otherwise, it would be impossible to determine if coarse-scale features corresponded to important fine-scale features, or artifacts of the transformation. In what is known as the scaling theorem, [Koenderink84], [Bebaud86], and [Yuille86] each proved that the Gaussian kernel uniquely holds this remarkable property.

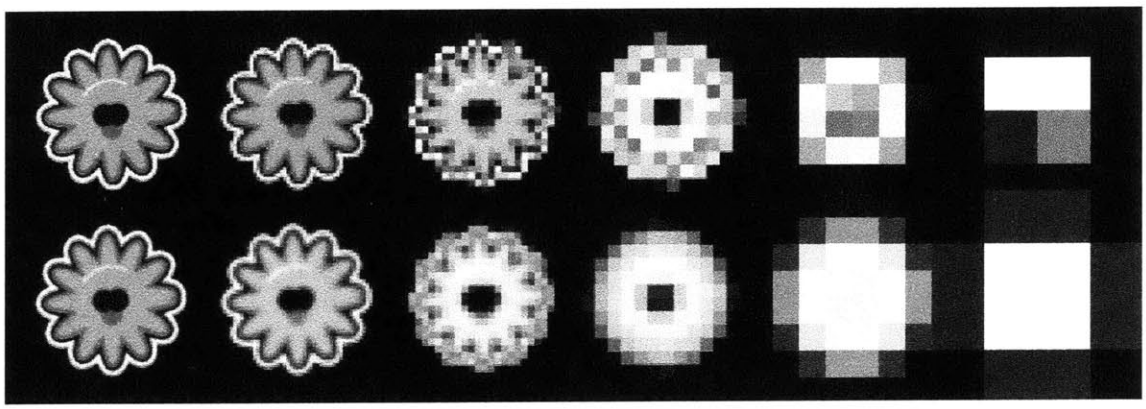

Figure 3.7. The Scaling Theorem. From left to right, progressive downsampling of an image. The bottom row depicts results using Guassian smoothing, while the top row does not. Observe the introduction of high-frequency spurious features in the third image from the left, top row.

3.3.6 Diagonalized NNPM

All that remains in completing our derivation of multi-scale NNPM is some means of combining the results found using fine and coarse scales. The output of NNPM is a spatial map of distances from normalcy. We create a probability of pathology by normalizing this map to scale from 0 to 1. Let us define the following:

P(A) = probability of pathology at the highest resolution

P(B) = probability of pathology at intermediate resolution

P(C) = probability of pathology at the lowest resolution

P(A,B,C) = probability of pathology

Operating on the assumption that using multiple scales is successful in isolating micro- and macro-texture, we treat the probabilities of pathology at each resolution as if they were independent. (Although not true in practice, we make this assumption for tractability.) Thus, we can combine the results obtained at each resolution by scaling each result to become a probability map, and then multiplying all the maps:

$$P(A,B,C) = P(A)P(B)P(C) \tag{3.8}$$

Finally, we must determine the value of the window size parameter, *W*. Imagine a matrix with a vertical axis of image resolution, and a horizontal axis of window width (2*r+1). Figure 3.8 arranges the resultant images from running NNPM into such a matrix. Instead of using identical window sizes at all scales (such as the red oval in figure indicates for a window radius of 2), we will prove that the diagonal blue oval is a better choice for us. We label this algorithm, where the window size increases monotonically with decreasing resolution, *diagonalized NNPM.*

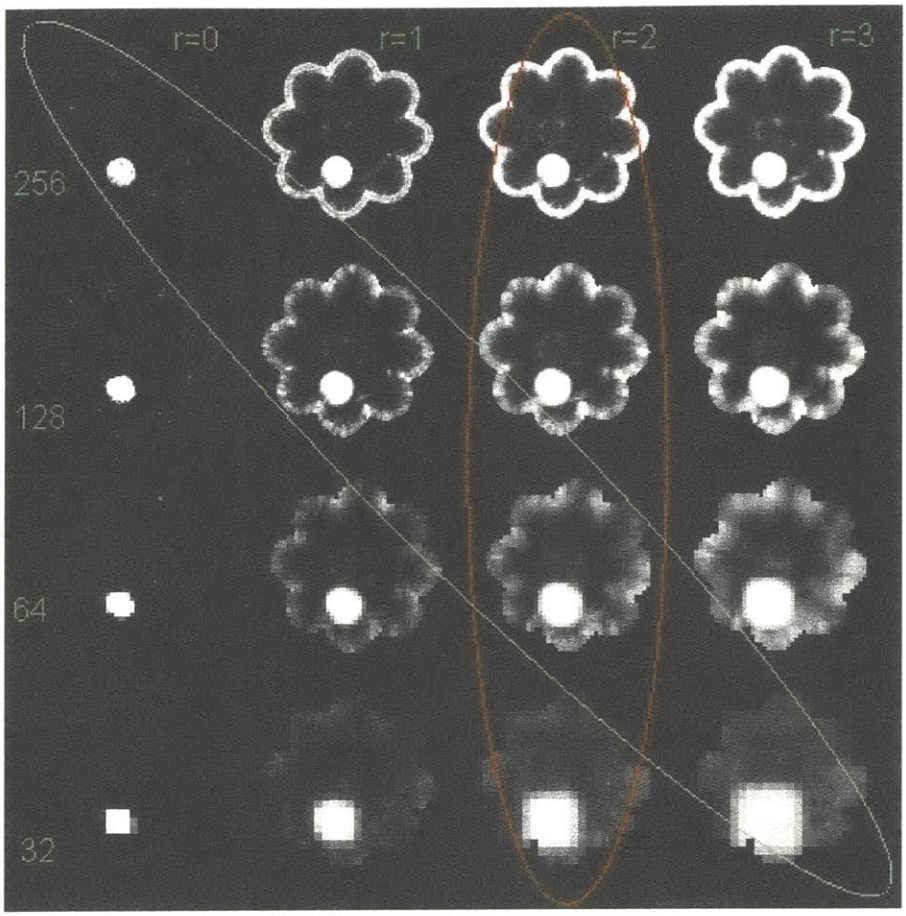

Figure 3.8. Diagonalized NNPM. The red oval represents basic multi-scale NNPM for a window size with radius 2, while the blue oval depicts diagonalized NNPM.

Statement:

In the Diagonalized NNPM algorithm, window size increases monotonically with decreasing resolution, resulting in larger windows at coarser resolutions.

Reasoning:

- Diagonalized NNPM combines the results obtained at each resolution by scaling each result to become a probability map, and then multiplying all the maps:

$$P(A,B,C) = P(A)P(B)P(C) \tag{3.9}$$

- The validity of this operation depends on the independence of each map.
- The independence of each map depends on the separation between micro- and macro-texture.
- Micro-texture is most isolated with a small window so that the Gaussian smoothing obscures the micro-features.
- Macro-texture is most isolated with a large window so that a given micro-feature within the window cannot exert a significant influence in the calculation of abnormality (equation 3.7).
- Thus, multiplicative combination of the maps is best achieved with window sizes that increase with coarser resolutions.

QED

Figure 3.9 demonstrates empirical results of applying this theorem to the synthetic data from Figure 3.4. Although it is dangerous to compare images that have been manually segmented and window/leveled, we would like to make an observation, regardless. The non-diagonalized result contains artifacts and an artificially larger tumor because the boundaries are more blurred. This is a consequence of failing to isolate the fine structure of boundary localization from the coarse structure of general tumor presence.

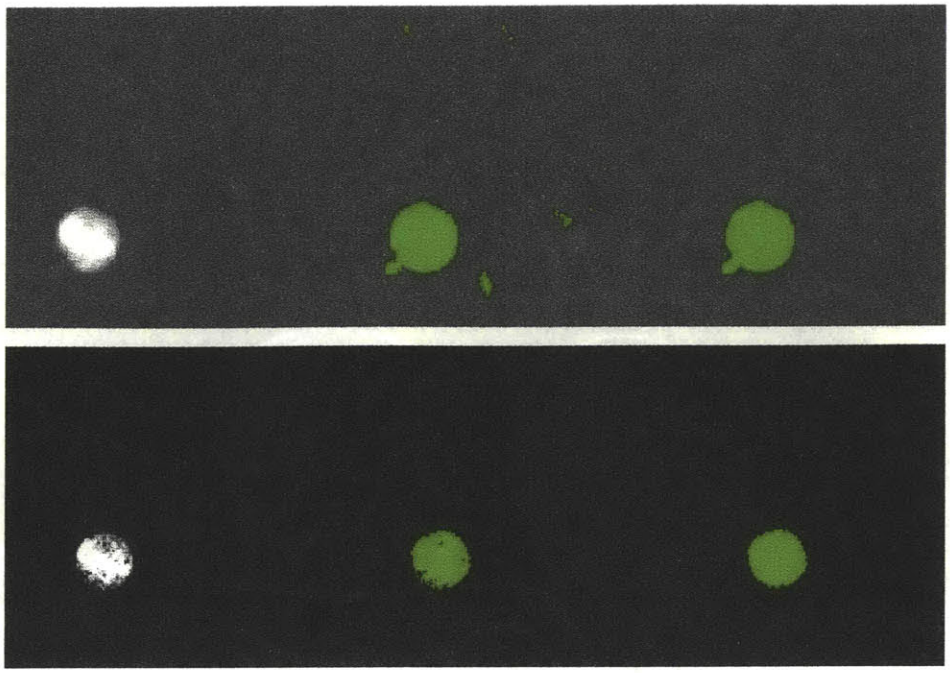

Figure 3.9. Diagonalized NNPM. The top row of images uses a probability map for pathology computed using the red oval in Figure 2.9, while the bottom row uses the blue oval corresponding with Diagonalized NNPM. From left to right, the 3 images are the map itself, segmentation using a threshold, and final segmentation following basic morphological operations.

3.3.7 NNPM Results on Real Data

We performed experiments by running diagonalized NNPM on every case in the tumorbase in addition to a healthy volunteer. The depicted results were generated by defining normal as the two best corresponding slices from the healthy hemisphere of the same patient. The diagonalization is performed using the following set of window radii from fine to coarse resolution: {1, 1, 2, 2}. The segmentation is performed fully automatically by applying a threshold to the 1% level of the map, and then keeping the largest island in the intracranial cavity. The layout of each of the next several figures is as follows:

- Left: Diagonalization matrix (same format as Figure 3.8)
- Upper right: Single abnormality map computed from the diagonalization matrix
- Lower right: Segmentation computed from the abnormality map

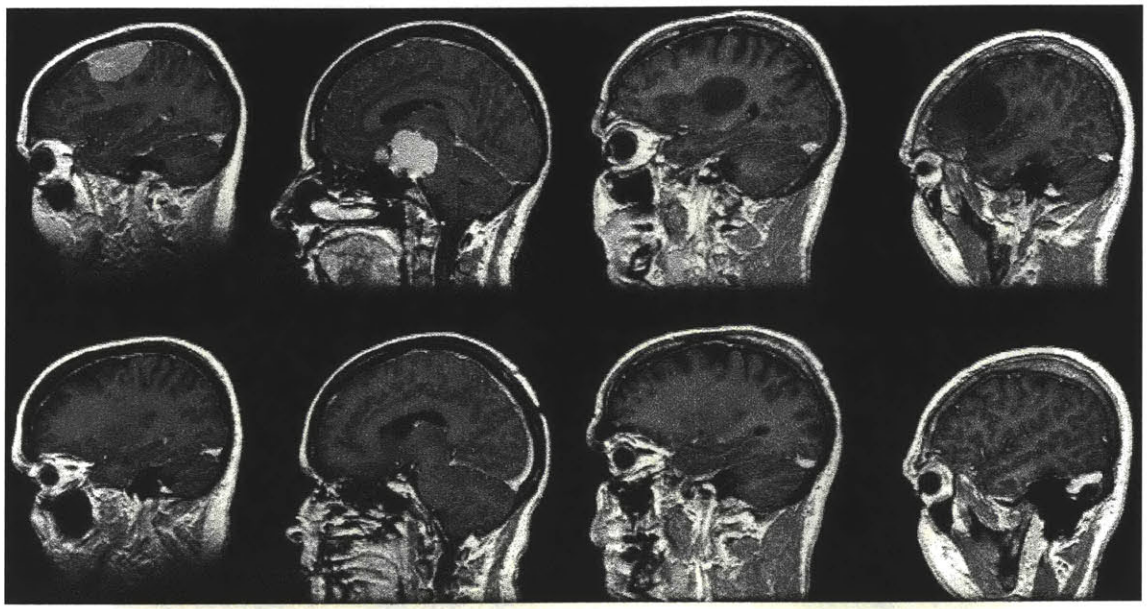

Figure 3.10. Defining Normal by Symmetry. For the first 4 cases in the tumorbase, the top row shows the central slice of the tumor, and the bottom row shows the corresponding slice in the other healthy hemisphere of the same patient.

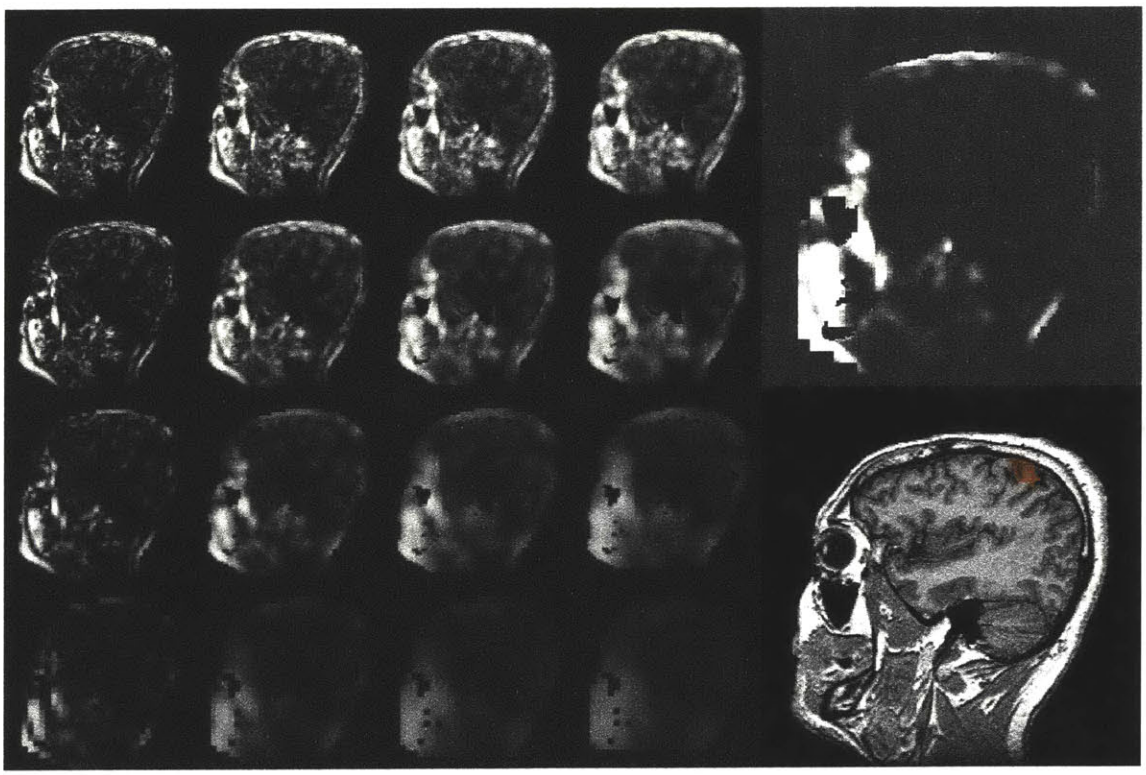

Figure 3.11. Healthy Volunteer. Mostly successful, although the fixed threshold detected a variation in cortical sulci.

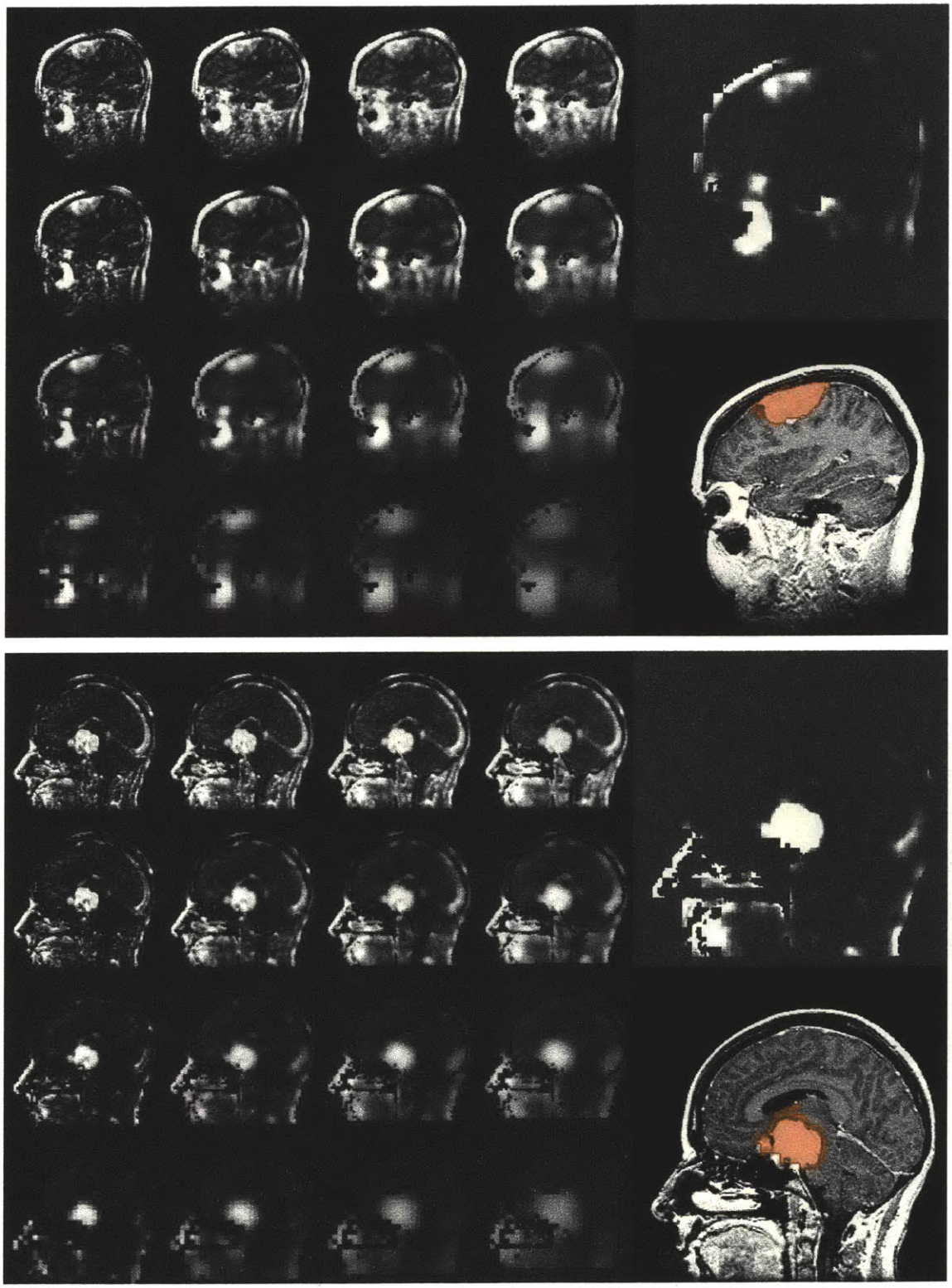

Figure 3.12. Meningiomas. Case 1 (top) and 2 (bottom) have hypointense tumors that are easily recognized as abnormal. Perfect boundary delineation needs user interaction.

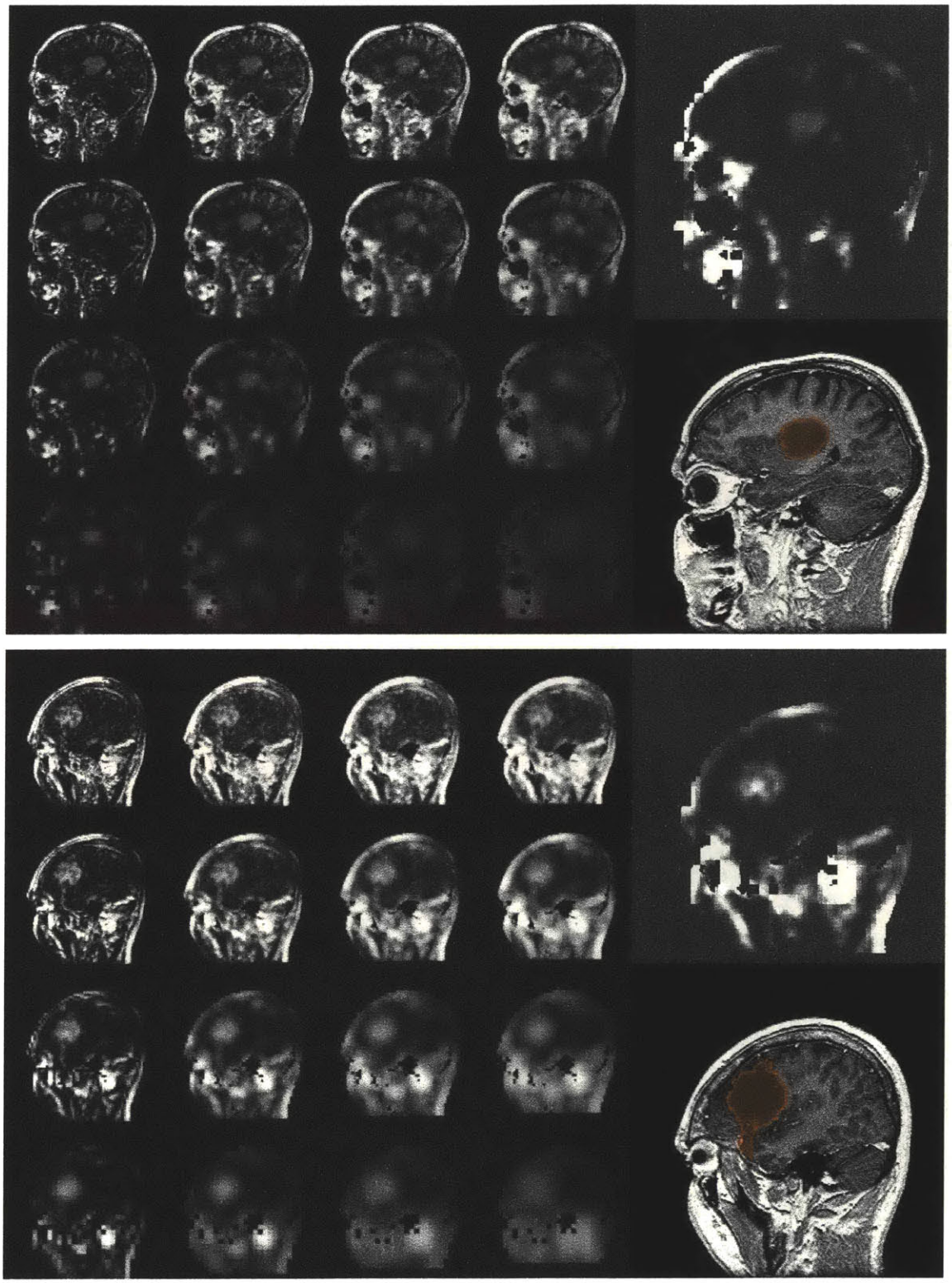

Figure 3.13. Low Grade Glioma. The hypointense tumors of cases 3-4 are segmented as well as the hyperintense ones, displaying the advantage of not training on tumors.

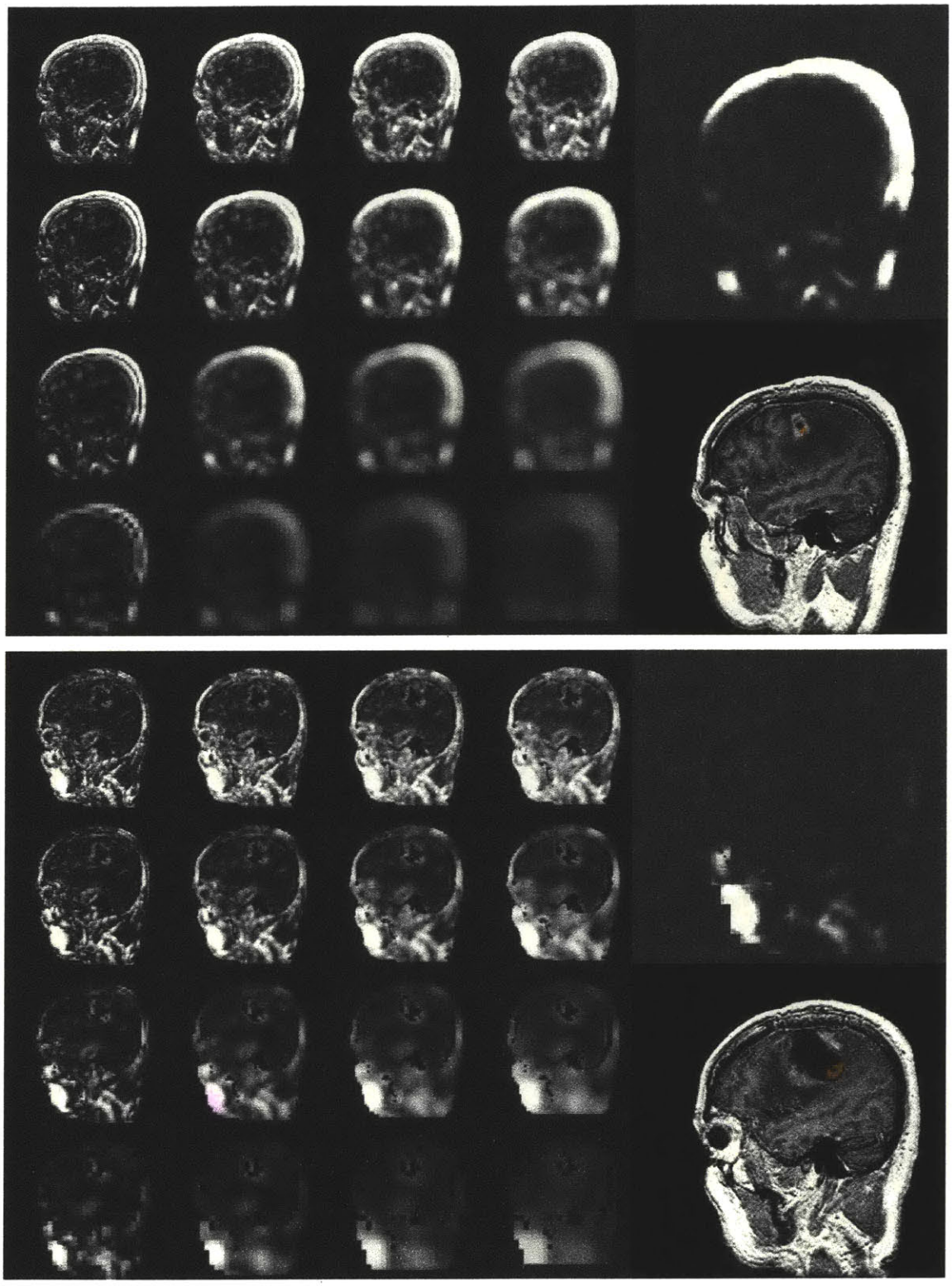

Figure 3.14. Astrocytomas. Cases 5 and 7 failed to produce suffient abnormality to cross the fixed threshold used for automatic segmentation of all cases.

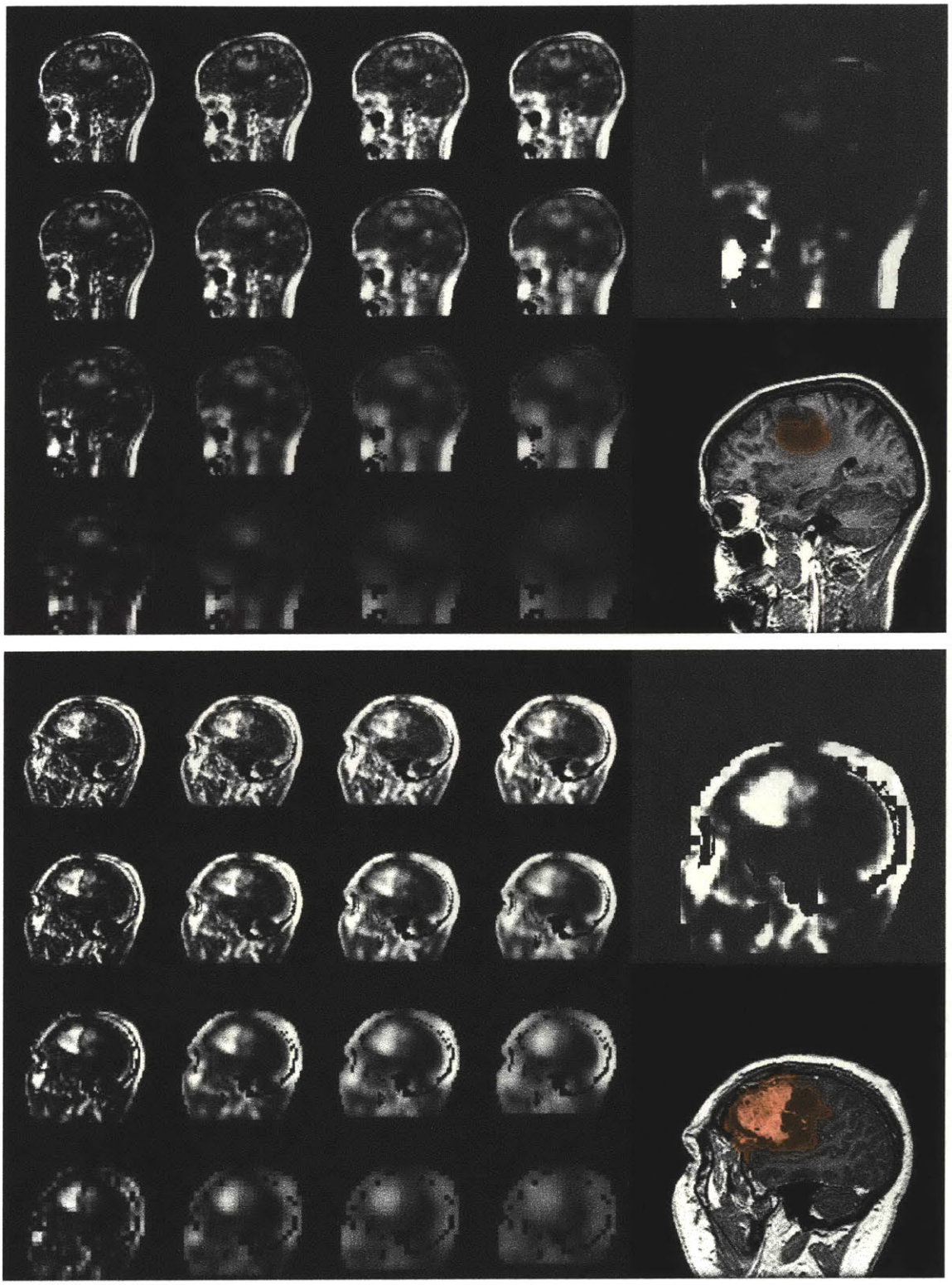

Figure 3.15. Heterogeneity. Cases 6 and 9 have very heterogenous tumors. Recognition of the entire tumor is possible on certain cases, which is at least superior to thresholds.

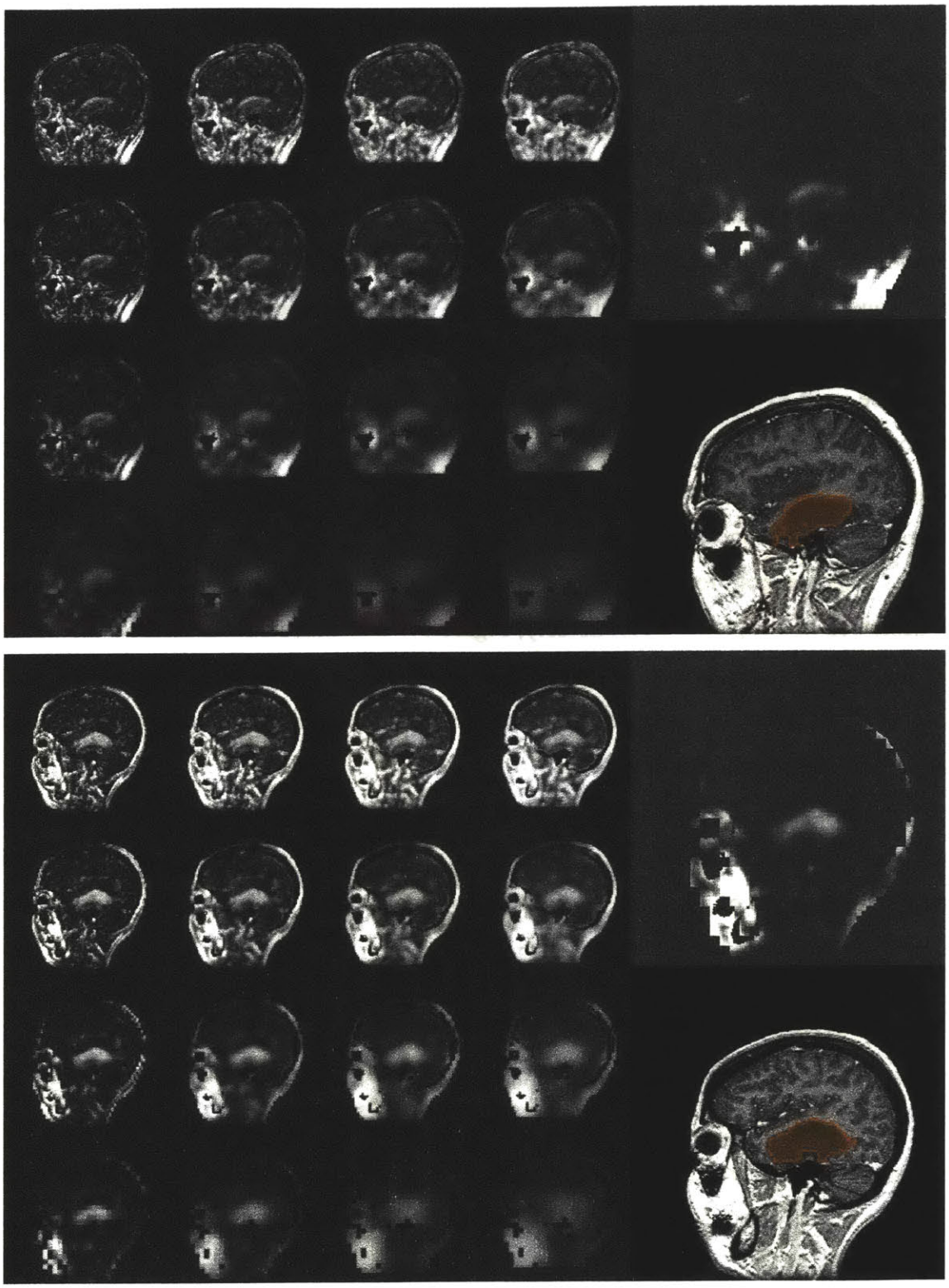

Figure 3.16. Cases 8 and 10 are typical of the fairly good results with lowgrade gliomas.

3.3.8 Discussion of Results for Diagonalized NNPM

In every one of the real data cases, the results of fully automatic segmentation using diagonalized NNPM are too inaccurate for clinical use. Regardless, the results are encouraging given the goal of this thesis, which is to solve the recognition problem for brain tumors. As described in Chapter 1, existing methods have largely focused on boundary delineation, leaving the recognition task for humans. With the exception of case #7, diagonalized NNPM correctly recognized the tumor well enough to initiate the boundary delineation process using one of the existing methods. For example, NNPM could be used to define a region of interest for applying a threshold, a seed point for region growing, or an initial boundary contour for curve evolution. Together, diagonalized NNPM and these methods can form an end-to-end solution for automatic recognition and delineation of brain tumors.

There is room for improvement following our initial experiments, and future work is described in Chapter 7. Most notably, Figure 3.4 demonstrated that remarkably better results can be achieved with synthetic data when a training set of 300 scans are used instead of 1. Meanwhile, our real data experiments were performed using only 2 slices from the healthy hemisphere.

Even with diagonalization, NNPM, as we have implemented it, is an imperfect solution to the simultaneous incorporation of context at all possible scales. We will attempt to improve on this shortcoming with our development of contextual dependency networks in the next section.

3.4 Contextual Dependency Network

The goal of this section is to build on our introduction of diagonalized NNPM to derive our Contextual Dependency Network (CDN). In applying multi-scale NNPM, we encountered the same frustrations – manifested as imprecise tumor boundaries – as described by [Stansfield80] in an MIT AI Lab project to create an artificial commodity expert:

> *Unfortunately, smoothing a graph results in an information loss. While smoothing does highlight large-scale features, the location of their boundaries is obscured.*

What I had hoped for was a series of progressively more abstract descriptions of a graph. The high levels of abstraction would describe only the major features and the lower levels would fill out the details.

3.4.1 Multiple Levels of Context

Recall the results of experimenting with various window sizes, which varied the breadth of the context incorporated by the algorithm. In computer vision, experiments such as this one are typically run to search a parameter space – window size, in this case. After finding the optimal parameter value using a training set, the algorithm is ready to be employed on the sample data sets. However, we discovered that no single window size produces adequate results with NNPM. Moreover, we discovered the double trouble that comes with increasing window size: larger windows imply more windows. This is because incorporating macro-texture also involves situating micro-texture.

Consequently, acknowledging that the primary shortcoming of NNPM is its limitation of being able to consider context on only one level at a time, we explored a multi-scale implementation of NNPM. Our goal was to isolate micro- and macro-texture in order to deal with each independently. However, multi-scale vision does not service all of our needs. We need to incorporate context at multiple levels in a manner conducive to answering our two questions of what is normal, and how to measure abnormality. Multi-scale methods force the coordinate system into the inference processing, but as we referred to earlier, images have voxels, brains do not. In the words of William James, "We must be careful not to confuse the data with the abstractions we use to analyze them." [Rice95] We would therefore rather compute measurements of normality on actual brain structures, such as cortical gray matter, than on some rectangular sub-regions of the image lattice.

3.4.2 NNPM with Non-rectangular Windows

One approach would be to relax the constraint that windows are shaped as rectangles. Then, each container of templates would be occupied by shapes with various sizes and orientations. In the spirit of multi-scale algorithms, the scope of these templates would vary as well. Some would describe detailed structures present at full resolution, while

others would characterize macro properties best analyzed with downsampled images. The determination of non-rectangular windows would be quite application-dependent and complex to train, so we seek another solution.

3.4.3 Hierarchy of Layers

Diagonalized NNPM was shown to possess broad recognition capabilities, but poor precise boundary localization. We seek a new system that meets both requirements, so we propose a solution with multiple levels: some for breadth, and some for precision. Beginning with the smallest possible region, and extending outward, we propose considering the levels of context listed in Table 3.1. The rightmost column lists the definitions of normalcy associated with each level. Our central argument in favor of such a framework is how conveniently these definitions accommodate reasonable answers to our two guiding questions.

Table 3.1. Levels of context that accomodate answering the two questions of what is normal, and how is abnormality measured.

#	Level of Context	Meaning of Context	Characterization of Normalcy
1	Voxel *(point)*	Intensity	Gaussian distributions over voxel value intensity.
2	Neighborhood *(local)*	Compatability	Gibbs distributions over compatability.
3	Intra-structure *(region)*	Shape	Gaussian distributions over shape descriptors, such as relative position of a voxel within its own structure.
4	Inter-structure *(global)*	Situation	Gaussian distributions over situational descriptors, such as relative position of a voxel's structure to other structures.

Ambiguity necessitates the incorporation of contextual information into the brain segmentation process. Consider the example of non-enhancing tumor tissue that mimics the intensity of healthy gray matter, but is too thick to be gray matter. The lowest level of context could first classify the tissue as gray matter, and a higher-level stage – through its

broader understanding of context – could correct the classifications of the lower level. Just as a voxel-wise classification must be computed prior to a neighborhood-wise refinement, a voxel's region must be classified before features regarding the size and shape (or other intrinsic properties) of that region can be computed. This is a concept of predicated context, where high-level vision is performed based on aggregated information from low-level vision. Therefore, we organize our levels of context into a hierarchical network, and label it as a Contextual Dependency Network (CDN). Furthermore, to accommodate intelligent interaction with users, we add a fifth layer on top, as shown in Table 3.2. Note that NNPM has difficulty with expressing predicated context. How does one express that edema always borders tumor, but tumors, and subsequently, edema, can be situated almost anywhere?

Table 3.2. A Contextual Dependency Network is a framework that features no decisions made by certain layers that permanently (and perhaps adversely) affect other layers. Information flows between the layers bidirectionally while converging toward a solution. (Rows are reversed in order from Table 3.1 to situate "high-level" layers above "low-level" layers.)

#	Layer	Definition	Our Simple Computation
5	User *(oracle)*	Spatially specific points clicked on by the user on the fly as corrective action.	Mouse clicks trigger re-iteration.
4	Inter-structure *(global)*	Relative position of a voxel's structure to other structures.	Distance from other region boundaries.
3	Intra-structure *(region)*	Relative position of a voxel within its own structure.	Distance from own boundary.
2	Neighborhood *(local)*	Classification of a voxel's immediate neighbors.	Mean Field MRF
1	Voxel *(point)*	Classification based on voxel's intensity.	EM, ML or MAP

3.4.4 Comparing CDN with Multi-Scale Vision

Our levels of context distinguish themselves in several important ways from traditional multi-scale vision, such as segmentation of image texture [Bouman91] or scale-space approaches to mammography [Karssemeijer95]. We have already mentioned that CDN carries greater independence from the coordinate system than traditional multi-scale vision. Moreover, unlike multi-scale vision that applies essentially the same processing at

each level such that the only differences are in resolution and perhaps parameters, CDN encourages entirely different algorithms to be applied at each level. Furthermore, unlike multi-scale vision where processing can proceed each level simultaneously, CDN levels are based on predication. That is, a given level cannot perform its processing until the level beneath it completes its processing. The reason is that the higher level processing is predicated on the lower level output. Finally, multi-scale vision is not designed to be iterated, which implies that information flows only one direction – from lower resolutions to higher resolutions. CDN can iterate to propagate information bi-directionally; after a higher level corrects a lower level's mistakes, the lower levels can be recomputed given their new high-level information. These distinctions are summarized in Table 3.3. In fact, CDN can be implemented in scale space. That is, a certain layer can perform its processing using multiple resolutions of the data.

Table 3.3. Constrasts between multi-scale vision and CDN.

Multi-scale Vision	Contextual Dependency Network
Region definitions are coordinate system dependent	Region definitions are object dependent
Identical processing at each level	Unique processing at each level
Levels can be computed simultaneously	Higher levels are predicated on lower levels
Information flows one direction	Iteration allows bidirectional information flow

3.5 Chapter Summary

The aim of this chapter was to revisit the image segmentation problem in hope of developing a more generally applicable approach. In contrast to treating the tumor segmentation problem as an exercise in discovering distinguishing features, we derived our unique approach for recognizing deviations from normalcy. Beginning with NNPM, we developed a framework for Contextual Dependency Networks that can incorporate context at multiple levels. Subsequent chapters will develop our first implementation of such a framework. This implementation is designed to be a simple proof of concept. Our

hope is that smarter components, when inserted into our framework, will further improve its effectiveness. To summarize the important principles asserted in this chapter:

3.1 For general applicability, tumor segmentation systems could recognize deviations from normalcy, rather than identifying known features of tumors.

3.2 Systems that recognize deviations from normalcy must answer the following two questions:

1.) What is normal?

2.) How is abnormality measured?

3.3 In NNPM, double trouble comes with increasing window size: larger windows imply more windows. This is because incorporating macro-texture also involves situating micro-texture.

3.4 In the Diagonalized NNPM algorithm, window size increases monotonically with decreasing resolution.

3.5 CDN incorporates multiple levels of predicated context as a step toward the goal of achieving recognition capabilities that are both broad and precise.

Chapter 4

CDN Layer 1: Voxel Classification

In this chapter, we introduce the first layer of our framework for a contextual dependency network. The role of the first layer is to produce a preliminary classification of each voxel so that the next layer has a starting point from which to consider immediate context. Without an initial context, the voxels must be considered in isolation, but the only information offered by individual voxels is their intensity. Hence, we seek answers to our two guiding questions of how to define what is normal, and how to measure the degree of abnormality, based only on intensity.

This chapter is organized to review the mathematical background for Bayesian classification and the expectation maximization (EM) algorithm, and then to address the difficulties encountered when applying these techniques to pathological, rather than healthy, brains. Specifically, we modify EM segmentation to avoid confusing the bias field with pathology. Then, we examine spatially varying priors and generalize their concept into probabilistic mappings between image space and model space. We then base the processing for each layer of CDN on these mappings. Next, we develop a method for computing a probability of pathology for CDN Layer #1. Finally, we conclude by evaluating our analytical models by inverting them to produce generative models.

4.1 Mathematical Background for Model-Based Classification

Understanding what is normal involves possessing some model of what should be expected, so we are interested in model-based mathematical techniques. As discussed in Chapter 3, Gaussian distributions handle these questions most elegantly, provided they

are applicable, which was shown in Chapter 2 to be the case for MRI signals with intensities well above the noise floor. We will therefore rely on Gaussian distributions for intensity models, and this section will discuss their application within classifiers.

4.1.1 Bayesian Classification

Bayesian classification provides a probabilistic approach to weighting the evidence supporting alternative hypotheses. The probability of a hypothesis is determined from both the observed data and prior knowledge, and these can be characterized by probability distributions. This prior knowledge can be represented in either, or both, of two ways:

- The **prior** *P(h)* for each candidate hypothesis is the probability of that hypothesis being true *prior* to observing any data *D*.
- The **likelihood** *P(D|h)* of each candidate hypothesis is the conditional probability, or *likelihood,* of the data given the hypothesis. This term is also referred to as the *measurement model* because we can measure it *a priori* in order to construct application-specific models.

Bayes' Theorem provides a quantitative method for computing the posterior probability from the prior and the likelihood:

$$p(h \mid D) = \frac{p(D \mid h)p(h)}{p(D)} \tag{4.1}$$

Using this equation, we can address the classification problem by searching for the *maximum a posteriori* (MAP) hypothesis from the set *H* of all candidate hypotheses:

$$h_{MAP} = \arg\max_{h \in H} p(h \mid D) = \arg\max_{h \in H} p(D \mid h)p(h) \tag{4.2}$$

When the priors are unavailable, or every hypothesis is equally probable, we can instead search for the *maximum likelihood* (ML) hypothesis. This is the hypothesis under which the observed data would be most likely to appear:

$$h_{ML} = \arg\max_{h \in H} p(D \mid h) \tag{4.3}$$

Because the logarithm function is monotonic, we can equivalently maximize the log likelihood:

$$L(h) \equiv \log p(D \mid h) \tag{4.4}$$

This is attractive because it makes the math more tractable in two ways. First, if the likelihood factors into multiplicative terms, then the effect of the logarithm is to separate the factors into additive terms that can be maximized independently – effectively decoupling the classification problem. Second, likelihoods tend to take exponential forms, such as Gaussian and Binomial distributions, and the logarithm operation conveniently converts exponents into multiplicative factors. The caveat is that $p(D|h)$ must be everywhere nonzero, which we can ensure in practice by substituting the smallest representable positive number for zero.

4.1.2 The EM Algorithm

Consider the problem of determining the probability densities that generated a certain data set. Given the general form of the densities, their governing parameters can be estimated using ML to maximize the likelihood of the data. Suppose, however, that some of the data is missing, hidden, or represented by latent random variables. Since we cannot compute the likelihood of unseen data, we instead compute its expected value, and *maximize* this *expectation*. Therefore, the name of this general approach is expectation-maximization (EM).

Following the notation of the original EM paper, [Dempster77], let the current set of parameters be denoted by ϕ, and a revised set that we are seeking to compute be denoted by ϕ'. Suppose that we have observable data y and latent data x that is not observed directly, but only indirectly through y. We would like to choose the parameters ϕ' that maximize $\log p(x,y \mid \phi')$, but we do not know $p(x,y \mid \phi')$ because x is unobserved. Consider what we do know, which is the marginal probability of the visible data y. The marginal density is found by integrating the joint density over all possible values of x:

$$\log p(y \mid \phi') = \sum_x \log p(x, y \mid \phi') \tag{4.5}$$

Given that x is a random variable, we can average the log likelihood $log\ p(x,y \mid \phi')$ over all possible values of x, weighting each according to its probability. This is accomplished by inserting a term for the probability of x into the summation in equation 4.5. (We express this probability as $p(x \mid y,\phi)$ instead of $p(x)$ to denote its conditional dependence.)

$$\langle \log p(x,y \mid \phi') \rangle = \sum_x p(x \mid y,\phi) \log p(x,y \mid \phi') \tag{4.6}$$

Observe that equation 4.6 represents the expected value of the log likelihood. The expectation is performed over the probability of the hidden variables, and another notation is:

$$\langle \log p(x,y \mid \phi') \rangle = \underset{p(x \mid y,\phi)}{E} \log p(x,y \mid \phi') \tag{4.7}$$

We repeat equation 4.7 once again just to use the notation of [Dempster77]. The authors label the expectation with the term $Q(\phi' \mid \phi)$ to denote that we are searching for a revised hypothesis ϕ' given the current hypothesis ϕ.

$$Q(\phi' \mid \phi) = E[\log p(x,y \mid \phi') \mid y,\phi] \tag{4.8}$$

We can then choose a new ϕ to maximize this expectation:

$$\phi' \leftarrow \arg\max_{\phi'} Q(\phi' \mid \phi) \tag{4.9}$$

Thus, the parameters ϕ' are set to the values that would make the complete data most likely. However, observe the circularity of the computation, and therefore the need for iteration. The probability of the hidden variables $p(x \mid y,\phi)$ is calculated using the observed data y and the *belief* that the current parameter hypothesis ϕ is correct. But equation 4.9 then updates ϕ, which alters that belief. Once ϕ has been improved, we can re-compute the expectation to better "fill-in" the hidden x. That, in turn, will allow us to recalculate a better ϕ. Iteration can continue between the following 2 steps until a local maximum has been reached.

E-Step:

Compute the **Expectation**, $Q(\phi' \mid \phi)$, using the current ϕ and visible data y.

M-Step:

Perform the **Maximization** to replace ϕ by the ϕ' that maximizes $Q(\phi'|\phi)$.

While this completes the general description of the EM algorithm, we would like to make some comments regarding its use in practice. First, while the theoretical goal of the E-Step is to compute the full expectation, in efficient implementations, we need only compute the probabilities of the hidden variables $p(x \mid y,\phi')$ for use by the M-Step. Second, to allow the M-Step the freedom to contain computationally simpler steps (with the penalty of slower convergence), it may compute a better ϕ, but not necessarily the one that maximizes Q. Such an approach is referred to as Generalized EM (GEM). A corresponding idea to partial maximization in the M-Step is to partially perform the E-Step as proven in [Neal98].

4.1.3 EM Segmentation

The EM algorithm was first applied to medical imaging to achieve image reconstruction. SPECT images can be computed by finding the most probable image that is consistent with the observed projection data [Lange84]. Later, [Wells96b] applied EM to medical image segmentation to simultaneously classify MR images while correcting for the magnetic field inhomogeneities. In this domain, the *visible variables* are the image intensities, the *hidden variables* are the tissue classifications, and the *parameters* govern the bias field that models the inhomogeneities. If the bias field were known, then the tissue classes could be estimated directly from the intensity-corrected image. On the other hand, if the tissue classifications were known, then the bias field could be estimated from the difference between the observed and predicted intensities. Therefore, the EM algorithm iterates as follows:

E-Step:

Compute the expected values of the tissue classifications assuming that the current estimate of the bias field is correct.

M-Step:

Calculate the bias field assuming that the tissue classifications are correct.

Although the calculation of the bias field is dependent on tissue classifications, we include it in layer 1 of CDN. The reason is that the bias field is computed to correct voxel intensities rather than add to the understanding of their meaning. If the bias field could be corrected for as a preprocessing step, then classification could proceed normally through the CDN.

[Wells96b] derived the EM segmentation algorithm from the standpoint of a MAP estimator of the bias field. In the appendix of this thesis, we present a slightly different derivation by deriving EM segmentation directly from [Dempter77]'s definition of EM based on ML estimation. Additionally, our derivation uses our imaging model from Chapter 2 to explain the validity of the various assumptions.

4.2 Robust Bias Estimation

4.2.1 Bias Correction

The bias field is most pronounced when surface coils are used, but brain scanning is typically performed with a birdcage head coil [Dongfeng91] that results in minimal bias effects. In this section, we will exaggerate the bias field on synthetic data to clearly illustrate how the algorithm negotiates these signal inhomogeneities. The bias field was simulated with linear and sinusoidal patterns as depicted in Figure 2.6. Figure 4.1 demonstrates that the impact of the bias field is that tissue classes cease to be linearly separable. Note that although the bias field is only marginally apparent in the original scan, it greatly corrupts the correctness of the segmentation.

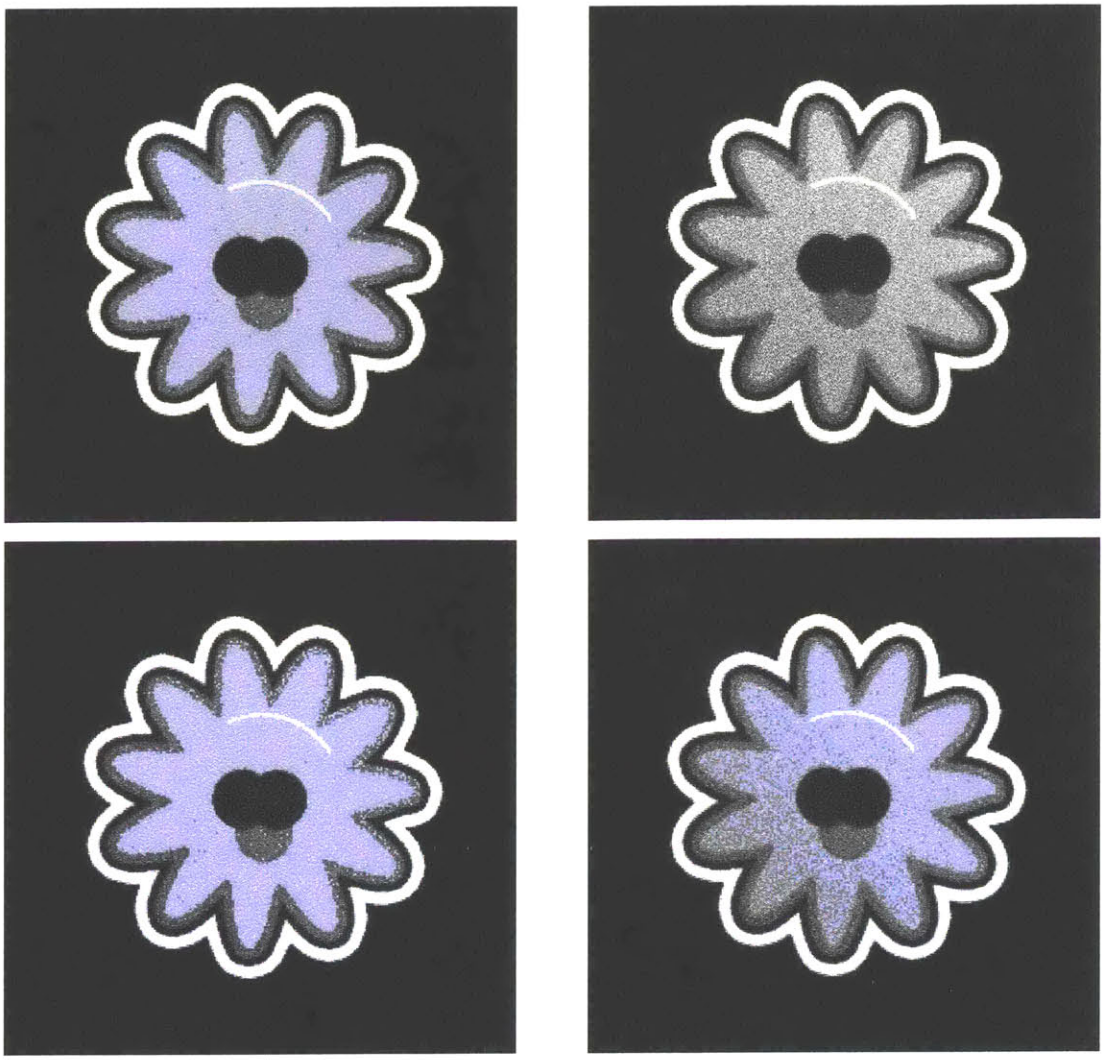

Figure 4.1. Effects of the Bias Field. We experimented with applying a threshold to segment white matter as blue. While this worked well with the original image (top left), the two images on the bottom show that attempts to threshold the biased image (top right) result in either misclassifying upper-right gray matter or lower left white matter.

Figure 4.2 demonstrates the impact of the bias field on the segmentation by showing the intermediate segmented results after several different iterations of the EM algorithm. For variety, the sinusoidally varying bias field was applied to these images.

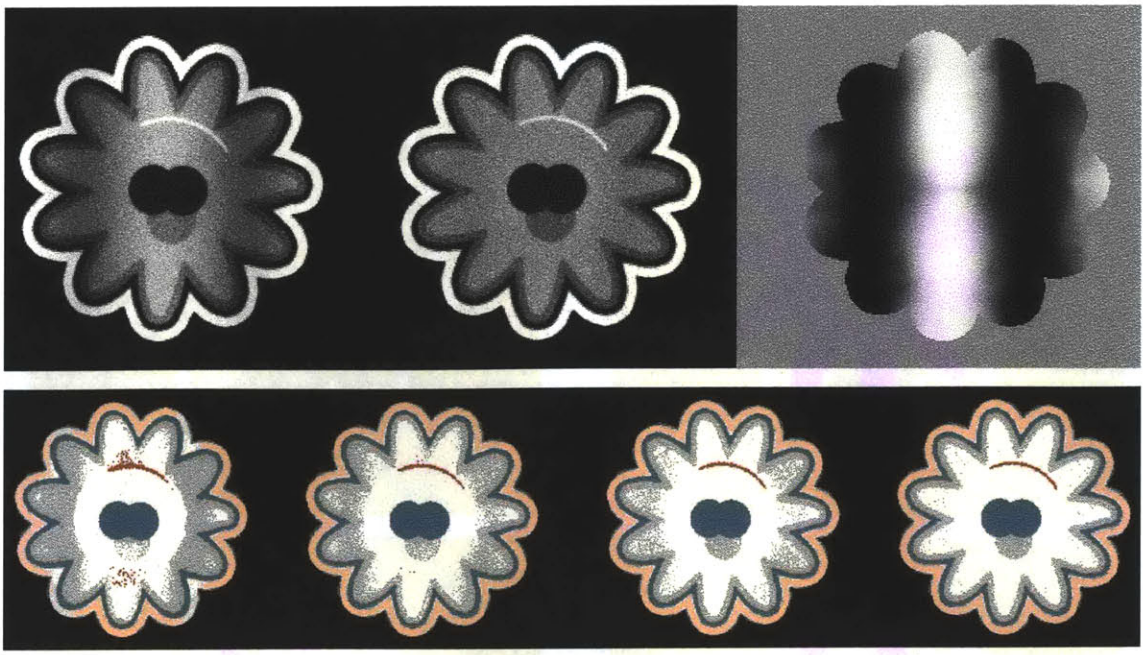

Figure 4.2. EM Bias Correction. (Top) From left to right are the corrupted original, the correcting image after performing EM segmentation, and the recovered bias field. (Bottom) From left to right are intermediate segmentation results after 1, 2, 3, and 10 iterations of the EM algorithm. Observe how the full extent of the white matter (white) is correctly discovered after 10 iterations.

4.2.2 Bias Correction Influenced by Pathology

Recall the six voxel intensity modifiers identified in Chapter 2 that cause the segmentation problem to be ill-posed. If we run the EM segmentation algorithm on a scan that contains pathology, then EM will attempt to remove the pathology by adjusting the bias field. To combat this, we weight each voxel's contribution to the bias field estimation according to its measurement of abnormality. Thus, voxels with high typicality contribute strongly to the estimation, while voxels that are almost certainly tumor are ignored. The degree of weighting can be set with a single parameter, or the parameter's value can change according to a schedule through the course of iterations. Figure 4.3 shows that such a weighting is not perfect, but it is an improvement toward resolving the ambiguity between bias and pathology. For the most part, bias is a very smooth, very slowly varying phenomenon, while tumors are not.

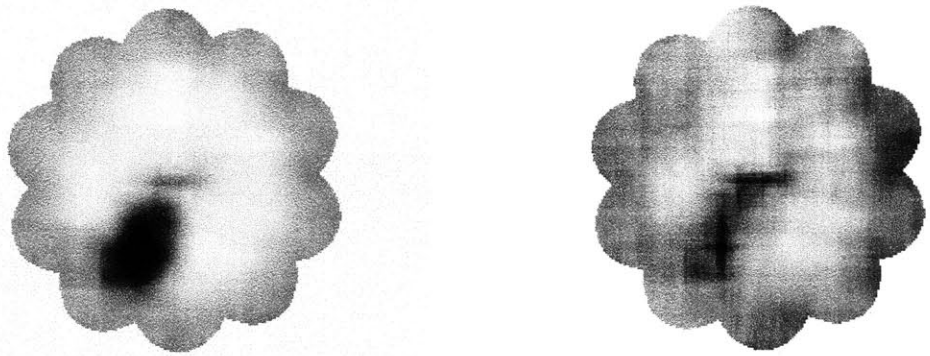

Figure 4.3. Weighting the Bias Computation with the Probability of Pathology. The image on the left is the result of running 10 EM iterations on a synthetic scan with a tumor but no bias field. The algorithm mistakes the tumor for a supposed bias. The image on the right is the bias computed after 10 iterations when the bias computation is weighted by the probability of pathology.

4.3 Spatially Varying Priors

Since Bayesian classification includes a term for *a priori* knowledge of tissue class likelihood, we desire meaningful values for this term. A *stationary prior* is *a priori* knowledge that is not spatially varying. Table 4.1 lists the prior computed from 300 synthetic brains by counting the number of voxels belonging to each tissue class. The final tallies were normalized to sum to 1 in order to express probabilities.

Table 4.1. Stationary Priors computed from 300 synthetic scans.

Tissue Class	Stationary Probability
Scalp	0.178
White matter	0.442
Gray matter	0.196
CSF	0.179
Vessel	0.005

Instead of keeping total counts, a localization model that is commonly known as a *spatially varying prior* can be computed by keeping separate counts for each voxel. The prior shown in Figure 4.4 was computed from the label maps of 300 synthetic brains by

counting the number of occurrences of each tissue class at each location, and then normalizing the result to form probabilities. That is, at each voxel location, the contributions from the 6 images of Figure 4.4 sum to 1.

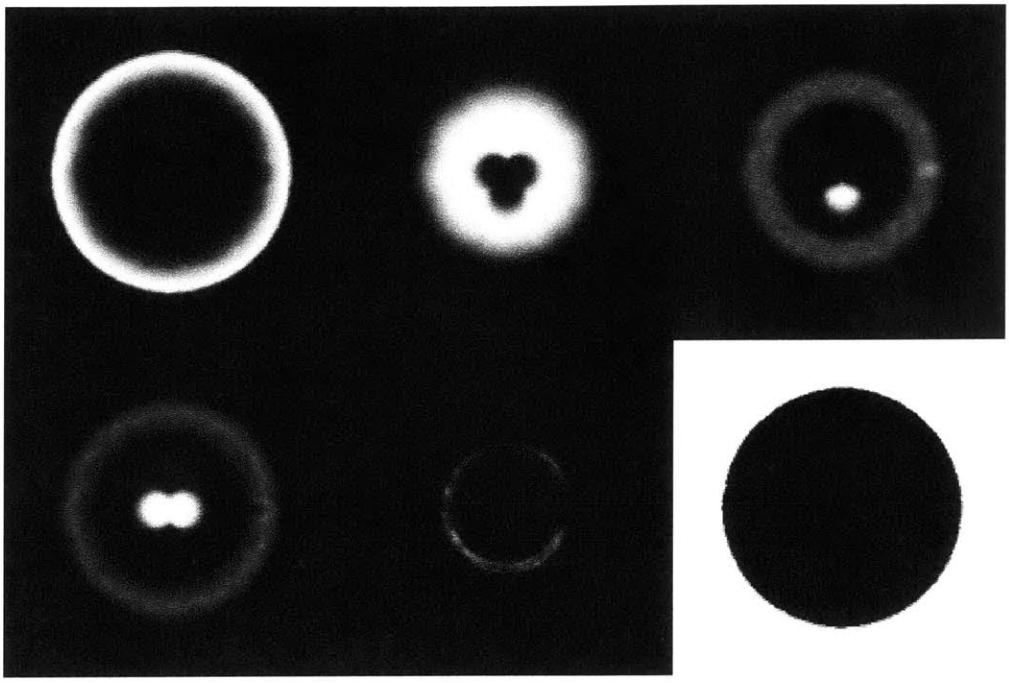

Figure 4.4. Spatially Varying Prior. Each image represents the probability of occurance of a certain tissue class at each voxel location. From top left to bottom right, are scalp, white matter, gray matter, CSF, vessels, and background.

Figure 4.5 illustrates the impact of applying the atlas depicted in Figure 4.4. Observe how EM classification errors decrease from left to right (top to bottom in Table 4.2). Apparent errors include mistaking scalp for vessel, white matter for gray matter, and tumor for CSF.

Table 4.2. Impact of Priors computed from results of Figure 4.6.

Prior	# Misclassified Voxels
None	2150
Stationary	2070
Spatially Varying	1529

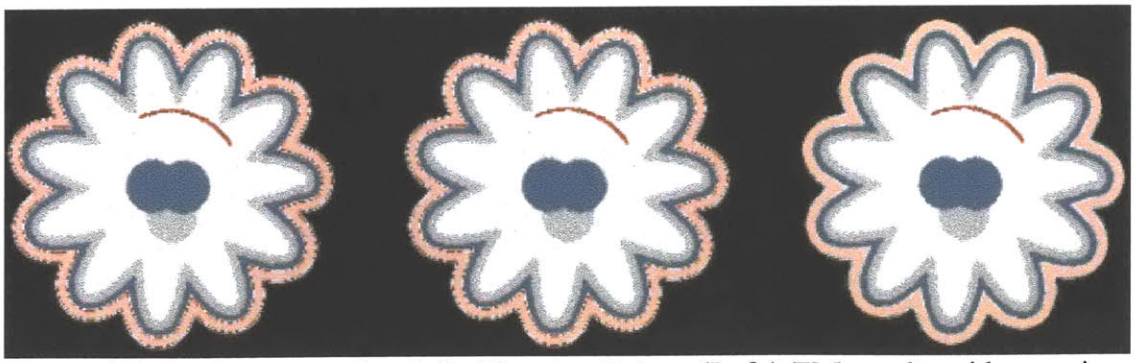

Figure 4.5. Stationary vs Spatially Varying Priors (Left:) EM results with no priors. (Center:) Results with stationary priors computed from a training set. (Right) Results with the spatially varying priors computed from the same training set of 300 images.

The literature argues that the incorporation of spatially varying priors adds context into the classification process. However, we include spatially varying priors here in CDN layer #1 because their usage is not dependent on an initial classification. It may add context, but it does not add *predicated context*. For example, a statistical atlas is aligned using a registration step that is based on the gray-level images, before any classification is performed. As another example, in [Kapur99]'s approach of using distance to major landmarks, the distances are computed before any initial classification is performed. This is possible because the major landmarks, such as skin and ventricles, may be easily segmented as a preprocessing step before the brain tissues of interest are brought into consideration.

There are at least four different approaches for replacing the stationary tissue class prior with a spatially varying prior in the calculation of the posterior probabilities:

1.) [Kapur99] localized anatomical structures relative to landmark anatomical structures within the same patient. Specifically, she used a joint probability distribution based on distances from ventricles and skin. By avoiding registration with an atlas, the method has the advantage of avoiding the dependence on the quality of the registration and on the similarity between the patient and the atlas – a notion that can change considerably in the presence of large pathology.

2.) A rigidly registered digital atlas has the advantage of adding a richer understanding of context to the computation then can be achieved using relative position to the patient's

own landmark structures. Such an atlas can be constructed from a highly detailed segmentation of a single scan [Kikinis96], or an average of a very large collection of scans [Evans93, SPM], as depicted in Figure 4.6.

3.) [Fischl02] used a hybrid approach to overcome the deficiencies of the first two methods. The atlas was constructed from only a few scans to avoid the blurring of fine structures in an averaged atlas, yet was not as susceptible to specific irregularities present in a single scan.

4.) [Pohl02, Rexilius01, Warfield01, Warfield98b] use a non-rigidly registered digital atlas because of the need for local agreement between the atlas and patient. This is an attempt to overcome the shortcomings of the first two methods. In this application however, non-rigid registration would incorrectly attempt to morph the healthy anatomy of the atlas to conform to the unusual morphology of the patient's pathology. Instead, we desire the atlas to provide the spatially dependent tissue probabilities as if the patient were healthy. One possible method would modify the non-rigid registration algorithm [Thirion98] to reduce the degree of warping in the presence of pathology. The rough location of the tumor can be identified very quickly through a first-pass of the segmentation algorithm using rigid registration. This information would then bias the non-rigid registration routine.

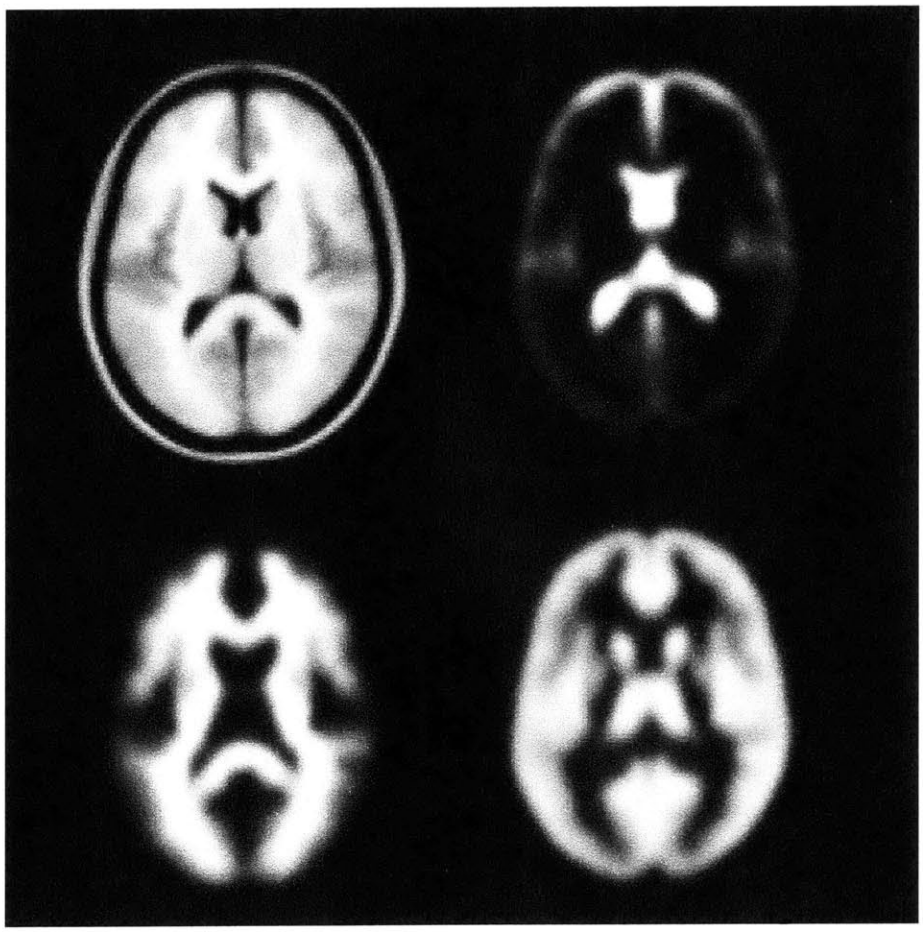

Figure 4.6. Spatially varying priors computed by averaging 305 scans. Clockwise from the top left: average scan, probability of CSF, probability of GM, probability of WM.

4.4 A Computational Paradigm for every CDN Layer

In this section, we examine the use of spatially varying priors for incorporating context, and we demonstrate why they are insufficient in meeting our requirements. To propose a solution using CDN, we generalize the concept of these priors into probabilistic mappings between image space and model space. We then base the processing for each layer of CDN on these mappings.

4.4.1 Mean Samples vs. Typical Samples

Consider spatially varying priors as a *localization* model. We can construct an *intensity* model in the exact same manner: for every voxel location, compute the mean intensity

across an ensemble of images. Figure 4.7 depicts such a model computed from the training set of 300 synthetic brain images. In addition to computing the mean, we also computed the variance at every voxel to form a voxel-wise Gaussian model. Could this intensity model be used to answer our two guiding questions? That is, does the mean image define "normal", and does the variance image enable measuring abnormality?

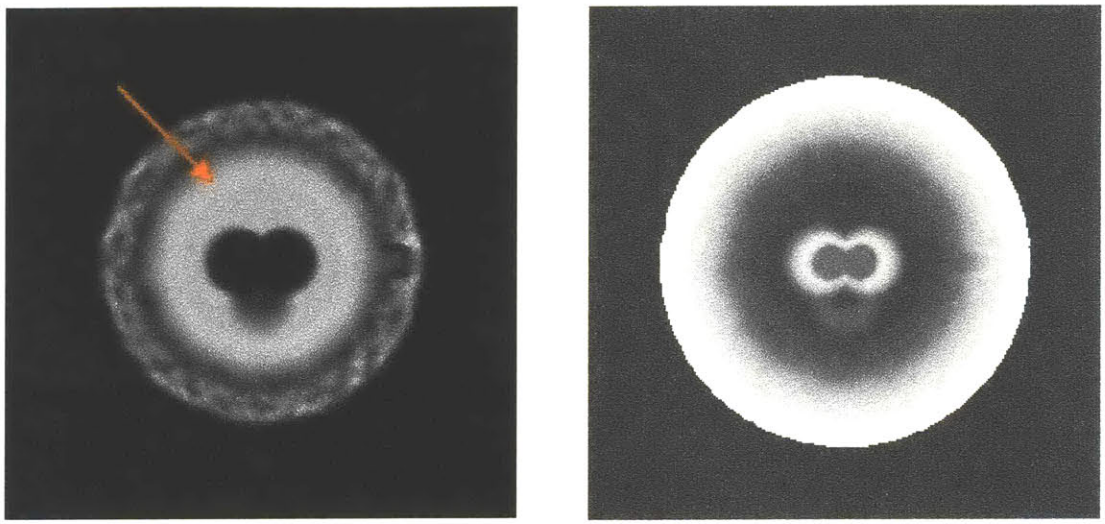

Figure 4.7. Voxel-wise Gaussian Model. The mean (left) and relative deviation (right) of 300 synthetic images. Relative deviation is the variance normalized by the mean for better display. The red arrow points out the faint glow of vessels that covers the outer half of the white matter anulus.

The peculiarity of this model becomes immediately apparent upon inspection of Figure 4.8. To state it bluntly, no synthetic brain looks like that. More formally, by defining *normal* to be the statistical average of all brains, *normal* is an impossible achievement. For example, whereas vessels are thin tubular structures of very bright intensity, they appear as a thick, very faint, ghostly glow in Figure 4.7. Consequently, using the above model causes any normal vessels to be identified as abnormal. Figure 4.8 demonstrates the results of using this model to measure abnormality on healthy and diseased synthetic brains. In addition to the vessel anomaly, observe that the outermost fringes of the healthy brain were measured to be nearly as abnormal as the tumor of the diseased brain. Thus, the model is too insensitive to the rare extent of the gyral protrusions of the healthy brain.

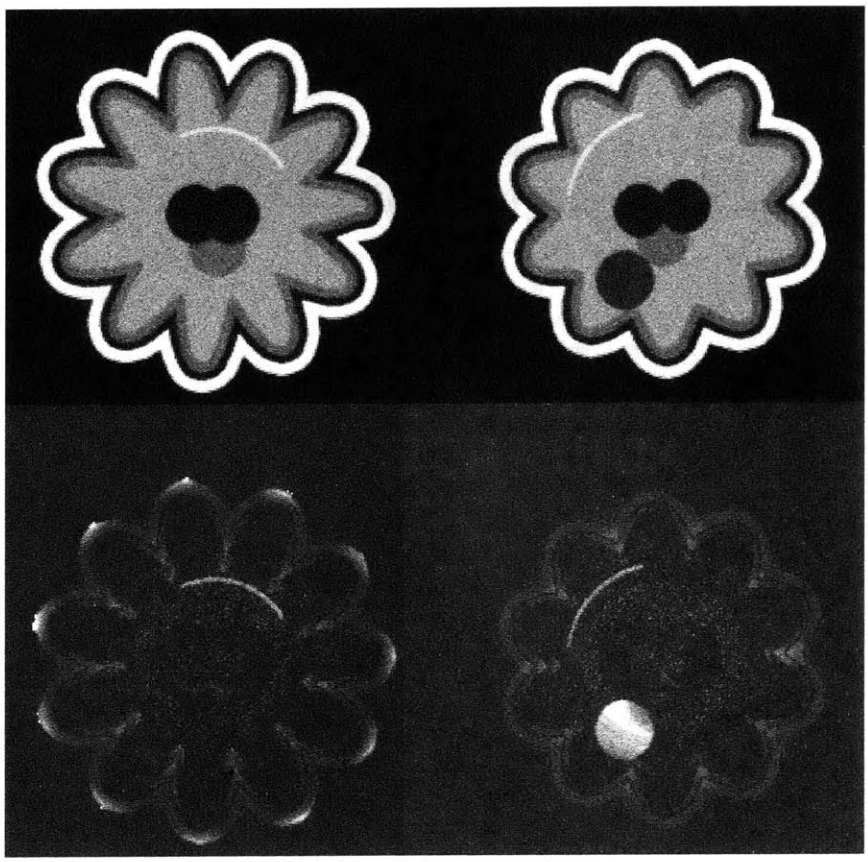

Figure 4.8. Measuring Abnormality with a Voxel-wise Model. (Top:) the input synthetic scans, one healthy and one diseased with a dark gray, circular tumor. (Bottom:) Mahalanobis distance measured using the model of Figure 4.7. The central issue that we wished to expose is that the vessels (thin bright arc subtending 40 degrees), regardless how healthy, are always recognized as abnormal by this model.

Figure 4.9. Average is Not Normal. Martha's Vineyard has a spinning light with one side red and one side white. While the average color is pink alright, that's the flower -- never the light.

Figure 4.9 suggests the same simple solution as that provided by spatially varying priors. An observer needs to map his/her observation to a model space consisting of two distinct models: red and white. The mapping exists to separate *time and color*. Instead of expecting pink light at all times, red light is expected to be emanating from the lighthouse during the first half of its rotary cycle, and white light from the second. Similarly, spatially varying priors separate *intensity and location*. Vessels would not be mistakenly labeled abnormal in Figure 4.8 if there were one model for their bright intensity, and another model for their expected whereabouts.

However, we argue that spatially varying priors suffer from the same problem that they purport to solve. Once intensity and location have been separated, there remains a need to also separate out size and/or shape. For example, the spatially varying prior of Figure 4.4 may solve the vessel anomaly of Figure 4.8, but it would also treat vessels several times too thick as completely normal, as illustrated in Figure 4.10. The reason is that the lighthouse analogy applies again, only at another level. Our voxel-wise Gaussian intensity model of Figure 4.7 failed because it relied on average intensity, and average is not normal. Similarly, a spatially varying prior relies on average position, and average is not normal.

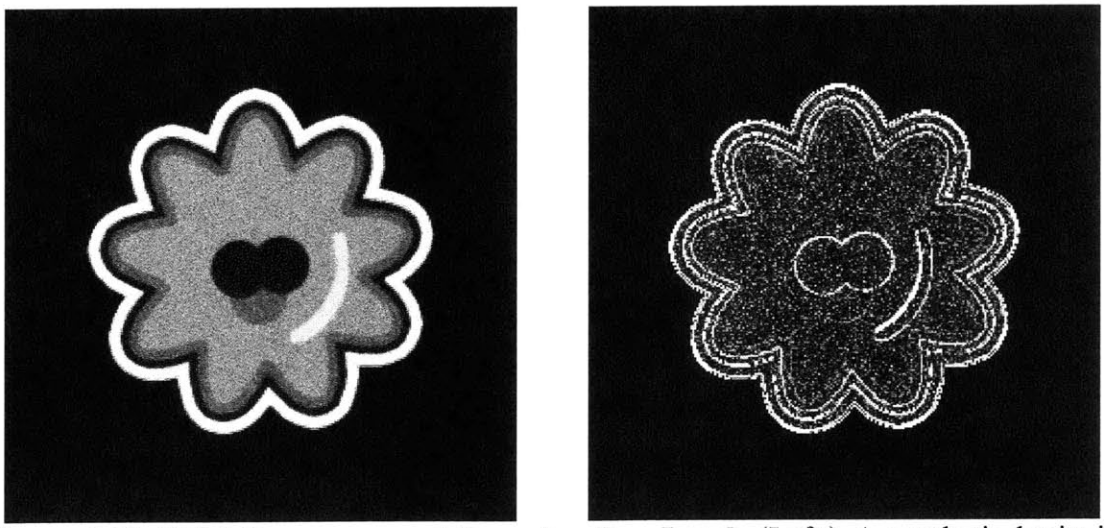

Figure 4.10. "Lighthouse anamoly" on Another Level. (Left:) A synthetic brain is generated with an unusually sized vessel (bright arc in lower right quadrant). (Right:) Similar to why the vessels, regardless how healthy, were always recognized as abnormal in Figure 4.8, wide vessels, regardless how abnormal, are always recognized as healthy here. (The vessel boundary is labeled abnormal due to partial volume artifacts, but not the vessel interior.)

In summary, we have defined spatially varying priors as a solution to the "lighthouse anomaly", but we have simultaneously criticized them for suffering from the same problem, but on another level. The solution that we propose in this thesis is to generalize the concept of a spatially varying prior, and apply it at more levels. These "levels" are naturally related to the CDN layers. Following the intuitive arguments of this section, the next section formally defines this generalization.

4.4.2 Probabilistic Mapping from Image Space to Model Space

Consider a general processing paradigm that consists of the following three elements:

1. Define *image space* to be the input data.
2. Define *model space* to be a set of distinct models such that the dimensionality of model space is less than or equal to that of image space.
3. Situate a *probabilistic mapping* between image space and model space.

Consider casting CDN Layer #1 into this paradigm. Let image space be the set of input voxel intensities, and model space be the set of Gaussian intensity distributions. To establish the dimensionality of model space, recall from our imaging model of Chapter 2

that the MR scan parameters are set so as to resolve all structures of interest. We thus know that each anatomical structure occupies multiple voxels, alleviating the need to have a separate intensity model for each voxel. As a direct implication, model space should have significantly lower dimensionality than image space. Specifically, if we assign one dimension of model space to correspond with each interesting tissue class, then model space becomes the set of parameters for a few Gaussian distributions (eg. $\{\Phi_{WM}, \Phi_{GM}, \Phi_{CSF}, \Phi_{Vessel}, \Phi_{Scalp}\}$). Then, the probabilistic mapping becomes the spatially varying prior of Figure 4.4. Interestingly, our experiment in Figure 4.7 with a Gaussian intensity model for each voxel location also represents an instantiation of this paradigm. Observe the relation between these two examples in Figure 4.11.

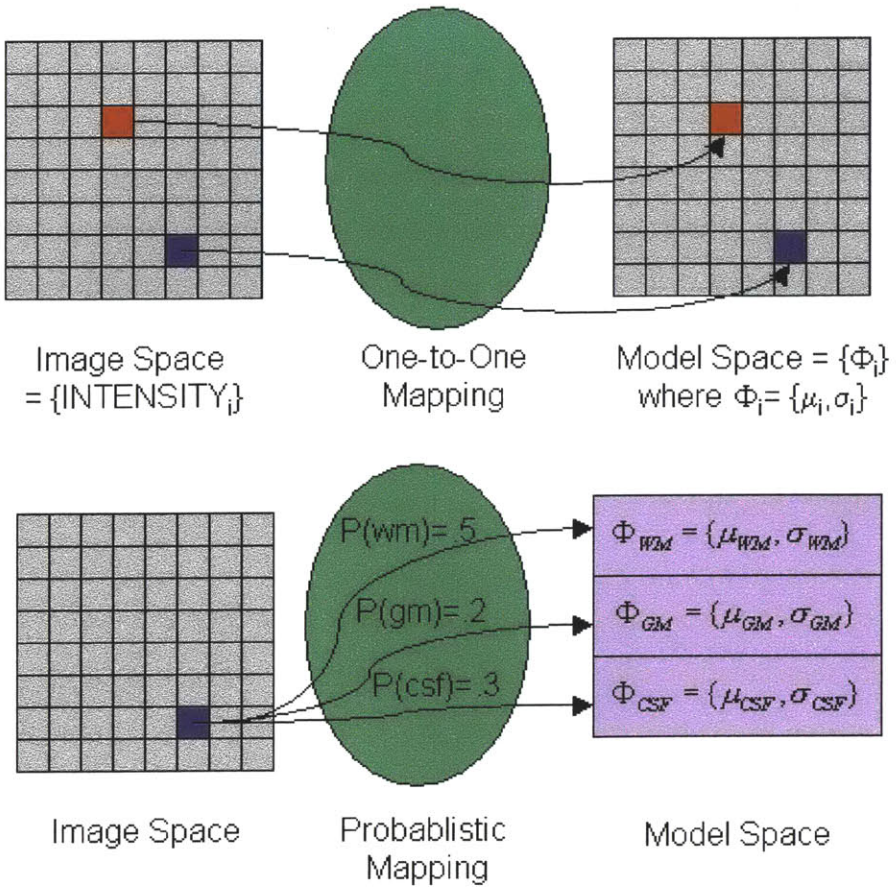

Figure 4.11. Mapping from Image Space to Model Space. (Top:) One-to-one mapping such as in Figure 4.7. (Bottom:) Probablistic mapping such as in Figure 4.4, which is a generalization of one-to-one mapping. Two abstractions were made in transitioning from the top paradigm to the bottom one. First, the dimensionality of model space was reduced to less than that of image space. Second, the one-to-one mapping was relaxed to be a probabilistic mapping between a *single* element of image space and *every* element of model space.

Next, let us examine the space complexity of the two example paradigms of Figure 4.11. Given an image dimensionality N, and model dimensionality M, the top paradigm requires O(N) space since M=N. The bottom paradigm, on the other hand, requires significantly more space: O(MN). The bottom's model space is smaller than the top's, but the top's mapping is trivial while the bottom's mapping acts as a seat of knowledge. Inference becomes possible based on this intelligent mapping. For example, the bottom paradigm can answer questions such as "Where can one expect to find white

matter?" or "Do vessels ever exist in the scalp?" but the top paradigm is incapable of performing such reasoning.

This general paradigm of situating a probabilistic mapping between an image space and a model space will frame the computation at every layer of CDN. For example, in CDN layer #1, the probabilistic mapping will be implemented as the spatially varying prior of Figure 4.4. Other layers in our implementation will be based on abstractions somewhere between the top and bottom paradigms of Figure 4.11. Model space will have different dimensionality and content for each layer. While model space contains models for intensity in layer #1, it will contain models for neighborhood interactions in layer #2, models for shape descriptors in layer #3, and models for inter-structure relationships in layer #4.

We conclude this section with a look at how the existing works in the field of normal brain segmentation can be described using our abstract paradigm. Table 4.3 categorizes several works referenced thus far in this thesis, and a brief discussion follows.

Table 4.3. Using the Paradigm of Probablistic Mapping to situate various works in the field.

Type of Mapping (image-to-model)	Common Name	Example
One-to-One		Figure 4.7
All-to-One	Homogenous	MRF, Chapter 5
Many-to-One	Heterogeneous	MRF, Chapter 5
Many-to-Many	Grid	[Fischl02]
Many-to-All	Spatially varying prior	[Leemput99b]
All-to-All, equally probable	No prior	[Cline90]
All-to-All, unequally probable	Stationary prior	[Wells96b]

Distinct Paradigm Instantiations:

- **One-to-One:** There is a unique model for each voxel of image space such that the mapping is one-to-one and onto. We proposed this in Figure 4.7 for illustrative purposes.
- **All-to-One:** One model applies to all elements of image space. Chapter 5 will discuss the example of a homogenous Markov random field where the model is a matrix of probabilities of tissue class interactions.

- **Many-to-One:** There are multiple models, and each one applies to a distinct subset of image space. Chapter 5 describes the example of a heterogenous MRF.
- **Many-to-Many:** Similar to Many-to-One except that the mapping is probabilistic. Multiple elements of image space map with some probability to each of multiple, but not all, elements of model space. For example, [Fischl02] is a heterogenous MRF with a model space of such large dimensionality that allowing each image voxel to map to all elements of model space would be too computationally intensive, especially with regard to space.
- **Many-to-All:** Similar to Many-to-Many except that multiple elements of image space map probabilistically to *all* elements of model space. This is feasible for model spaces of very small dimensionality, such as a typical spatially varying prior that characterizes only a handful of tissue classes.
- **All-to-All:** All image voxels have a non-zero probability of mapping to each element of model space. Examples include stationary priors in MAP classification and the lack of a prior in ML prior. The lack of a prior is the degenerate case of equal probabilities.

4.5 Computing a Probability of Pathology

The previous sections have explored how to assign probabilities of tissue class membership to each voxel. The probabilities are derived from statistical models of healthy tissues. However, we need a method for assigning a probability of pathology, for which no model is available. In this section, we define such a method.

4.5.1 Computing Abnormality

A probability of pathology can be computed based on the inadequacy of model space to explain the appearance of the voxel. In our case, the model space for voxel intensity is a multi-modal Gaussian distribution, with the typical profile from training data graphed in Figure 4.13. (Although partial volume artifacts (PVA) have the effect of "filling" in the valleys between the pure Gaussian distributions, we ignore this fact in CDN Layer #1 because CDN Layer #4 will address PVA.) We can use such a distribution to compute the probability that a given sample was not generated from the distribution.

Since the figure illustrates that there is negligible overlap of the tails of the distributions, we can simply the computation by calculating the probability that a given sample was not generated by the closest univariate Gaussian model.

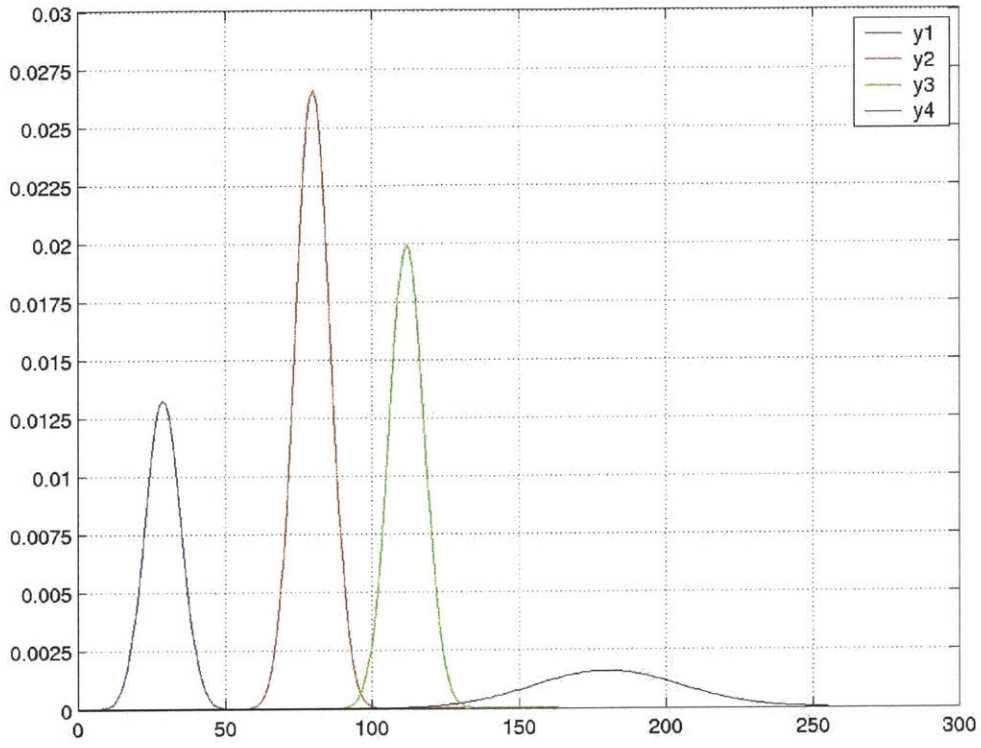

Figure 4.12. Typical Intensity Distribution for Post-Contrast Brain MRI. From left to right, are the Gaussian intensity distributions for CSF, gray matter, white matter, and vessels.

Let M denote Mahalonobis distance as defined in equation 3.1, then the probability of abnormality is defined as integrating the area under the Gaussian curve between +/-M of the mean, divided by the total area under the curve [Rice95]. This nonlinear function has the advantage of asymptotic growth, graphed in Figure 4.13, and it's desirable for expressing the abnormality measurement as a probability. Linear functions fail to express the fact that there is a point at which "very abnormal" becomes no different from "very, very abnormal".

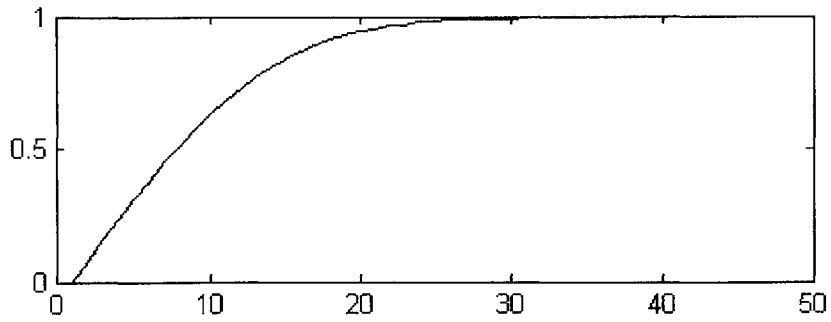

Figure 4.13. Asymptotic Growth of the probability of tenths of Mahalanobis distances.

The equations below summarize this calculation, where $p_i(T)$ represents the probability of tumor at voxel i, and $P_<(M)$ represents the probability of occurrence of a Mahalanobis distance of M or less. Let L denote the set of all possible labels, x the bias-corrected intensity, and λ the shift from the origin in terms of standard deviations.

$$p_i(T) = \min_{l \in L} P_< \left(\sqrt{\frac{(x_i - \mu_l)^2}{\sigma_l^2}} - \lambda \sigma_l \right) \tag{4.10}$$

$$\forall i: \; tumor_i \leftarrow p_i(T) > \max_{l \in L} p(l \mid x_i) \tag{4.11}$$

The logic underlying the inclusion of the offset λ is that there exists some small distance from the mean within which we are comfortable considering the sample to be entirely normal. This is justified based on the asymptotic equipartition property [Cover91] that enables us to divide a sequence of samples into two sets: the typical set and non-typical set.

We are now prepared to perform segmentation that incorporates measurement of abnormality. Figure 4.14 displays results using the spatially varying priors of Figure 4.4 and the intensity models of Chapter 2. Classification of healthy tissues is performed by selecting the maximum likely tissue class at each voxel. Pathology is included by labeling a voxel as tumor (rendering it green) whenever the probability of pathology exceeds that of all healthy tissue classes.

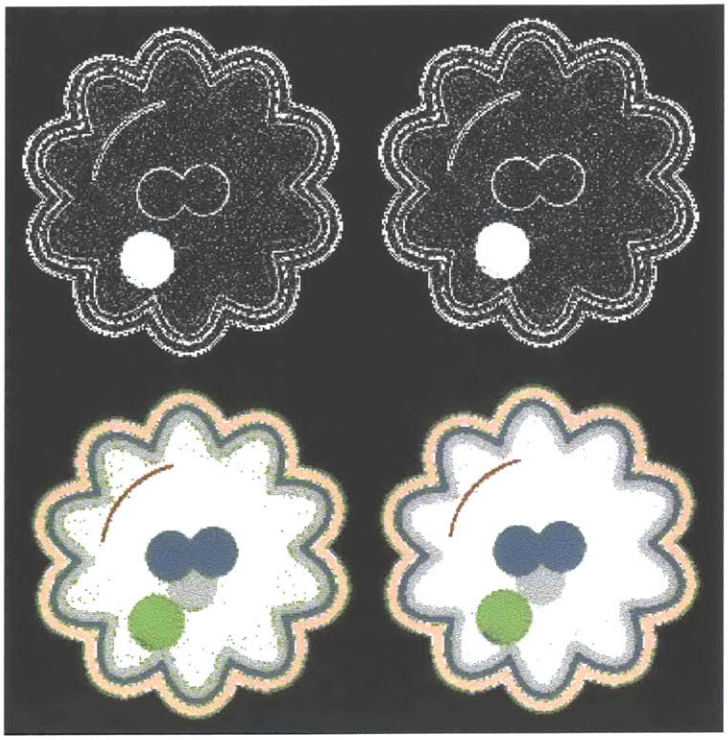

Figure 4.14. Asymptotic Abnormality. The top images are the probability of Mahalanobis distance, and the bottom images are the corresponding segmented label maps. Tissue classes include pathology (green), white matter (white), gray matter (gray), CSF (blue), vessels (red), and scalp (tan). The left side was generated using a baseline of 0 standard deviations, and the right side was generated using a baseline of 2 standard deviations. The apparent abnormality produced by the presence of partial volume artifacts will be handled in Chapter 6. For all subsequent experiments in this thesis, we used a baseline of 2 standard deviations for measuring intensity abnormality, and 0 for shape abnormality (Chapter 6). These parameter values were determined from running the algorithm on a healthy scan.

4.5.2 Comparing NNPM with Probabilistic Models

As an aside, we run an experiment in this section to compare computing abnormality based on NNPM vs. Gaussian models. Recall that NNPM defines normal as a set of example templates, and the abnormality of a sample is measured by the distance between the sample and the nearest template. Since the performance of the algorithm is critically dependent on the selection of the members of the training set, we ran an experiment to measure its sensitivity to training set size for images of size 1 voxel. Table 4.4 lists the results in a manner conducive to making comparisons with probabilistic models.

The experiment involved analyzing the set of samples in the first column using NNPM. The experiment was repeated with 4 different training set sizes: 1, 10, 100, and 1000. The templates that populated each training set were drawn from a random sampling of a white matter distribution with mean 120 and variance 33.

Upon inspection of the table, NNPM's shortcoming of treating all templates as equally normal is clearly evident. With only one template, the measured abnormality merely grows linearly with distance. As the number of templates increases, the measurement's linear march begins from an increasingly larger initial distance from the mean value of 120 (10 for 10 templates, 15 for 100 templates, and 20 for 1000 templates).

Table 4.4. Sensitivity of NNPM to Training Set Size. In the righthand columns, M stands for Mahalanobis distance, and $P_<(M)$ represents the probability of occurrence of a Mahalanobis distance of M or less.

Sample Values	**Measured Abnormality with NNPM (RMS error)**				**Measured Abnormality with Gaussian Model**		
	1 template	**10 templates**	**100 templates**	**1000 templates**	**M**	**$P_<(M)$**	**$P_<(M-2)$**
120	0	0.9	0.0	0.0	0.0	0.00	0.00
115	5	0.0	0.3	0.0	0.9	0.63	0.00
110	10	0.8	0.3	0.0	1.7	0.91	0.00
105	15	5.8	0.9	0.3	2.6	0.99	0.45
100	20	10.8	5.9	1.6	3.5	1.00	0.84
95	25	15.8	10.9	6.6	4.4	1.00	0.98

From the results in the "1000 templates" column of Table 4.4, observe that NNPM treats "slightly abnormal" (sample value of 115 or 110) as effectively normal. This "shifting from the origin" is similar to our use of λ in equation 4.10. In contrast to NNPM, the right-hand side of Table 4.4 exhibits results of employing a Gaussian model. As with RMS error, the Mahalanobis distance also increases linearly with deviation from the mean. But because the distribution is known, these distances can be fit with a probability measurement. The probability is computed as the area under the Gaussian curve between +/-M of the mean, divided by the total area under the curve [Rice95].

From this discussion, we observe that subtracting some small number of standard deviations from the Mahalanobis distance before computing the probability is a way of combining the advantages of both NNPM and Gaussian model-based approaches by. As a result, the two gray columns in Table 4.4 appear more similar than the comparison between any other Gaussian model column and the gray NNPM column.

Figure 4.15. An Abnormality Function that is Shifted from the Origin, Exponentially Rising, and Asymptotic. Along the Maine coast, the photographer selected a rock on which to stand where splashes were an abnormal occurrence. Small sprays were perfectly acceptable, but larger spashes were greeted with rising intolerance. Once soaked by a wave, however, becoming any more wet was irrelevant.

4.6 Generative Models of Normal Anatomy

How can we assess how well our models encode descriptions of normal anatomy? One answer is to reverse the recognition process from being an analytical one to being a generative one. It is not clear that recognizing deviations from normalcy necessarily requires the computer to have some notion of a "generative model" of the brain. Regardless, it is instructive to consider the impacts that variations in the degree of context have in producing a generative model.

Toward this end, we experimented with inverting the analytical process of CDN layer #1 by using Monte Carlo simulations [Papoulis91]. In each experiment, the

generation proceeded at each voxel location by drawing a tissue class at random given the prior probabilities, and then generating an intensity using the Gaussian distribution for the selected class. Figure 4.17 displays the results, from which a number of observations can be made. Most apparent is the impact realized through the addition of context. Furthermore, it is evident that higher CDN layer #1 has no knowledge of the sizes and shapes of structures. Observe how the vessel is scattered into isolated points rather than a tube. As another comparison, Figure 4.16 includes the results of the generative model produced by the one-to-one mapping paradigm of Figure 4.7.

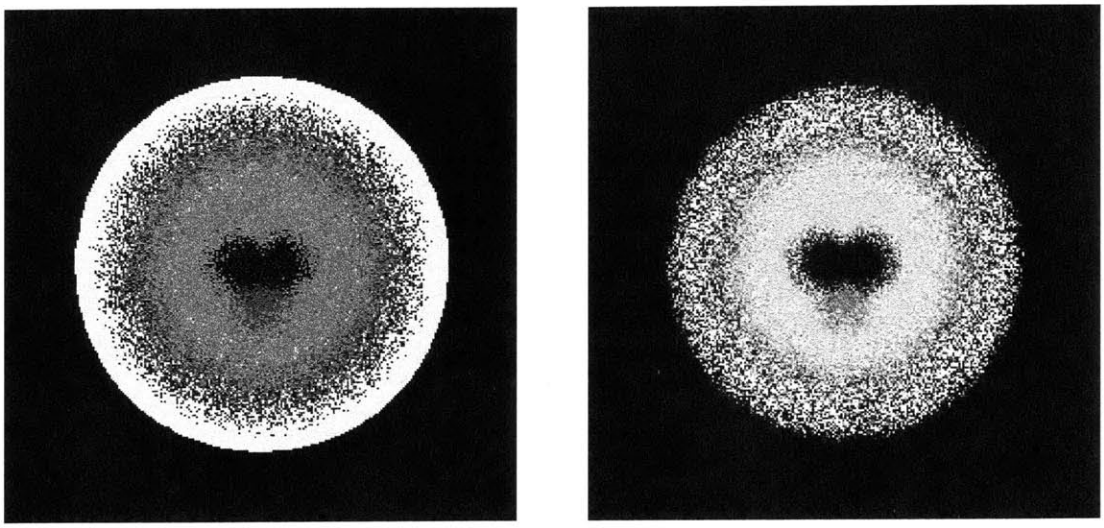

Figure 4.16. Generative Models: Paradigm Comparison. Generative models computed from the top paradigm of Figure 4.11 (right) and bottom paradigm (left). Note that the image on the left has no discernable vessel voxels, and there is no means of generating a companion label map as was done for Figure 4.15. This is because of the absence of the "seat of knowledge" represented by having a probablistic mapping.

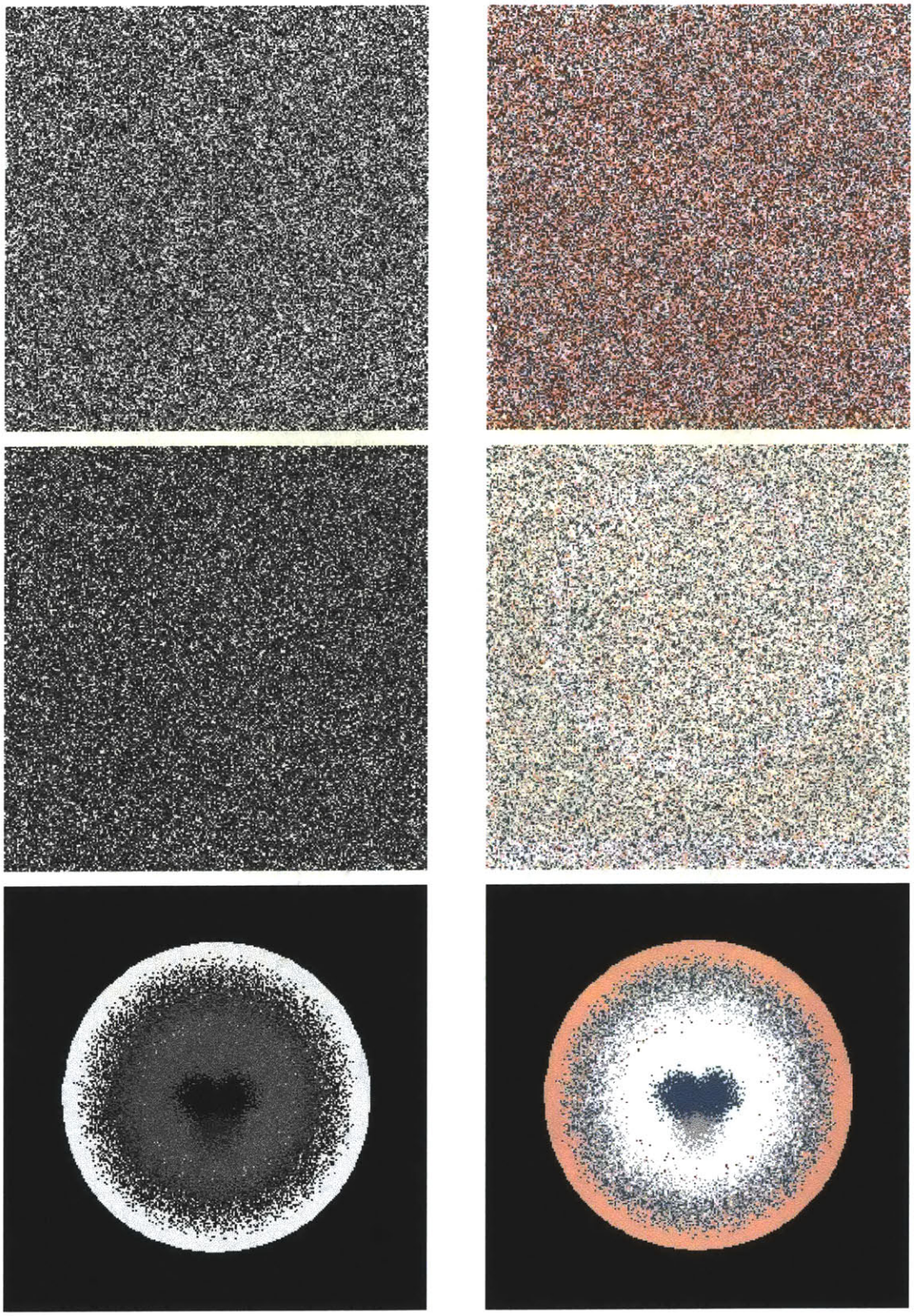

Figure 4.17. Generative Models of the Brain. The generative model produced the intensity images on the left, and segmentation on the right. The top row used no prior, the center row used a stationary prior, the bottom row used the spatially varying prior.

4.7 Chapter Summary

Within this chapter, we introduced the first layer of our framework for a contextual dependency network. The role of the first layer is to produce a preliminary classification of each voxel so that the next layer has a starting point from which to consider immediate context. After reviewing Bayesian classification and EM segmentation, we examined the "lighthouse anomaly" to explain a fundamental flaw in the field's trend of developing spatially varying priors. To propose a solution using CDN, we generalized the concept of these priors into probabilistic mappings between image space and model space. We then based the processing for each layer of CDN on the abstract concept of these mappings. Finally, we defined a method for computing a probability of pathology based on the inadequacy of model space to explain the appearance of the voxel.

To summarize the important principles asserted in this chapter:

4.1 Each voxel's contribution to the EM-based bias estimation is weighted by its typicality in order to produce an estimation that is robust to pathology.

4.2 Recognizing deviations from normalcy using statistical models requires separating intensity information from location.

4.3 Model-based segmentation methods can be described as some form of probabilistic mapping between image space and model space.

4.4 Our imaging model specifies that model space should have lower dimensionality than image space.

4.5 A function for computing a probability of pathology is based on integrating the area under the tails of Gaussian distributions, and is thus shifted from the origin, exponentially rising, and asymptotic.

Chapter 5

CDN Layer 2: Neighborhood Classification

In this chapter, we introduce the second layer of our framework for Contextual Dependency Networks. While the first layer classified voxels in isolation, the second through fourth layers will add the consideration of context – immediate and broad. Immediate context will be the subject of this chapter, as we will consider the classification of each voxel's neighbors. This approach resolves some of the residual ambiguity remaining after classifying voxels based strictly on the basis of visual, rather than spatial, information.

Consider a segmented image, or a collection of labeled voxels, to be a collection of random variables – one per voxel. Specifying how probabilities should be computed for events involving subsets of these random variables requires a probabilistic model. Layer #1 adopted a naively simple probabilistic model: statistical independence between subsets of size 1. In layer #2, we will be specifying contextual constraints using a probabilistic model referred to as a Markov Random Field (MRF). MRFs conveniently model the mutual influence between voxels systematically using rational principles rather than *ad hoc* heuristics.

This chapter is organized to introduce the foundations of MRFs, derive iterated condition modes, and the mean-field approximation, experiment to contrast these two techniques, and finally apply to EM segmentation of pathological brains.

5.1 Foundations of Markov and Gibbs Random Fields

5.1.1 Random Fields and the Labeling Problem

A random field F is a collection of random variables F_i defined on a discrete lattice S such that there is one random variable corresponding to each site in S. Given a set of labels L, the labeling problem involves finding a many-to-one mapping, $f : S \rightarrow L$, that assigns a label w_i to each random variable. Such a realization, w of F, is referred to as a configuration, and the set of all possible configurations is the configuration space W. Given M labels and a lattice of m nodes, the formal definitions are as follows:

$$\begin{aligned}
&\text{Lattice:} && S = \{i \mid i \in 1..m\} \\
&\text{Random Field:} && F = \{F_i \mid i \in S\} \\
&\text{Labels:} && L = \{w_i \mid w_i \in \{1..M\}\} \\
&\text{Configuration:} && w = \{w_i \mid i \in S, w_i \in L\} = \{w_1, w_2 \ldots w_m\} \\
&\text{Configuration Space:} && W = LxLxL...xL = L^m
\end{aligned} \tag{5.1}$$

Furthermore, let N be a *neighborhood system* for S that specifies which sites are "neighbors" of each site. Then the pair {S, N} defines a graph with nodes for each site in S, and links between neighbors. Define a *clique* to be any subset $c \subseteq S$ where c is either a single site, or a set of sites such that every pair of distinct sites in c are neighbors. Formal definitions are listed below, and Figure 5.1 depicts an example.

$$\begin{aligned}
&\text{Neighborhood of site i:} && N_i = \{j \mid j \in S, j \neq i, i \in N_j\} \\
&\text{Neighborhood System:} && N = \{N_i \mid i \in S\} \\
&\text{Local Configuration:} && w_{N_i} = \{w_i \mid i \in N_i\} \\
&\text{Set of cliques of size 1:} && C_1 = \{i \mid i \in S\} \\
&\text{Set of cliques of size 2:} && C_2 = \{\{i, j\} \mid i \in S, j \in N_i\} \\
&\text{Set of all cliques:} && C = C_1 \cup C_2 \cdots
\end{aligned} \tag{5.2}$$

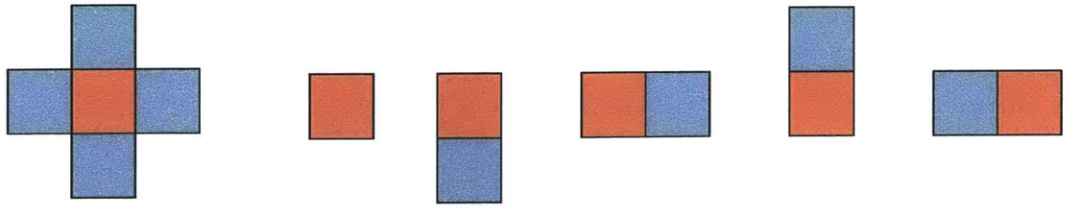

Figure 5.1. Cliques in 2-D. (Left:) A red site and its blue neighbors. (Right:) Set of cliques of size 1 and 2.

5.1.2 Probabilistic Approach to Incorporating Context

5.1.2.1 Local vs. Global Contextual Constraints

Contextual constraints that consider the likelihood of labelings of voxels can be expressed either locally or globally. Local expressions of contextual constraints are cast in the form of conditional distributions $P(w_i \mid w_{S-\{i\}})$ relating the label of a given voxel w_i to its surroundings (every other voxel in the image) $w_{S-\{i\}}$. Global expressions involve the entire image, whose likelihood is represented by the joint distribution $P(w)$, which asserts the probability of the joint event $P(F_1 = w_1, F_2 = w_2, \cdots, F_m = w_m)$.

Table 5.1. Contextual Constraints

Scope	Distribution	Expression
Local	Conditional	$P(w_i \mid w_{S-\{i\}})$
Global	Joint	$P(w)$

5.1.2.2 Conditional vs. Statistical Independence

CDN layer #1 ignored mutual influences between voxels by considering the likelihood of the label of a given voxel $P(w_i)$ independent of the labeling of all other voxels. This assumption of statistical independence permitted the joint distribution to be computed as the product of the conditional probabilities:

$$P(w_i \mid w_{S-\{i\}}) = P(w_i) \qquad (5.3)$$

$$P(w) = \prod_{i \in S} P(w_i)$$

Consequently, the computation of Bayesian classification given the image data x could be computed in a very straightforward manner because maximizing the global MAP classification was equivalent to maximizing each local (one voxel) MAP classification.

$$\begin{aligned} P(w|\ x) &= P(x \mid w)P(w) \\ &= \left(\prod_{i\in S} P(x_i|\ w_i)\right)\left(\prod_{i\in S} P(w_i)\right) \\ &= \prod_{i\in S} P(x_i|\ w_i)P(w_i) \end{aligned} \tag{5.4}$$

The two assumptions that support the middle line of equation 5.4 are summarized in Table 5.2. Based on our imaging model in Chapter 2, we know it is safe to assume the *conditional independence* of $P(x \mid w)$ because MR scan parameters are set to restrict the Gibbs ringing that would cause a voxel's intensity to be influenced by that of its neighbors. Although a voxel's intensity is dependent on only it's own classification, it's classification is dependent on the classifications of its neighbors. This is because MR scan parameters are set to resolve the structures of interest, resulting in some degree of voxel homogeneity. Hence, this chapter must strive to relax the assumption of *statistical independence* of $P(w)$.

Table 5.2. Conditional vs. Statistical Independence. When the assumption of independence is valid, we can substitute the expression under the heading "Equivalence when Independent" in place of the expression under "Distribution". The column labeled "Imaging Model" indicates whether our imaging model specifies that the given type of independence is valid in our application.

Independence	Distribution	Equivalence when Independent	Imaging Model
Conditional	$P(x \mid w)$	$\prod_{i\in S} P(x_i\|\ w_i)$	Valid
Statistical	$P(w)$	$\prod_{i\in S} P(w_i)$	Invalid

As a consequence, two challenges arise:

1. There is no obvious method to deduce the joint distribution $P(w)$ from the set of conditional probabilities $P(w_i \mid w_{S-\{i\}})$, which are subject to interlocking consistency conditions.

2. Conditional priors must be available with neighborhoods large enough to model interesting classes of images, but small enough to ensure trainability and feasible computational loads.

The first challenge has been addressed through a means of specifying the joint distribution directly as a *Gibbs* distribution, which is equivalent to modeling F as a *Markov Random Field.* In the next subsection, we will present this mathematical formalism for modeling the *a priori* probability of contextual dependent patterns. This will permit global inferences to be made based on local properties, which are more directly observed than global information.

Regarding the second challenge, consider a practical example of a problem with exponential time complexity. Given a typical MRI scan consisting of 124 slices with 256x256 resolution, the segmentation problem with M tissue classes yields $M^{8,126,463}$ distinct surroundings $w_{S-\{i\}}$. (We define "surroundings" using this notation as the labeling of every voxel in the image other than the one at location *i*.) Even with spectacular computing speed, fitting a distribution to each of these possible surroundings in such a way as to avoid sampling error, would require a formidable amount of training data. Regardless, Figure 5.2 offers a visual appreciation of the benefit of context. In this thesis, we will overcome the second challenge by restricting the neighborhood size, and relying on higher levels of CDN to resolve some of the associated deficiencies.

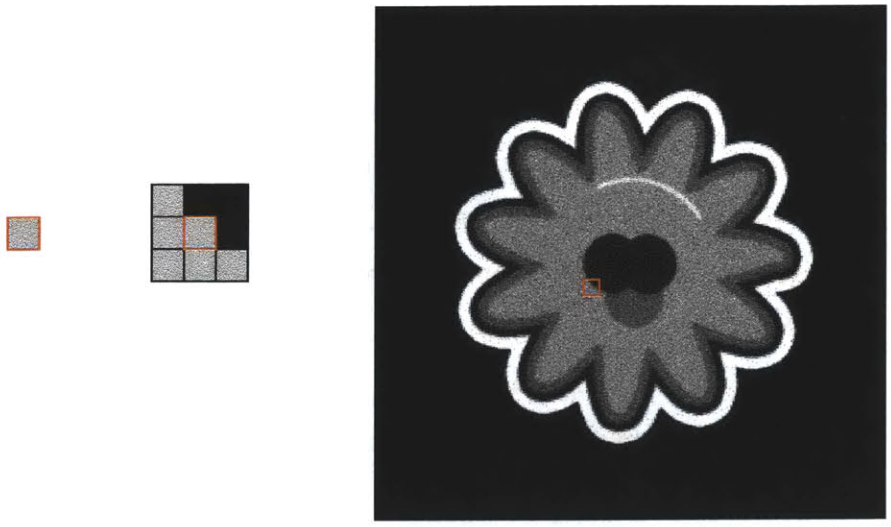

Figure 5.2. Benefit of Context. Mapping a single voxel's gray level to one of a set of discrete labels encounters ambiguity that can be resolved by incorporating neighborhood information. From left to right: (w_i), $(w_i \mid w_{N_i})$, and $(w_i \mid w_{S-\{i\}})$, where the first two are enlarged. Imagine asking yourself what the center voxel (outlined in red) represents given the surroundings shown.

5.1.3 Markov Random Fields

A Markov Random Field (MRF) is a multidimensional generalization of Markov chains defined by conditional probabilities associated with spatial neighborhoods. A random field, F, is a *Markov Random Field* on a lattice S with respect to a neighborhood system N if and only if the following two conditions are satisfied. The first condition, representing *Markovinaity*, asserts that a variable's dependence on all other sites is equivalent to its dependence on only its neighbors. The second condition, *positivity*, is simply a computational formality.

Markovinaity: $$P(w_i \mid w_{S-\{i\}}) = P(w_i \mid w_{N_i}) \tag{5.5}$$

Positivity: $$P(w) > 0, \quad \forall w \in W \tag{5.6}$$

An MRF is *homogeneous* if $P(w_i \mid w_{N_i})$ is independent of the location of site i within the lattice. We will revisit this concept in Section 5.2.3.

Figure 5.3. MRFs at Sea. Lobster bouys, similar to those adorning the wall of Rockport, Massachusetts' "Motif #1", are colored according to the scheme uniquely registered by their owner. Dense arrays of bouys can often be seen peppering the sea off the New England coastline to denote the ownership of the attached lobster traps. The network of bouys can be conceptualized as a Markov random field because neighboring bouys are likely to be colored similarly, but knowing the color scheme of distant buoys is unhelpful in discerning those close at hand.

5.1.4 Gibbs Random Fields

A random field, F, is a *Gibbs Random Field* (GRF) on a lattice S with respect to a neighborhood system N if and only if its configurations satisfy the Gibbs distribution. Through its formulation as a negative exponential, the Gibbs distribution exerts that configurations with lower energy $U(w)$ are significantly more probable. The distribution also allows for an external source to dictate a control parameter called temperature T,

which is a holdover from its origins in modeling molecular systems in statistical physics. T remains useful in computer vision for controlling the sharpness of the distributions during simulated annealing schemes [Geman84]. For example, all configurations have nearly equal probabilities when $T >> \max(U(w))$, but the distributions concentrate on the global energy minima as T approaches 0.

$$P(w) = \frac{\exp\left(-\frac{1}{T}U(w)\right)}{\sum_{w' \in W} \exp\left(-\frac{1}{T}U(w')\right)} \tag{5.7}$$

The denominator, often denoted by Z and referred to as the *partition function* in the literature, is a summation over all possible configurations in order to make the probabilities sum to 1. The energy of a configuration is defined as the sum of clique potentials over C, the set of all cliques on S.

$$U(w) = \sum_{c \in C} V_c(w) \tag{5.8}$$

Since the $V_c(w)$ depends on only those sites i for which $i \in c$, we expand this equation below for the case of cliques of size 1 and 2. The last line of the equation below is true for cliques of any size. This fact is the key to computability: *the joint energy is the sum of the energy at each site.*

$$\begin{aligned} U(w) &= \sum_{i \in C_1} V_1(w_i) + \sum_{i,j \in C_2} V_2(w_i, w_j) \\ &= \sum_{i \in S} V_1(w_i) + \sum_{i \in S} \sum_{j \in N_i} V_2(w_i, w_j) \\ &= \sum_{i \in S} \left(V_1(w_i) + \sum_{j \in N_i} V_2(w_i, w_j) \right) \\ &= \sum_{i \in S} U_i(w_i \mid w_{N_i}) \end{aligned} \tag{5.9}$$

For convenient future reference, we express the local energy separately:

$$U_i(w_i \mid w_{N_i}) = V_1(w_i) + \sum_{j \in N_i} V_2(w_i, w_j) \tag{5.10}$$

The value of each clique potential is determined from application-specific modeling. Observe that $V_c(w)$ can take on a finite set of values when there exists a finite set of label values. A GRF is said to be *homogenous* if $V_c(w)$ is independent of the position of clique c within the lattice, and it is *isotropic* if $V_c(w)$ is independent of the orientation of c.

5.1.5 Markov-Gibbs Equivalence

Although an MRF is characterized by local properties (Markovinaity for w_i) and a GRF is characterized by global properties (Gibbs distribution for w), the two were shown to be equivalent by the Hammersley-Clifford theorem [Besag74]. Both components of the MRF-GRF equivalence are exploited for computing (MRF) and modeling (GRF). The Markovinaity property permits massively parallel computation with one processor per voxel, while the Gibbs distribution provides a convenient formalism for specifying the joint probability $P(w)$ by specifying clique potentials $V_c(w)$ to encode *a priori* knowledge about interactions. Because the major topic for designers is defining the forms and parameters of $V_c(w)$, it is insightful to reproduce one direction of the proof here based on [Li01], and we refer the reader to [Besag74] for the other.

Theorem:

F is a MRF on lattice S with respect to neighborhood system N if and only if F is a GRF on S with respect to N.

Proof of GRF $\Rightarrow$ MRF:

Let $l = \{w_1, w_2, ...w_{i-1}, l_i, w_{i+1}, ...w_m\}$ denote a configuration that is identical to configuration w except perhaps at site i. Then using Bayes' Rule:

$$P(w_i \mid w_{S-\{i\}}) = \frac{P(w_i, w_{S-\{i\}})}{P(w_{S-\{i\}})} = \frac{P(w)}{\sum_{l \in L} P(l)} \tag{5.11}$$

Let $P(w)$ be a Gibbs distribution, and let A be the subset of cliques in C that contain site i, and let B be the subset of C that excludes i:

$$P(w_i \mid w_{S-\{i\}}) = \frac{\frac{1}{Z}\exp\left(-\frac{U(w)}{T}\right)}{\sum_{l_i \in L}\frac{1}{Z}\exp\left(-\frac{U(l)}{T}\right)} \tag{5.12}$$

$$= \frac{\exp\left(-\sum_{c \in C} V_c(w)\right)}{\sum_{l_i \in L}\exp\left(-\sum_{c \in C} V_c(l)\right)}$$

$$= \frac{\exp\left(-\sum_{c \in A} V_c(w)\right)\exp\left(-\sum_{c \in B} V_c(w)\right)}{\sum_{l_i \in L}\exp\left(-\sum_{c \in A} V_c(l)\right)\exp\left(\sum_{c \in B} V_c(l)\right)}$$

Since $V_c(w) = V_c(l)$ for any clique that excludes i, the rightmost exponentials cancel, producing a Gibbs distribution dependent only on neighbors of i:

$$P(w_i \mid w_{S-\{i\}}) = \frac{\exp\left(-\sum_{c \in A} V_c(w)\right)}{\sum_{l_i \in L}\exp\left(-\sum_{c \in A} V_c(l)\right)} \tag{5.13}$$

$$= P(w_i \mid w_{N_i})$$

QED

For future reference, we wish to rewrite equation 5.13 in terms of the local, conditional energy from equation 5.10:

$$P(w_i \mid w_{N_i}) = \frac{\exp\left(-U_i(w_i \mid w_{N_i})\right)}{\sum_{l \in L}\exp\left(-U_i(l \mid w_{N_i})\right)} \tag{5.14}$$

5.2 MRF Design

We have seen that the Hammersly-Clifford theorem reveals that the conditional MRF distributions $P(w_i \mid w_{N_i})$ are synonymous with modeling the joint $P(w)$ as a Gibbs distribution. The question, then, is how should one go about designing the clique potentials for the Gibbs energy function? MRF models favor certain classes of patterns encoded by the designer into the MRF to be associated with higher probabilities. One could argue that this leaves an opportunity for *ad hoc* methods to specify $V_c(w)$ to achieve the desired system behavior. Indeed, as in any optimization problem, the designer selects models or distributions, and solves for the parameters that optimize the solution given the particular model. With respect to the model itself, however, [Geman84] suggests "a general theory of interactive, self-adjusting models that is practical and mathematically coherent may lie far ahead." Unfortunately, any Bayesian classification scheme that employs invalid assumptions of statistical or conditional independence in order to circumvent the computation of the joint likelihood may be considered *ad hoc*.

Given a model, determining its optimal parameters can be achieved through incorporating knowledge accumulated from one or more of the following sources:

- The designer's knowledge of the imaging process
- An EM-based approach to unsupervised segmentation
- Training on manually labeled images

The latter two deserve more discussion, and will be covered next.

5.2.1 MRF Parameter Estimation

[Zhang92] and [Langan92] pioneered the use of mean-field approximations (described later) within an EM-based approach for model-based image segmentation. The idea was that parameters for the Gaussian image intensity model $\mathbf{\Phi}_x = \{(\mu_l, \sigma_l^2) \mid l \in L\}$ are independent of the parameters for the MRF model $\mathbf{\Phi}_w$. This allows the M-step to separate the processes of finding the ML estimate of each set of parameters. To state this

mathematically, the stochastic model based on the composite parameter vector $\mathbf{\Phi} = [\mathbf{\Phi}_x, \mathbf{\Phi}_w]^T$ can be expressed as:

$$P(x, w \mid \mathbf{\Phi}) = P(x \mid w, \mathbf{\Phi}_x) P(w \mid \mathbf{\Phi}_w) \tag{5.15}$$

Then the M-step separately finds the ML estimate of the parameters for the intensity data:

$$\hat{\mathbf{\Phi}}_x = \arg\max_{\mathbf{\Phi}_x} E[\log P(x \mid w, \mathbf{\Phi}_x)] \tag{5.16}$$

$$\mu_l = \frac{\sum_{i \in S} x_i P(w_i = l \mid x_i, \mathbf{\Phi}_x)}{\sum_{i \in S} P(w_i = l \mid x_i, \mathbf{\Phi}_x)}$$

$$\sigma_l^2 = \frac{\sum_{i \in S} (x_i - \mu_l)^2 P(w_i = l \mid x_i, \mathbf{\Phi}_x)}{\sum_{i \in S} P(w_i = l \mid x_i, \mathbf{\Phi}_x)}$$

and the MRF:

$$\begin{aligned} \hat{\mathbf{\Phi}}_w &= \arg\max_{\mathbf{\Phi}_w} E[\log P(w \mid \mathbf{\Phi}_w)] \\ &= \arg\max_{\mathbf{\Phi}_w} E\left[\frac{\exp(-U(w \mid \mathbf{\Phi}_w))}{Z(\mathbf{\Phi}_w)}\right] \\ &= \arg\max_{\mathbf{\Phi}_w} \left[\left\langle -U_{MF}(w \mid \mathbf{\Phi}_w) \right\rangle - \log Z_{MF}(\mathbf{\Phi}_w)\right] \end{aligned} \tag{5.17}$$

Where U_{MF} and Z_{MF} are the mean field approximations to U and Z that will be derived later.

5.2.2 MRF Parameter Training

While the previous section entertained unsupervised image segmentation, our application of recognizing deviations from normalcy requires collecting *a priori* knowledge of normalcy. Therefore, the parameters for our MRF model are derived through training on a manually labeled image. Our model emphasizes continuity, but instead of unilaterally discouraging pairs of differing labels, we will allow pairs to have a certain labeling according to its observed presence in the training population. Following the literature's notation of reserving **J** to denote "bond strength", define **J** to be the pairwise interaction matrix. (For clarity of function, we can refer to it informally as the "**J**ives with" matrix.) The square matrix **J** contains one row for each label, where each row i contains the

probabilities of the given label occurring in a clique with various other labels. (The index i here refers to an index into matrix **J**, not an image).

$$\mathbf{J}(i,j) = P(w_i \mid w_j) \tag{5.18}$$

The probabilities within **J** are discovered through training on example data. Training consists of incrementing as expressed below, followed by normalizing each row to sum to 1.

$$\sum_{i \in S} \sum_{j \in N_i} inc(\mathbf{J}(i,j)) \tag{5.19}$$

Note that **J** is not symmetric, as proven in Figure 5.4.

$$\mathbf{J} = \begin{bmatrix} & Gray & Blue \\ Gray & 0.1 & 0.9 \\ Blue & 0.1 & 0.9 \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} & Gray & Blue \\ Gray & 0.9 & 0.1 \\ Blue & 0.9 & 0.1 \end{bmatrix}$$

Figure 5.4. Assymetric J. Pairwise interaction matrices computed from the images on the left and right, are displayed to the right of their respective images. The were generated at random with a 9:1 probability ratio. The experiment demonstrates that the more blob-like an image is, the more **J** approaches the identity matrix.

Because low Gibbs energy corresponds to likely configurations, the pairwise clique potential can be expressed as a function of the inverse of the probabilities encoded into **J**. The logic underlying our selection of *f()* for our CDN implementation will be discussed in Section 5.5.

$$V_2(w_i, w_j) = -f(\mathbf{J}(i,j)) \tag{5.20}$$

For intractably large neighborhood sizes, an efficient approximation was proposed by [Pickard77] to estimate the joint probability of a large neighborhood by assuming conditional independence:

$$P(w_{N_i}) = P(w_i)\prod_{j\in N_i} P(w_j \mid w_i) \qquad (5.21)$$
$$= P(w_i)\prod_{j\in N_i} \mathbf{J}(i,j)$$

5.2.3 Mapping Image Space to Model Space

We now describe the framework of CDN layer #2 in terms of Chapter 4's paradigm of probabilistic mappings between image space and model space. In layer #2, the models are the class interaction matrices. If there is one **J** for the entire image, as in [Kapur99], then the literature refers to it as a "homogenous MRF". From our vantage point described in Table 4.3, we refer to it as an "all-to-one" mapping from image space to model space. Alternatively, an MRF with multiple matrices, where one particular matrix is specified for each image partition, is considered a "heterogeneous MRF" in the literature. This corresponds to the "many-to-one" mapping in Table 4.3.

The results of training a "many-to-one" mapping compared to an "all-to-one" mapping are demonstrated in Figures 5.5 and 5.6. The image is roughly partitioned into two regions: cortex and sub-cortex. The mapping is computed respective to CDN layer 1's mapping (the spatially varying prior). So in our implementation, it is brought into correspondence with the image during the same rigid registration step. Since ventricles border white matter in the anterior portions, and gray matter in the posterior regions, more accurate classifications would result with more than two class interaction matrices, which we leave for future work in Chapter 7.

$$\mathbf{J} = \begin{bmatrix} & SCA & WM & GM & CSF & VES \\ SCA & .94 & .00 & .00 & .06 & .00 \\ WM & .00 & .95 & .03 & .01 & .01 \\ GM & .00 & .06 & .88 & .06 & .00 \\ CSF & .06 & .01 & .06 & .87 & .00 \\ VES & .00 & .36 & .00 & .00 & .64 \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} & SCA & WM & GM & CSF & VES \\ SCA & .94 & .00 & .00 & .06 & .00 \\ WM & .00 & .96 & .03 & .00 & .01 \\ GM & .00 & .07 & .87 & .06 & .00 \\ CSF & .08 & .00 & .08 & .84 & .00 \\ VES & .00 & .36 & .00 & .00 & .64 \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} & SCA & WM & GM & CSF & VES \\ SCA & .00 & .00 & .00 & .00 & .00 \\ WM & .00 & .98 & .01 & .01 & .00 \\ GM & .00 & .04 & .94 & .02 & .00 \\ CSF & .00 & .03 & .01 & .96 & .00 \\ VES & .00 & .00 & .00 & .00 & .00 \end{bmatrix}$$

Figure 5.5. Many-to-One Mapping. A heterogenous MRF is implemented by mapping image space to model space (the set of class interaction matrices). Compare the **J**'s computed from the portions of the image shaded red in the top row (complete), middle row (cortex) and bottom row (subcortex). The tissue classes included were scalp (SCA), white matter (WM), gray matter (GM), cereobro-spinal fluid (CSF), and vessel (VES).

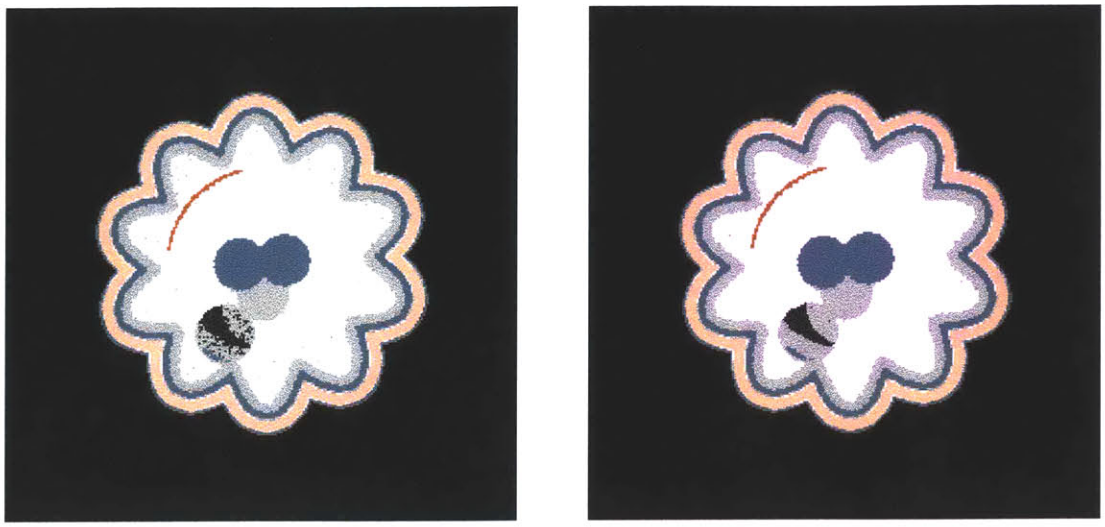

Figure 5.6. Impact of Heterogeneity. The results of using a heterogenous MRF (right) show improvement over using a MAP classification (left) by correcting misclassifications in the white matter, gray matter, and tumor. However, the heterogenous MRF showed little improvement over the homogenous MRF in this case. Greater benefit would be realized by using a model space with higher dimensionality.

5.3 MRF Optimization

The goal of optimization is to find the most probable configuration $\hat{w}$ of a given MRF, which is equivalently the lowest energy configuration of a GRF:

$$\text{MRF:} \quad \hat{w} = \arg\max_{w \in W} P(w) \tag{5.22}$$

$$\text{GRF:} \quad \hat{w} = \arg\min_{w \in W} U(w) \tag{5.23}$$

5.3.1 Optimization Methods

Finding the global minimum often requires an exhaustive search to find all of the local minima, followed by a comparison of these from which to select the global minimum. In MRF optimization, a brute force search would compute $P(w)$ for each permissible configuration w, and select the instantiation corresponding with the highest $P(w)$. Brute force search is intractable due to this problem's combinatorial search space: m voxels each with M possible labels. Therefore, a wide variety of approximate methods have been

applied, and we refer the reader to [Li01] for a comprehensive review. In general, optimization techniques vary depending on the following four categories:

- **Constrained or Unconstrained** – constrained techniques are limited to searching a subspace of the total search space
- **Continuous or Discrete** – techniques that optimize a continuous set of labels can compute and descend the gradient of the energy function such that the next configuration w' is displaced from the current configuration f by a fractional step λ along the energy gradient: $w' \leftarrow w - \lambda \nabla E$
- **Deterministic or Stochastic** – stochastic methods generate the next configuration at random from sampling a distribution over w_{N_i}
- **Locally or Globally Optimal** – locally optimal, or greedy, methods seek a lower energy configuration at each step by requiring $E(w') < E(w)$. Globally optimal methods relax this requirement subject to certain conditions. In this way, they allow for an escape mechanism out of local minima.

5.3.2 Optimization of MAP-MRF Problems

Note that solving equation 5.22 or 5.23 would determine the most probable *a prior* configuration. But within the context of this thesis, the object of optimization is to find the *maximum a posteriori* configuration. Hence, consider the posterior probability to be a single Gibbs distribution:

$$
\begin{aligned}
P(w \mid x) &= \frac{P(x \mid w)P(w)}{\sum_{w' \in W} P(x \mid w)P(w)} \\
&= \frac{\exp(\log P(x \mid w))\left(\frac{1}{Z_p}\exp(-U(w))\right)}{\sum_{w' \in W} \exp(\log P(x \mid w))\left(\frac{1}{Z_p}\exp(-U(w))\right)} \\
&= \frac{\exp(-(U(w) - \log P(x \mid w)))}{\sum_{w' \in W} \exp(-(U(w) - \log P(x \mid w)))} \\
&= \frac{1}{Z}\exp(-U(w \mid x))
\end{aligned}
\tag{5.24}
$$

The last line is clearly a Gibbs distribution, and it energy function is:

$$
U(w \mid x) = U(w) - \log P(x \mid w) \tag{5.25}
$$

Therefore, maximizing the posterior probability can be equivalently formulated as minimizing the posterior energy, which is derived from combining the likelihood and prior energies. We can now recast equation 5.23 to solve for the maximum *a posteriori* configuration instead of the *a prior* configuration. We will refer back to this equation:

$$
\begin{aligned}
\hat{w} &= \arg\min_{w \in W} U(w \mid x) \\
&= \arg\min_{w \in W} [U(x \mid w) + U(w)]
\end{aligned}
\tag{5.26}
$$

5.4 Factorizing the Joint Distribution

As mentioned earlier, solving equation 5.26 is an exponentially complex problem because there exist a combinatorial number of elements in configuration space W. Therefore, we are interested in approximations that *factorize* the joint probability into a product of local conditional probabilities. Table 5.3 summarizes three approaches to factorization, which requires decoupling the interactions between sites. While the first technique considers no neighborhood interactions, the next two techniques solve the consistency relations between different variables through an iterative scheme. Given an initial configuration, the values of each variable are updated sequentially as if they were decoupled from the other variables. These two techniques will be covered next.

Table 5.3. Approximations for Factorizing the Joint Distribution

Assumption	Local Conditional	Joint Prior
Statistical Independence	$P(w_i \mid w_{S-\{i\}}) = P(w_i)$	$P(w) = \prod_{i \in S} P(w_i)$
Iterated Condition Modes	$P(w_i \mid w_{S-\{i\}}) = P(w_i \mid w_{N_i})$	$P(w) \approx \prod_{i \in S} P(w_i \mid w_{N_i})$
Mean Field Approximation	$P(w_i \mid w_{S-\{i\}}) = P(w_i \mid \overline{w}_{N_i})$	$P(w) \approx \prod_{i \in S} P(w_i \mid \overline{w}_{N_i})$

5.4.1 Iterated Condition Modes

[Besag86] developed Iterated Condition Modes (ICM) as a computationally efficient alternative to the stochastic and globally optimal method of simulated annealing in [Geman84]. The idea is to iteratively update the current labeling at voxel i in light of all available information, which includes the image data x and the current labeling elsewhere $w_{S-\{i\}}$. We derive the following *update equation* using Bayes' rule, the assumption of conditional independence in equation 5.4, and the assumption of Markovinaity in equation 5.5:

$$
\begin{aligned}
P(w_i \mid x, w_{S-\{i\}}) &\propto P(x \mid w_i, w_{S-\{i\}}) P(w_i \mid w_{S-\{i\}}) \\
&= P(x \mid w) P(w_i \mid w_{S-\{i\}}) \\
&= \left(\prod_{i \in S} P(x_i \mid w_i) \right) P(w_i \mid w_{S-\{i\}}) \\
&\propto P(x_i \mid w_i) P(w_i \mid w_{S-\{i\}}) \\
&= P(x_i \mid w_i) P(w_i \mid w_{N_i})
\end{aligned}
\tag{5.27}
$$

Therefore, ICM sidesteps the combinatorial computation of maximizing the joint probability $P(w \mid x)$, in favor of maximizing the local conditional probabilities $P(w_i \mid x_i, w_{N_i})$ sequentially. As a consequence, ICM is a deterministic, greedy search algorithm that converges to a local minimum.

5.4.2 Mean Field Approximation

Interestingly, the origin of MRFs within statistical physics suggests insightful analogies with CDN-based image segmentation. We note these parallels by adding the italicized text to the following from [Parisi88]:

> The aim of statistical mechanics *(image segmentation)* is to derive thermodynamic properties *(classifications)* of macroscopic bodies *(image regions or neighborhoods)* starting from a description of the motion *(intensity)* of microscopic components *(voxels)*. This would be an impossible and hopeless task if one took the normal approach of mechanics *(brute-force search)*, since the number of degrees of freedom *(size of configuration space)* is so huge: probabilistic methods are mandatory. The problem can be divided into two parts: (a) Find the probability distribution of microscopic components *(local conditional distributions)*. (b) Compute the macroscopic properties of the system *(image)* from the microscopic probability distributions.

Given these similarities, it is appropriate to adopt the mean field approximation from statistical physics. The intuition behind mean field theory is that within dense random fields, each random variable is subject to influences from several other variables. If each influence is weak, and the influences are additive (such as the noise in our imaging model from Chapter 2), then fluctuations from different sites tend to cancel each other, as shown here for weak fluctuations α at neighboring sites A and B:

$$\begin{aligned} &\text{Additive fluctuations:} && (A+\alpha)+(B-\alpha) = A+B \\ &\text{Multiplicative fluctuations:} && (A+\alpha A)+(B-\alpha B) = A+B+error \end{aligned} \tag{5.28}$$

This permits each variable to be roughly characterized by its mean value. Because the mean value of each variable is unknown and related to the mean values of other variables,

finding the mean field at site i requires finding the mean field at its neighbors. Solving these consistency relations between different variables can be accomplished through iteration.

Figure 5.7. Smooth Sailing with the Mean Field Approximation. MIT sailors begin forming a long, straight line of boats across the Charles River. The Markov approach would advise a sailor to match course by monitoring only the 2 boats immediately fore and aft. Since each boat experiences different instantaneous wind and waves, exactly mimicing the steering moves of neighboring boats would fail to hold the line. Instead, sailors rely on the mean field approximation by matching course with the average observed headings of their neighboring boats.

To summarize these ideas mathematically, the mean field approximation assumes that the influence of $w_j, j \neq i$ can be approximated by the influence of $\overline{w}_j$. This permits the factoring of the joint probability in Table 5.3, and changes equation 5.10 to:

$$U_i(w_i \mid \overline{w}_{N_i}) = V_1(w_i) + \sum_{j \in N_i} V_2(w_i, \overline{w}_j) \tag{5.29}$$

Subsequently, equation 5.14 changes to:

$$P(w_i \mid \overline{w}_{N_i}) = \frac{\exp\left(-U_i(w_i \mid \overline{w}_{N_i})\right)}{\sum_{l \in L} \exp\left(-U_i(l \mid \overline{w}_{N_i})\right)} \tag{5.30}$$

Next, we need to compute the statistical average at each site, for which we can rely on the formula for expected value:

$$\overline{w} = \sum_{w \in W} wP(w) \tag{5.31}$$

Then we apply the mean field approximation stated in Table 5.3, and exploit the factorization that it allows:

$$\overline{w}_i \approx \sum_{w_i \in L} w_i P(w_i \mid \overline{w}_{N_i}) \tag{5.32}$$

Regarding representation, this thesis will adopt the standard convention in the literature of utilizing indicator vectors. Then $\mathbf{w}_i$ denotes one of the basis vectors that completely span the orthogonal state space **L** of M-dimensional *indicator vectors*:

$$\mathbf{L} = \left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\} \tag{5.33}$$

Therefore, the kTH component $\overline{w}_{ik}$ of $\overline{\mathbf{w}}_i$ represents the probability that site *i* is a member of the kTH class. Given this representation, we can complete our derivation of a mean field update equation by substituting equation 5.30 into 5.32:

$$\begin{aligned} \overline{\mathbf{w}}_i &\leftarrow \sum_{\mathbf{w}_i \in \mathrm{L}} \mathbf{w}_i P(\mathbf{w}_i \mid \overline{\mathbf{w}}_{N_i}) \\ \overline{w}_{ik} &\leftarrow P(\mathbf{w}_i \mid \overline{\mathbf{w}}_{N_i}) \\ \overline{w}_{ik} &\leftarrow \frac{\exp\left(-U_i(\mathbf{w}_i \mid \overline{\mathbf{w}}_{N_i})\right)}{\sum_{\mathbf{l} \in \mathrm{L}} \exp\left(-U_i(\mathbf{l} \mid \overline{\mathbf{w}}_{N_i})\right)} \end{aligned} \tag{5.34}$$

For further reading on mean field theory, refer to books by [Chandler87] and [Parisi88]. In addition to mean values, computations of correlations and moments are derived in [Elfadel93]. For globally optimal image segmentations, simulated annealing schemes that employ the mean field approximation are presented in [Lin97], [Noda99], and [Cho00].

5.5 Experimental Comparisons

In this section, we evaluate the efficacy of the algorithms presented in this chapter. Experiments will showcase the differences between ICM and mean field approximations, explore the sensitivity to MRF model parameters, and demonstrate the value of MRF theory relative to simple smoothing.

5.5.1 Simple Smoothing

For comparison with MRF theory, we experiment with an unsophisticated method that is conceptually appealing for its simplicity. A plausible approach to smoothing a discrete labeling might involve taking a majority vote of the labels within each neighborhood:

$$w_i \leftarrow \arg\max_{l \in L} v_i(l) \tag{5.35}$$

Where the voting function is:

$$v_i(l) = \sum_{j \in \{N_i \cup i\}} \delta(l - w_j) \tag{5.36}$$

Where we employ the discrete Dirac delta function:

$$\delta(x) = \begin{Bmatrix} 1, x = 0 \\ 0, else \end{Bmatrix} \tag{5.37}$$

Observe that simple smoothing considers only the hard classifications as if the image data had been discarded after the first iteration. Consequently, running for a sufficient number of iterations will result in a complete "filling-in" of all gaps between structures (see Figure 5.8). While this could be useful prior to performing operations such as CDN layer #3's computation of shape descriptors, there is a tendency to make gross classification errors.

5.5.2 ICM

In contrast to simple smoothing, we now seek a solution grounded in MRF theory. Given the ICM optimization algorithm, we need only design the clique potentials. We will base our design on that which will accommodate convenient comparisons with MAP classification and simple smoothing.

Beginning with an initial labeling that was computed without neighborhood interactions, the ICM iteration proceeds to update the label at each voxel sequentially. We derive this update equation by beginning with equation 5.27, and substituting equation 5.14 and 5.10:

$$\begin{aligned} P(w_i \mid x_i, w_{N_i}) &\propto P(x_i \mid w_i) P(w_i \mid w_{N_i}) \\ &\propto P(x_i \mid w_i) \exp\left(-U_i(w_i \mid w_{N_i})\right) \\ &= P(x_i \mid w_i) \exp\left(-V_i(w_i) - \sum_{j \in N_i} V_2(w_i, w_j)\right) \\ &= P(x_i \mid w_i) \exp\left(-V_i(w_i)\right) \exp\left(-\sum_{j \in N_i} V_2(w_i, w_j)\right) \end{aligned} \tag{5.39}$$

Observe that this would perform MAP classification given the following designs of the clique potentials:

$$\begin{aligned} V_1(w_i) &= -\ln(P(w_i)) \\ V_2(w_i, w_j) &= 0 \end{aligned} \tag{5.40}$$

Given the above choice for V_1, we observe that a non-zero choice for V_2 effectively appends an additional term for prior knowledge about neighborhoods to the MAP classification equation:

$$P(w_i \mid x_i, w_{N_i}) \propto \text{(likelihood) * (singleton prior) * (neighborhood prior)} \tag{5.41}$$

To conceptualize how these terms influence the computation, let us revisit our choice for the design of V_1. Setting V_1 to be the negative prior probability would have resulted in a middle term of equation 5.39 that varies exponentially within the range of $[1,e]$. By instead choosing the negative log probability, the term varies linearly on $[0,1]$. The form

of variation – exponential or linear – is of minor importance as long as the variation is monotonic. Those seeking efficient implementations might prefer the linear design to avoid the expensive computation of the exponential function. More importantly, the critical difference in the two designs is that the latter has the ability to "sink" the total posterior probability to zero – a multiplicative singularity. We believe the "sinking" property is desirable for the singleton prior because our spatially varying priors from Chapter 4 contain zeros to express impossible locations for certain tissues. We do not, however, desire the "sinking" property for our neighborhood prior, so we will keep its exponential form. Alternatively, we could have followed the implementation of [Kapur99] to introduce a parameter α on [0,1) that controls the strength of the influence of the neighborhood prior by replacing it with the following term:

$$(1\text{-}\alpha + \alpha^* \text{ (neighborhood prior)}) \tag{5.42}$$

Now that we have examined the differences between ICM and MAP, we wish to compare ICM with simple smoothing. Toward this end, we seek a design for V_2 that allows for the closest comparison possible. We optimize based on equation 5.26:

$$w_i \leftarrow \arg\max_{l \in L} \exp\left(\ln P(x_i \mid l) + \ln P(l) - \sum_{j \in N_i} V_2(w_i, w_j) \right) \tag{5.44}$$

$$w_i \leftarrow \arg\max_{l \in L} \left(Soft(l, w_i, x_i) + \sum_{j \in N_i} -V_2(w_i, w_j) \right)$$

Where *Soft()* denotes a function of the "soft" values of probabilities rather than the "hard" values of a discrete delta function. Given this form, we can now solve for V_2:

$$\begin{aligned} V_1(w_i) &= -\ln(P(w_i)) \\ V_2(w_i, w_j) &= -\delta(w_i - w_j) \end{aligned} \tag{5.45}$$

Comparing this result with equation 5.36 reveals that the advantage of ICM over simple smoothing is that ICM's iterations continue to use "soft" values (likelihood and prior probabilities) for the single-site potentials. ICM also continues to include the image data at every iteration in contrast to simple smoothing, which discards the image following the initial classification step.

$$\text{Simple smoothing:} \quad w_i \leftarrow \arg\min_{l \in L} -\delta(l - w_i) - \sum_{j \in N_i} \delta(l - w_j) \tag{5.46}$$

$$\text{ICM:} \quad w_i \leftarrow \arg\min_{l \in L} -Soft(l, w_i, x_i) - \sum_{j \in N_i} \delta(l - w_j)$$

As a further step away from simple smoothing, we can replace the "hard" version of V_2 with the "soft" probabilities gathered from the training data. We elect not to use *ln(**J**)* to deny the neighborhood term the "sinking" property. Then:

$$\begin{aligned} V_1(w_i) &= -\ln P(w_i) \\ V_2(w_i, w_j) &= -\mathbf{J}[w_i, w_j] \end{aligned} \tag{5.47}$$

Similarly, we will next see how mean field techniques use "soft" values at the neighboring sites instead of ICM's use of hard classifications there.

5.5.3 Mean Field

For the closest comparison with ICM, we will adopt the following model to express equation 5.29 (mean field version of 5.10):

$$\begin{aligned} V_1(\mathbf{w}_i) &= -\ln P(\mathbf{w}_i) \\ V_2(\mathbf{w}_i, \overline{\mathbf{w}}_j) &= -(\mathbf{w}_i^T \mathbf{J} \overline{\mathbf{w}}_j) \end{aligned} \tag{5.48}$$

Where the quadratic product can be expanded using $\mathbf{J}_{\mathbf{w}_i}$ to denote the $\mathbf{w}_i^{\text{TH}}$ row of $\mathbf{J}$:

$$\sum_{j \in N_i} V_2(\mathbf{w}_i, \overline{\mathbf{w}}_j) = -\sum_{j \in N_i} \mathbf{w}_i^T \mathbf{J} \overline{\mathbf{w}}_j = -\sum_{j \in N_i} \mathbf{J}_{\mathbf{w}_i}^T \overline{\mathbf{w}}_j = -\sum_{j \in N_i} \sum_{l \in L} J_{w_i l} \overline{w}_{jl} \tag{5.49}$$

Beginning with an initial labeling that was computed without neighborhood interactions, the iteration proceeds to update the label at each voxel sequentially. We

derive this update equation very similarly to how we derived it for ICM in equation 5.39, except that now we substitute equation 5.30 (mean field version of 5.14):

$$\begin{aligned} P(\mathbf{w}_i \mid x_i, \overline{\mathbf{w}}_{N_i}) &\propto P(x_i \mid \mathbf{w}_i) P(\mathbf{w}_i \mid \overline{\mathbf{w}}_{N_i}) \\ &\propto P(x_i \mid \mathbf{w}_i) \exp\left(-U_i(\mathbf{w}_i \mid \overline{\mathbf{w}}_{N_i})\right) \\ &= P(x_i \mid \mathbf{w}_i) \exp(-V_1(w_i)) \exp\left(-\sum_{j \in N_i} V_2(\mathbf{w}_i, \overline{\mathbf{w}}_j)\right) \\ &= P(x_i \mid \mathbf{w}_i) P(\mathbf{w}_i) \exp\left(\sum_{j \in N_i} \mathbf{w}_i^T \mathbf{J} \overline{\mathbf{w}}_j\right) \end{aligned} \tag{5.51}$$

We then optimize based on equation 5.34 (which is the mean-field version of 5.26 for ICM). This is a major departure from equation 5.46 for ICM and simple smoothing optimization, and the reason is that we are computing a mean, rather than a maximum or minimum, configuration.

$$\overline{w}_{ik} \leftarrow \frac{P(x_i \mid \mathbf{w}_i) P(\mathbf{w}_i) \exp\left(\sum_{j \in N_i} \mathbf{J}_{\mathbf{w}_i}^T \overline{\mathbf{w}}_j\right)}{\sum_{\mathbf{l} \in \mathbf{L}} P(x_i \mid \mathbf{l}) P(\mathbf{l}) \exp\left(\sum_{j \in N_i} \mathbf{J}_{\mathbf{l}}^T \overline{\mathbf{w}}_j\right)} \tag{5.53}$$

Table 5.4 summarizes the formulation of the Simple Smoothing, ICM, and Mean Field approximations. Table 5.5 and Figure 5.8 compare the results on synthetic data where ground truth is known.

Table 5.4. Comparison of MRF Algorithms. Consider a *hard* function to be a delta function instead of a probability, and a *hard* function parameter to be a classification instead of a probability.

Algorithm	Single site	Neighborhood Function	Neighborhood Parameters
Simple Smoothing	Hard	Hard	Hard
ICM without training	*Soft*	Hard	Hard
ICM with training	*Soft*	*Soft*	Hard
Mean Field with training	*Soft*	*Soft*	*Soft*

Table 5.5. Comparison of MRF Results. Calculations are made from Figure 5.8.

Algorithm	# Incorrect Voxels	% of MAP	# Incorrect Excluding PVA	% of MAP	Time (ms)	% of MAP
MAP	1908	100	313	100	120	100
Simple Smooth	1069	56	83	27	270	225
ICM	1614	84	128	41	400	333
ICM trained	1627	85	140	45	400	333
MF	1606	84	127	41	1270	1058
MF trained	1617	85	136	43	1270	1058

The above table reveals that Simple Smoothing outperformed ICM and MF in both accuracy and run time. The reason for this surprising performance is that these are the "healthy" phantoms exhibiting tissues of perfect piecewise homogeneity. While simple smoothing excels in simple applications, mean field will be the algorithm of choice for our application in Chapter 6. Figure 5.8 demonstrates the difference in the spatial distribution of the misclassifications. While Simple Smoothing created a block of misclassified white matter, ICM and MF left speckle by being more sensitive to the original gray values. Additionally, we note that training over large global regions produced no measurable improvement in correcting local misclassifications.

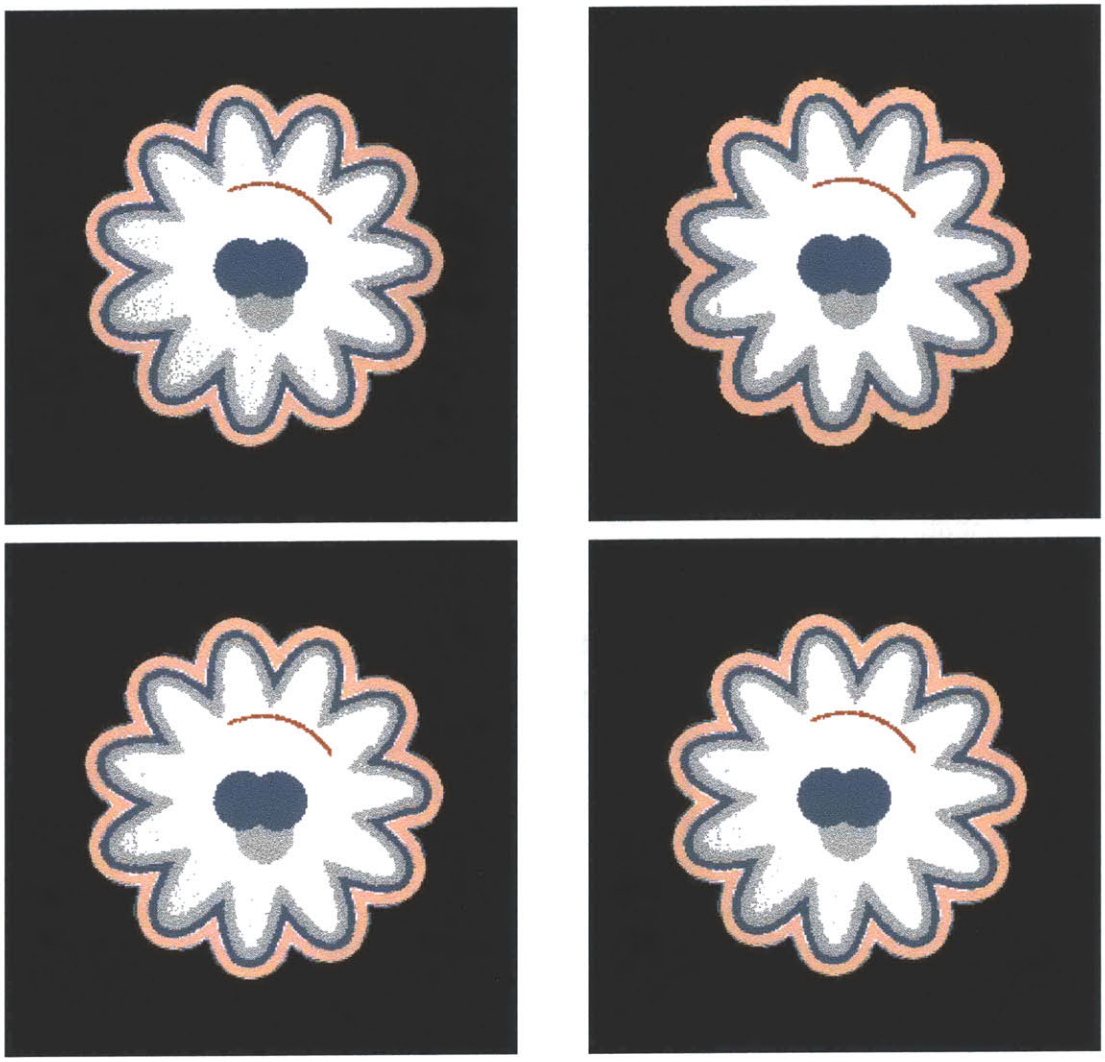

Figure 5.8. MRF Results. The input image used was the one corrupted with a linear bias ramp to produce more salt-and-pepper noise. (Top Left:) classification using Chapter 4's MAP. (Top Right:) Simple smoothing has filled in all gaps between structures, producing a very smooth, albeit erroneous segmentation. (Bottom Left:) ICM corrects without making the errors that SS does. (Bottom Right:) Mean field is only slightly better than ICM, yet runs several times slower (to perform the inner product at each voxel).

5.6 Recognizing Deviations from Normalcy

Armed with the second layer of CDN, we can now revisit the results of Chapter 4 computed with only one layer. The main improvement is that the consideration of context on the neighborhood level serves the following two purposes:

1. Reduces the "salt and pepper" noise. This is critical for preparing the data for Layer #3's analysis of region-level properties, which are sensitive to misclassification.
2. Corrects misclassifications caused by partial volume artifacts.

Naturally, it is desirable to achieve the aforementioned two improvements on the classification of pathology as well as healthy tissue. To include pathology in the MRF processing, we add a class of "weights" as follows:

1. Compute the soft weights for M healthy tissues as in Chapter 4.
2. Compute the probability of pathology as in Chapter 4.
3. Combine these two sets of $(M+1)$ weights, and renormalize them.
4. Perform MRF processing on the complete set of weights.

Furthermore, to facilitate bi-directional communication between the two CDN layers, we perform a few "Outer Iterations". Within each such iterations, a number of EM iterations are performed followed by a few MRF iterations. We converged on the iteration schedule reported in Table 5.5 to maximize both efficacy and efficiency.

Table 5.6. Bi-directional Communication between CDN Layers. Layer #1 passes its result to Layer #2, which returns its result back for re-processing by Layer #1.

Outer Iteration #	EM Iterations	MRF Iterations
1	10	3
2	1	3
3	1	3

Figure 5.9 showcases the impact of Layer #2, and Figures 5.10-5.12 depict the results at each iteration to offer insight into the origin of the results of Figure 5.9.

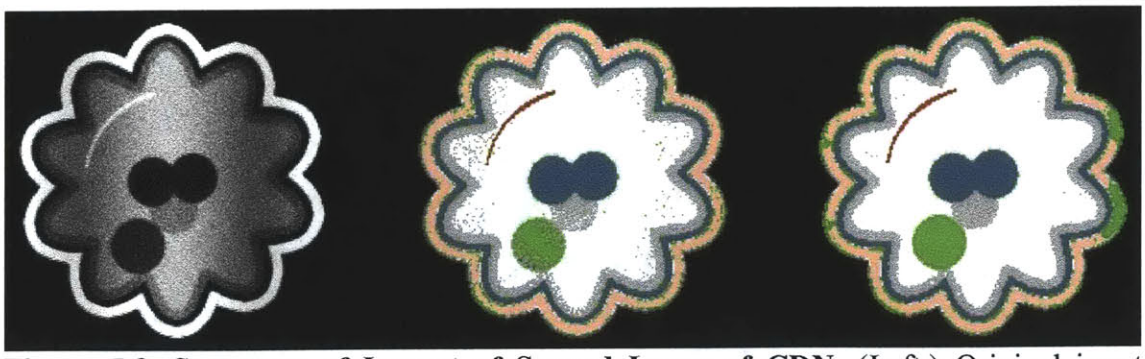

Figure 5.9. Summary of Impact of Second Layer of CDN. (Left:) Original input image that has been corrupted by a sinusoidally-varying bias field. (Center:) Result of segmentation using only Layer #1. (Right:) Result of integrating Layer #1 with Layer #2. Observe the strong reduction of noisy scatter.

Figure 5.10 shows the stages of intermediate results encountered during the computation of the center image of Figure 5.9. The processing is performed by a system consisting only of CDN Layer #1. At each EM iteration, the tissue classification, bias field estimation, and probability of pathology are computed. The probability of pathology weights the bias field estimation during the next iteration. Observe that the bias field is correctly characterized except at the object fringes. The fringe artifacts are the result of our computational speed-up where we perform no processing outside the patient's scalp. This is an acceptable trade-off in our application where skin segmentation is irrelevant.

Figure 5.10. EM Iterations with Layer #1 Only. From left to right, and top to bottom, are the results following EM iterations #1-7 and 10. Depicted are pathology (green), white matter (white), gray matter (gray), CSF (blue), and scalp (tan).

Figure 5.11 begins to illustrate the intermediate results of a system consisting of CDN Layers #1 and #2. As above, the system progressively discovers the full extent of the bias field based on the assumption that it is smoothly varying. Observe that the classification of the voxels within the tumor largely pass through these iterations unchanged. This is a combination of the facts that the tumor tissue is not slowly varying from healthy tissue, and that the tumor tissue's deviation from normalcy is weighting the bias field computation.

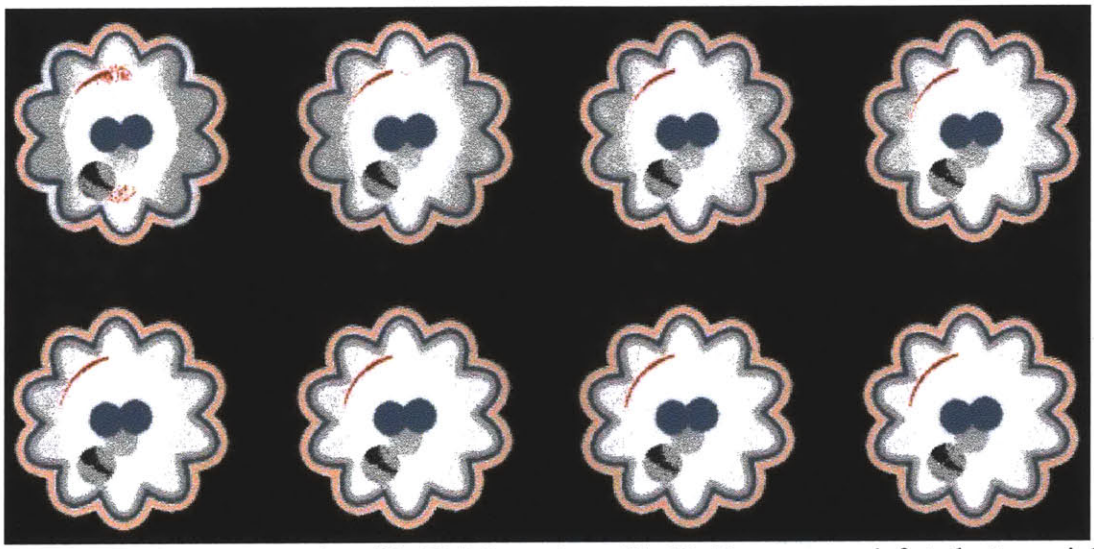

Figure 5.11. Outer Iteration #1, EM Iterations #1-10. From upper left to bottom right are the results after EM iterations 1,2,3,4,5,6,7,10. This is the first processing performed by Layer #1 before the probability of pathology is computed, and before Layer #2.

Figure 5.12 completes the intermediate results of CDN Layer #1-2 that Figure 5.11 began to show. First, observe the twin components of the impact of the 3 iterations of Layer 2's MRF. From left to right, the "salt and pepper" noise is reduced in the healthy tissues while the scattered distribution of abnormal voxels (green) coalesce into tumor masses. Second, observe the impact of Layer 2 providing feedback to Layer 1. This bi-directional communication is evident across the rows of the figure. The more correct classifications are enabling more correct bias field estimation, producing better tumor delineation. Flaws still remain in the ambiguity between bias field and pathology which is largely, but never completely, resolved.

Figure 5.12. MRF Iterations during each Outer Iteration. From left to right are the results following each MRF iteration of Layer #2. Each row corresponds to a different Outer Iteration, proceeding from top to bottom. Note the impact of Layer #2 communicating back to Layer #1 (the difference between the top and bottom rows).

5.7 Chapter Summary

We presented a review of established techniques for taking a probabilistic approach to incorporating immediate context into the segmentation paradigm. We created the method of simple smoothing as a "straw man" for pointing out the benefit of the mathematical formalism of MRF modeling. We used the difference between ICM and simple smoothing as an analogy for understanding the difference between mean field approximations and ICM. Finally, we performed experiments to analyze the impact of the second layer of CDN.

To summarize the important principles asserted in this chapter:

5.1 Our imaging model suggests assumptions of conditional independence, but not statistical independence.

5.2 There are two main purposes of CDN layer #2: reduce "salt and pepper noise", and correct misclassifications due to partial volume artifacts.

5.3 The simple smoothing, ICM, and Mean Field approximations use progressively "softer" functions and function parameters – moving from discrete mathematics to probabilities.

5.4 Pathology is included in layer #2 by relaxing the weights computed by normalizing the combination of the posterior probabilities and the probability of pathology.

5.5 Bi-directional communication between layers #1 and #2 can be achieved with 2-3 outer iterations.

Chapter 6

CDN Layers 3-5: Intra-Structure, Inter-Structure, and Supervisory Classification

In this chapter, we introduce the top three layers of our framework for Contextual Dependency Networks. While the bottom two layers classify voxels based only on their immediate context, the top three layers consider much broader contextual information to see the "big picture".

Recall how Figure 5.2 illustrated the value of context in the image segmentation problem. We have already referred to its benefit in resolving ambiguity by converting an ill-posed problem to a well-posed one. Moreover, we especially need to rely on context given our approach of recognizing deviations from normalcy. Other methods train on tumors by learning from many examples. Although not applied to tumors, [Miller02] introduced a method of training from one example for hand-written character recognition. The idea was that by studying the variability of a set of known characters, a novel character could be recognized by assuming it had similar variability. By comparison, we are learning tumors not from many examples, and not from one example, but rather, we are *learning from zero examples.* Our algorithm must have sufficient knowledge of healthy brains to identify any pathology at first sight. This is akin to the FBI identifying counterfeit money not by studying all the possible instances of fakery, but by thoroughly studying the genuine article.

Incorporating context is problematic in two aspects: knowledge representation and information processing. In the present implementation of our framework, we seek both very compact representations, such as parametric probability densities, and very

efficient processing, such as algorithms with linear time complexity. Such simplicity hinders performance, but validates the framework to pave the way for future work. This chapter is organized to derive our theory for linear time complexity processing, called the "ACME Segmenter", and then unveil each of the top three CDN layers.

6.1 The "ACME Segmenter"

6.1.1 The Complexity of Context

Chapter 5 discussed the combinatorial limitations of incorporating context into the labeling problem. That is, a data set of N voxels and M tissue classes has M^N possible distinct instantiations. A brute force maximum likelihood segmenter could compare an input image to each possible instance, and output the most likely instance given its compatibility with the input data. Such an approach is appealing because it classifies each voxel within the context of the entire image, but its exponential time complexity renders it impractical.

Therefore, we modeled the image as a Markov random field in order to consider each voxel within a context smaller than the entire image. Our results of Chapter 5 demonstrated that a very small neighborhood proved valuable in handling noise, but our experiments with Diagonalized NNPM in Chapter 3 revealed that a much larger neighborhood is required to recognize deviations from normalcy. The key difference between handling noise and recognizing abnormality is that the neighborhood size in handling noise was a constant, c, while the neighborhood size for recognition is some fraction, f, of the image size, N. This is clearly the case if the width of the white matter were to be a factor, for example. When every voxel considers information from just c other voxels, the time complexity is *linear* at $O(cN)$, but considering information from fN other voxels advances the time complexity to *polynomial* at $O(fN^2)$.

Therefore, we seek an algorithm that incorporates context as broad as a large-neighborhood Markov random field would, but with linear time complexity. We derive such an algorithm in the next subsection, named the "ACME Segmenter".

Table 6.1. The Complexity of Context

Time Complexity	O	Algorithms
Exponential	$O(M^N)$	Brute force Maximum Likelihood
Polynomial	$O(fN^2)$	MRF with large neighborhoods
Linear	$O(cN)$	MRF with small neighborhoods, and ACME

6.1.2 Derivation of the "ACME Segmenter"

In Chapter 1's review of related work, we noted that many approaches in the literature incorporate context using morphological operations [Jain95] and/or a series of *ad hoc* heuristics. To avoid bringing such a criticism upon ourselves, we develop a theory that provides a formalism for performing these types of computation by organizing the operations in a logical manner. We propose adopting the computational model employed by the scientific community, which we define based on the following observations:

1. Let the community perform parallel computation where each individual scientist represents a separate computational node. Proceeding in isolated parallelism, each computational node computes based on the information known uniquely to it, such as the sum of its experiences, interests, and education.
2. The computational nodes communicate with one another. However, it is too inefficient for each node to communicate everything it knows to every other node, or even one other node.
3. Neighboring nodes communicate more to each other than to distant nodes, producing a local coherence. (Researchers build on one another's work within the same group.)
4. While nodes keep some information entirely to themselves (rough notes and rough code), and share some with nearby neighbors (refined code and lab discussions), they submit an even smaller amount of information (conference submissions) to a global collection.
5. Each node does not have access to the globally contributed information (submitted papers) of every other node. Instead, some processing occurs at a global level (peer reviews) to reduce the amount of information (reject papers)

and compute some metrics (award papers, citations) on the global information to better facilitate its usage by nodes everywhere.

Figure 6.1 illustrates this computational model graphically.

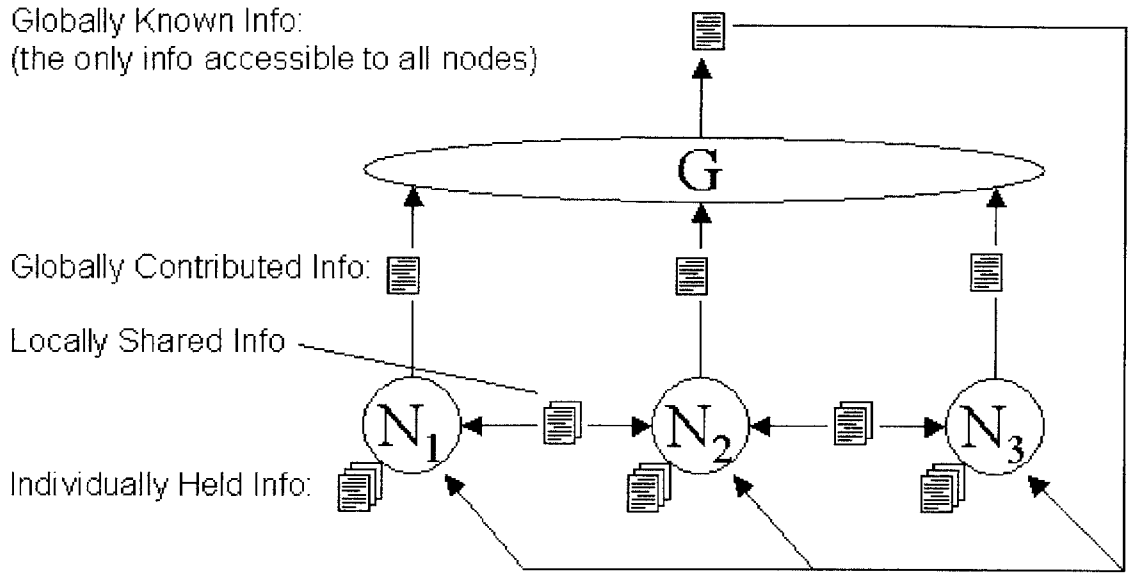

Figure 6.1. Academic Computation Model. Massively parallel computation exists where each node (N_1, N_2, N_3) shares a subset of its knowledge with its neighbors, and an even smaller subset with the global community. However, unlike the direct sharing with neighbors, some processing occurs (by G) on the collection of globally contributed information before its dissemination. This is far more efficient that requiring that each node process the global information independently.

Consider reducing CDN to the academic model. Let there exist a separate computational node for each image voxel. Proceeding in isolated parallelism, each computational node computes the classification vector $\boldsymbol{W}$ based on the information known uniquely to it, such as the input image intensity at that voxel, and the spatially varying prior at that voxel. This was CDN Layer #1. Better results can be achieved when the computational nodes communicate with one another, as in CDN Layer #2. Since is too inefficient for each node to communicate everything it knows to each other node, neighboring nodes communicate a subset of their knowledge: their computed classification vectors $\boldsymbol{W}$ (in the case of a mean-field MRF), but not their input image intensities or spatially varying priors.

Based on this reduction, we can conclude that broad context should be incorporated into CDN Layers #3 and #4 based on Figure 6.1. Since less information should be contributed to global processing than to neighborhood processing, we will allow each node to contribute its classification label, w, rather than its classification vector, $\boldsymbol{W}$. As a result, the total "global contribution" is essentially a segmented image. Global processing by some function, G, drastically reduces the information that each node incorporates into its subsequent processing. Just as no researcher reads all scientific papers, ACME achieves linear time complexity by not forcing each node to consider the entire segmented image. The exact nature of the global processing to be performed by G on the segmented image, and how each node will incorporate the global information, will be developed in the next few sections. The time complexity of the total algorithm will be linear as long as the selected G has linear complexity.

Based on the academic analogy, we could call this algorithm the "ACME Segmenter" as an acronym for the "Academic Computation Model for Efficiency". Alternatively, since this computational model is arguably employed by corporate enterprises, commercial markets, financial markets, the military, politics, and throughout organized society, perhaps the acronym stands for "A Computational Model for Everything". In this thesis, the true meaning of the acronym remains secret in accordance with the long tradition of ACME's usage[1].

6.1.3 Incorporating the Globally Processed Information

How should each node incorporate the output of G into its processing? The answer can be straightforward depending on the form of the nodal processing. For example, with an active contour, such as a snake or level set method, one can append an additive term to its energy function. With a Bayesian scheme, such as our implementation of CDN Layers

[1] According to the American Heritage Dictionary, acme has Greek origin meaning "the point of utmost attainment; peak". Apparently, early business school textbooks used Acme as a business name in some examples. Rumor has it that it was an acronym standing for "A Company Manufacturing Everything". Sears-Roebuck used Acme as one of their in-house brand names in the early 1900s, just like they use "Craftsman" today. Warner Brothers supposedly took the name from Sears and used it for the mail-order company in the Road Runner and Wile E. Coyote cartoons. However, its meaning was never disclosed.

#1-2, one can append a multiplicative term to the prior. Recall that our analysis of MRFs in Chapter 5 led to deriving equation 5.41, which we repeat in more general terms below:

$$\text{(posterior)} \propto \text{(likelihood)*(singleton prior)*(neighborhood prior)*(global prior)} \quad (6.1)$$

In essence, each layer provides the spatially varying prior for the layer below. Such an approach also circumvents the problems associated with atlas registration errors when spatially varying priors are required to resolve ambiguity. The expression is valid provided that the assumptions of conditional independence apply.

6.1.4 Comparing ACME with other Methods

We now compare other methods with the "ACME Segmenter".

6.1.4.1 High-Level Expert System

[Clark98] presented a technique for segmenting brain tumors that relied upon a combination of morphological operations and a high-level expert system. As will be detailed in the next section, the ACME global processing (G in Figure 6.1) is not a rule-based expert system that could be construed from a series of *ad hoc* heuristics. Rather, G performs analytical computation, such as probabilistic treatment of shape descriptors.

More importantly, the ACME computation retains a "soft" nature by feeding the global output back into the local processing nodes. This is a stark contrast to a "hard" expert system that makes decisions that irreversibly discard information.

6.1.4.2 MRF with Larger Neighborhood

Consider endowing CDN Layer #2's MRF with a neighborhood sized sufficiently large to encompass the largest structure radius. Motivations against this include the complexity argument presented in Section 6.1.1, and also the ease of measuring abnormality. Using MRF neighborhoods alone, the concept of normalcy would be captured more implicitly rather than explicitly. We would lose the benefit of being able to easily answer our two guiding questions from Chapter 2. Explicit high-level properties can be fit with probability distributions that allow us to define what is normal, and quantitatively measure abnormality.

6.1.4.3 Multiscale MRF

[Leuttgen93] developed multiscale representations of Markov random fields. These are substantially more computationally efficient than well-known MRF models. As described in Chapter 3, the main shortcoming that we find with multiscale techniques is that we much prefer to carve up image space along structural boundaries rather than along arbitrary divisions of the lattice. ACME allows us to compute regional properties of structures with full resolution.

6.1.4.4 ATM-SVC

The ATM-SVC algorithm applied to brain tumors by [Kaus00] was briefly described in Chapter 1. It is similar to ACME in the sense that some global processing is performed after classification, and prior to another iteration of the complete algorithm. This is accomplished by warping a binary brain template to the binary output classification. However, it does not model the independent processing nodes as ACME does. As a KNN-based classifier, it ignores the bias field and assumes binomial distributions rather than Gaussian distributions in a Bayesian framework. Neighborhood intensity interactions are not considered.

6.1.4.5 EM-MF

Chapter 1 described several EM- and MRF-based methods for segmentation of brains without tumors ([Kapur99], [Leemput01b]) and with tumors ([Moon02]). Beyond the local MRF neighborhood, these methods incorporate context only through their use of a geometric prior. In the case of [Leemput01b] and [Moon02], the prior is the same one employed by our CDN Layer #1. In the case of [Kapur99], the prior was computed relative to the scalp and ventricles, but these were segmented *a priori* before any EM and MRF processing began. In summary, all of these methods do not acquire a concept of context from analyzing the image data itself. That is the central benefit of our CDN Layers #1-2 feeding into Layers #3-4.

6.1.5 Designing the Global Processing

Now that we have described the ACME model of Figure 6.1, we are positioned to develop Layers #3 and #4 of CDN in the next sections. Note that while every ACME

segmenter is a CDN, not every CDN is an ACME segmenter. For example, a CDN where nodes communicate all of their knowledge to a neighbor, or to *G*, would not be ACME-compliant. As another example, a CDN that forced each node to perform its own global processing (often resulting in polynomial or even exponential time complexity) would not be an ACME segmenter.

The purpose of the ACME model is to assist with our development of CDN Layers #3-4 so that their design is less arbitrary, and not as open to *ad hoc* heuristics and inefficient schemes. We therefore seek to constrain the search space over allowable functions for *G*. Toward this end, the CDN framework provides some assistance by forcing the designer to separately consider layers #3 and #4 – dividing the computation into what can be computed about a structure, and what can be computed about its relationship to other structures. As an additional constraint for this thesis, we will explore only implementations of *G* that result in solutions with linear time complexity. This restriction will hinder the quality of our results, but we will strive for the best results given efficient computation.

6.2 CDN Layer 3: Intra-Structure Classification

Consider the example of non-enhancing tumor tissue that mimics the intensity of healthy gray matter, but is too thick to be gray matter. CDN Layers #1-2 would first classify the tissue as gray matter, but Layer #3 – through its broader understanding of context – could correct the misclassifications of the first two layers. In this example, tissue thickness is regarded as a *region-level property*. It is a metric computed over all voxels that share a certain tissue type. For such a metric to be computed, classification must first be performed by the CDN's first two layers. That is, Layer #3 is predicated on Layer #2.

6.2.1 Computation of Region-level Properties

Shape is a region-level property, as intensity is a voxel-level property, and homogeneity is a neighborhood-level property. By *region-level property*, we refer to any information describing the nature of aggregate collections of similar voxels. Since our goal in this thesis is to clearly present a framework, and demonstrate it with a simple, easily-conceptualized implementation, we are manually selecting simple shape descriptors as our region-level properties. Future implementations of the framework can follow Chapter

2 in employing mathematical methods such as PCA to automatically discover underlying structure from training data.

Simple shape descriptors can take the form of coefficients for combining a series of basis functions, or as measurements of curvature, or as distances. Applying distances in the form of thickness requires definition of two separate surfaces between which to compute object thickness [Yezzi01]. Since such definitions are not clear for all brain structures, we first experiment using shortest distance-to-boundary of a given voxel's structure. This metric is quite different from thickness, but readily computable. Note, for example, that all voxels within a sphere would have identical thickness properties, but the distance-to-boundary property varies radially. Therefore, we propose using an approximate thickness metric of *maximum* distance-to-boundary. This simplification applies when we assume spherical topologies for brain structures of interest. We select it over alternatives given its speed (run times listed in Table 6.2), compatibility with our framework (it suggests a form of G that introduces broad context for piecewise homogenous scenes), straightforward implementation, and its empirical impact on results (nicely complements intensity and neighborhood coherence).

The distance-to-boundary metric is computed by performing a distance transform on the segmented structures. There are generally two approaches available for performing distance transforms: approximate and absolute. The Chamfer distance [Borgefors86] presents an approximate algorithm that is favorable given that its run time is fast and consistent independent of the image topology. Chamfer distance is computed by convolving the segmented image with a triangular mask in the forward direction, and another in the backward direction. On the other hand, the fastest current algorithm for computing absolute Euclidean distance (square root of the sum of squares) for data of the extent typically encountered in medical imaging is [Saito94]. Table 6.2 lists the empirical results comparing run time for the two techniques on our specific 3-D domain, and Figure 6.2 displays the results in pictorial form.

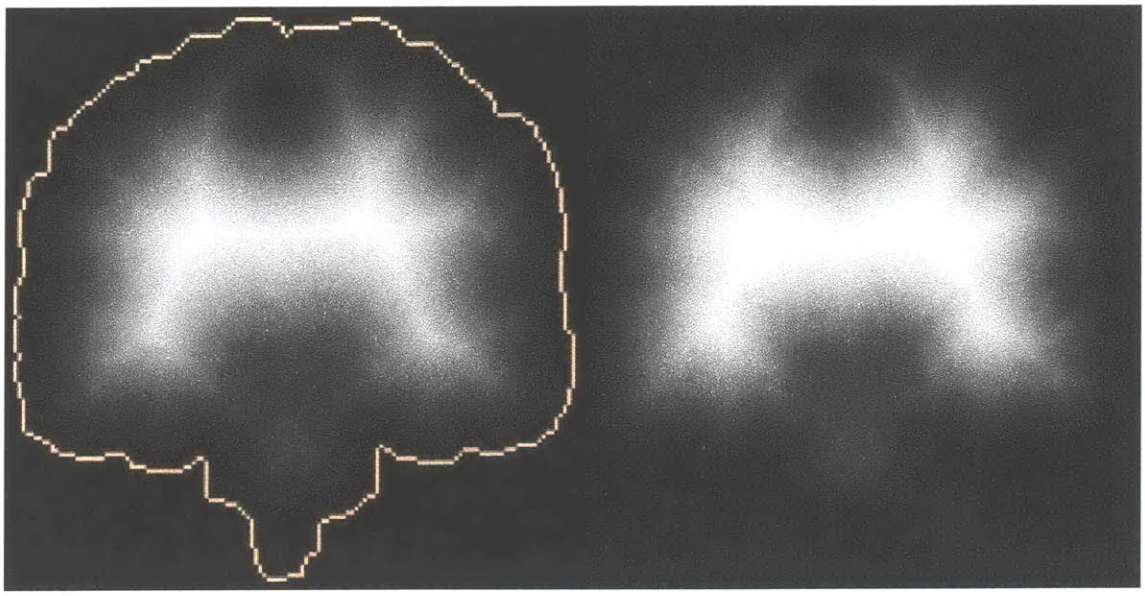

Figure 6.2. Distance Transforms (Left:) Euclidean distance to the ICC border outlined in tan. (Right) Chamfer distance with slight artifacts of radial streaking.

Table 6.2. Distance Transforms: run time (seconds) on real 3-D brain atlas. Observe that Euclidean run time increases with increasing object sizes, while Chamfer run times are independent of object topology.

Image	Chamfer	Euclidean (Saito)
CSF	*3.5*	*2.0*
White Matter	*3.5*	*2.1*
Gray Matter	*3.5*	*2.2*
ICC	*3.5*	*2.6*

The per-class probability distributions for the distance to each structure's own boundary, $p(r|w)$ are readily computed from a sample segmented scan presented as training data, and results are given in Table 6.3 and Table 6.4.

Table 6.3. Distance-to-Boundary measurements from a real brain atlas.

Image	Mean	Standard Deviation
CSF	2.67	2.73
White matter	5.36	3.14
Gray matter	1.89	1.06
Vessel	1.03	0.11

The distribution for maximum distance-to-boundary cannot be computed from a single scan, but only from a large population. Without a suitable real data set available, we trained using synthetic data and empirical measurements of real data. Table 6.4 presents the values used during the experiments throughout the remainder of this thesis.

Table 6.4. Maximum Distance-to-Boundary

Image	Synthetic Brains		Real Brains	
Image	**Mean**	**Standard Dev.**	**Mean**	**Standard Dev.**
Scalp	4	2	-	-
White matter	19	4	12	4
Gray matter	9	4	4	2
CSF	12	4	7	2
Vessel	1.4	1	4	2

6.2.2 A Probablistic, Topological Atlas in Addition to a Geometic Atlas

The atlas used in Layer 1 (described in Section 3.5) can be regarded as a geometric atlas because it encodes the geometric relationships between brain structures. In contrast, Layer 3 can, in general terms, be thought of as incorporating a probabilistic, topological atlas. Such an atlas can be constructed by fitting probability distributions to spatially varying shape descriptors.

Continuing with our simple example of using distances to structure boundaries, consider using different distributions for cortical gray matter than sub-cortical gray matter. The sheet-like nature of cortex would be represented by its very small distances, while the more spherical topology of sub-cortical structures would be encoded with a much broader distribution. Some geometric component is still required to map image space to atlas space. In the current example, atlas space would consist of two distinct distributions for gray matter, and the mapping from image space to atlas space would appear the same as depicted in Figure 5.5 for mapping to an atlas space of differing tissue class interaction matrices, **J**.

6.2.3 An Implemention of "G" Based on the Metric of Maximum Distance-to-Boundary

The role of G of an ACME segmenter is to instill each processing node with an understanding of its broad context while sparing it from performing computations of high complexity. We design a G_3 and G_4 to govern the 3rd and 4th CDN layers respectively. Our specific implementation of G_3 for use in the following experiments performs the steps listed below. Recall that the information contributed to G_3 from each node is its MAP classification (tissue label) computed from the results of CDN Layer #2.

1. Run connected component analysis to produce a voxel-wise labeling of the islands of each tissue type (Fig 6.3: 2nd column, 2nd row). We used a 3-D neighborhood size of 6 for efficiency.
2. Compute the distance transform on each island to produce a map of distance-to-boundary (Fig 6.3: 2nd column, 1st row). We used the Euclidean distance algorithm of [Saito94].
3. For each island, find its maximum distance-to-boundary (Fig 6.3: 2nd column, 2nd row). Compute the probability of abnormality of this distance according to equation 4.12. Assign this probability to the value of every voxel in the island (Fig 6.3: 1st column, 3rd row)

Figure 6.3 illustrates the technique by displaying the intermediate results from each step of G_3's processing. The processing applies to all tissue classes, although only gray matter results are shown for brevity. The synthetic case features a tumor with an intensity distribution identical to that of healthy gray matter, and therefore indistinguishable by intensity, but an outlier with respect to shape.

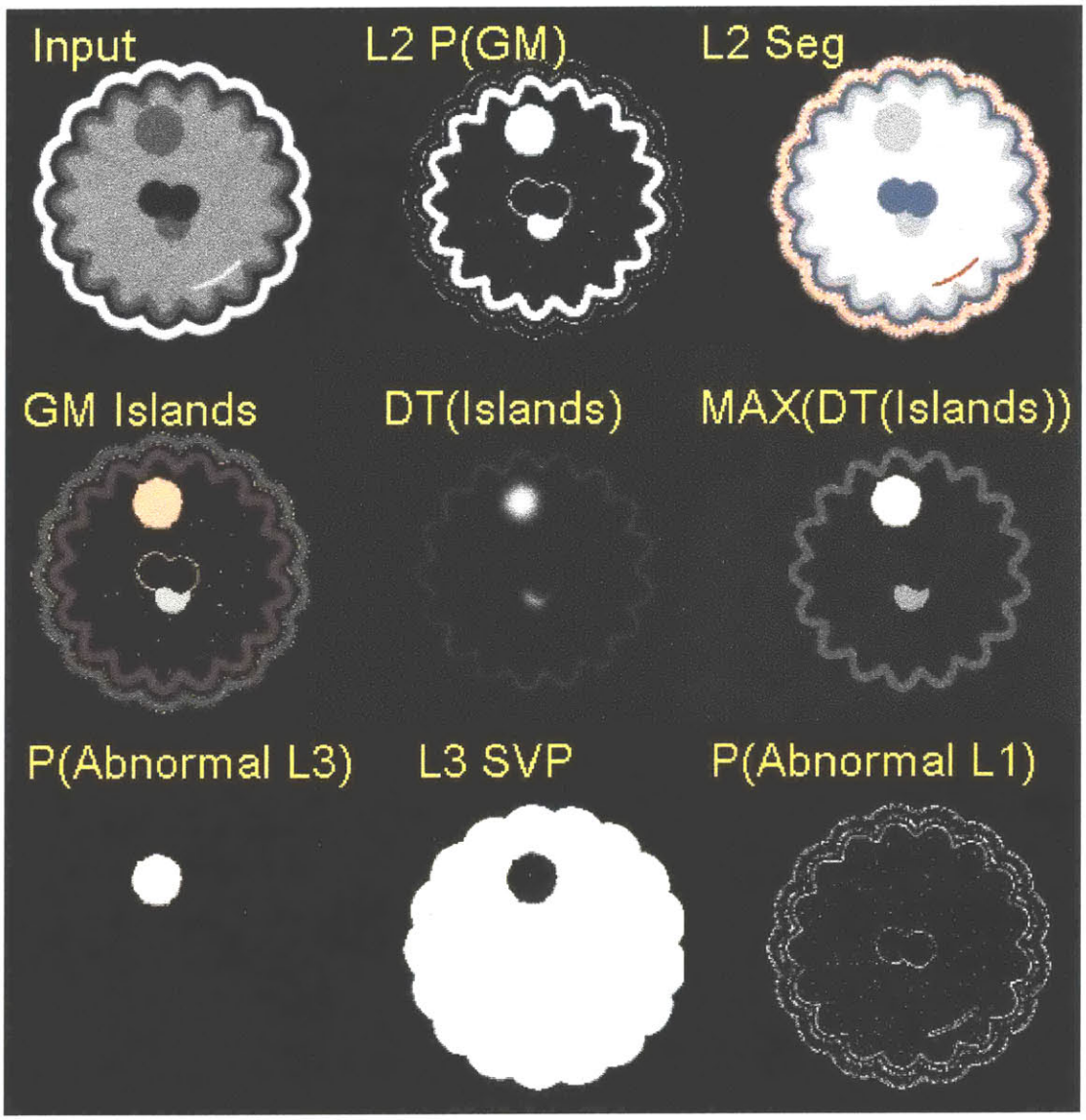

Figure 6.3. ACME *G* for Layer #3. Top row, left to right, are the input image, the posterior probability of gray matter computed by layers #1-2, and the segmentation following layer #2. The tumor is incorrectly classified as gray matter at this point, so layer #3 will create a spatially varying prior (SVP) to be applied in the next iteration of layers #1-2. The next rows show the intermediate steps of G: identification of gray matter islands (uniquely colored), distance transform of the islands, maximum distance found within each island, probability of this distance occuring in gray matter, and the spatially-varying prior computed from complement of this probability. Neighborhood interactions smooth out the effects of sharp priors. The last image is the probability of abnormality computed by layer #1, illustrating that tumor intensity appeared normal.

6.2.4 Incorporating the Output of G_3

Following Figure 6.1 and equation 6.1, the output of G_3 is incorporated into the next iteration of CDN layer #1 by creating a spatially varying prior for typicality. Since

typicality is the complement of abnormality, this prior appears as in the middle, bottom of Figure 6.3, and is computed simply as:

$$\forall s \in S: \quad SVP_{Shape}(s) = 1 - P_{Abnormal\ L3}(s) \tag{6.2}$$

Figure 6.4 illustrates the impact that SVP_{Shape} has in conjunction with spatially varying prior on intensiy, $SVP_{Intensity}$.

$$\forall s \in S: \quad SVP\ (s) = SVP_{Shape}(s)^* SVP_{Intensity}(s) \tag{6.3}$$

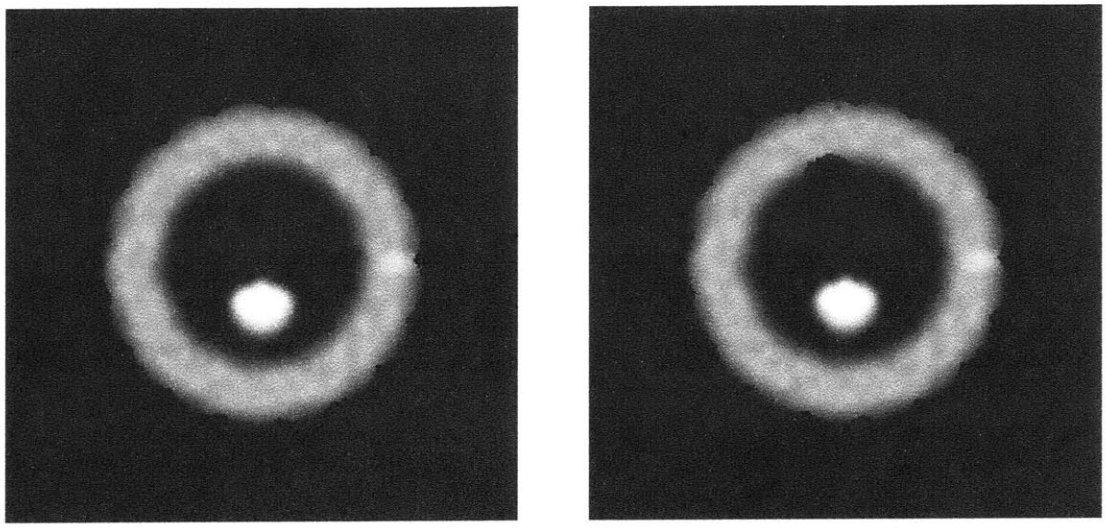

Figure 6.4. Patient-Specific SVP for Gray Matter. (Left:) Intensity SVP that would be used in layer #1 in the absense of layer #3. (Right:) New SVP used in the second iteration of layer #1 formed by combining information regarding both intensity and shape. Observe the dampening impact of the shape prior near the image top.

The predication of layer #3 on layers #1-2 demands that the spatially varying shape prior not take effect until the second "outer iteration" (execution of all layers). We have not addressed, however, during which iterations the spatially varying intensity prior is valid. While the referenced related works depend on a spatially varying intensity prior from the outset, it would not be appropriate for us to do so within our framework. The intensity prior imposes localization information for healthy classes into the scene recognition process. Since pathology is not represented within the prior, the prior adversely affects its recognition. Figures 6.5-6 express this concept pictorially, and a mathematical derivation follows later.

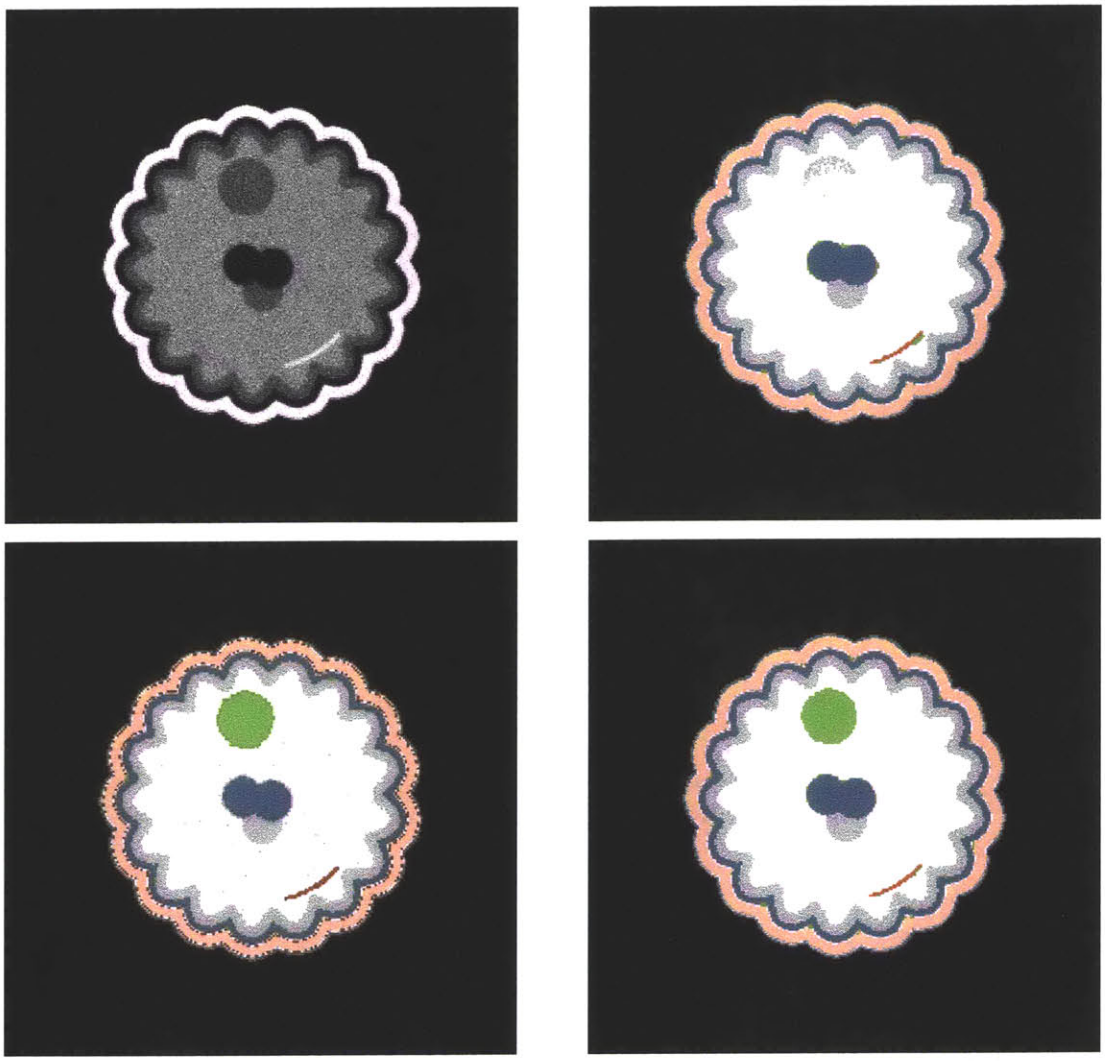

Figure 6.5. Invalid Application of an Intensity SVP. (TOP:) CDN segments the input image on the left to produce the segmentation on the right. The segmentation divides the tumor into two parts of gray matter and white matter -- according to their respective positions within the intensity SVP (refer back to Figure 6.7 for the gray matter SVP). The only pathology (green) in the final segmentation is partial volume artifacts surrounding the ventricles (blue) and vessel (red).

(BOTTOM:) The segmentation is performed using only stationary intensity priors on the left, which results in spurious gray matter speckle within white matter, and vessel speckle within skin. The segmentation on the right, however, is correct except for PVA which will be corrected by CDN layer #4. The success of this segmentation is based on its use of a stationary intensity prior during the first outer iteration, and a spatially varying prior during the second outer iteration. In essense, the algorithm "peeks" at the image before imposing its preconceptions.

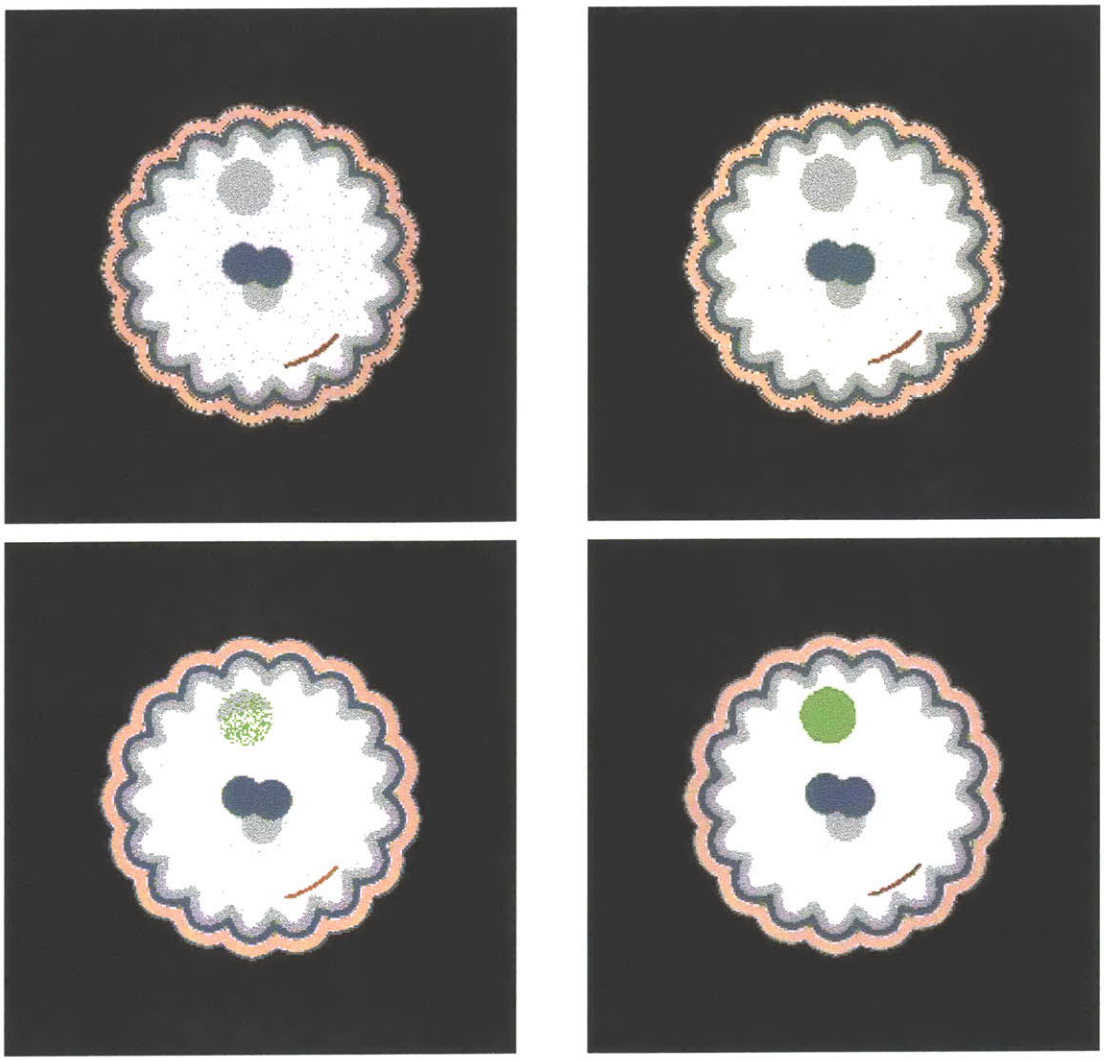

Figure 6.6. Sequential Intensity Prior. From top to bottom, left to right, are snapshots of intermediate results during the process of segmenting with a sequential intensity prior. The top row depicts the results of using a stationary prior during the first outer iteration, and the bottom row continues with the second outer iteration using a spatially varying prior. The left images are taken at the conclusion of layer #1, and those on the right reveal the results after layer #2. Observe that the voxels with the most abnormal intensities were identified as abnormal in layer #1, and because of their close relationship to other voxels in the tumor with respect to both intensity (through the input image intensities) *and* shape (through the shape SVP), the MRF propagated the pathology classification throughout the entire structure very rapidly.

The above figure explored the case of pathology with an intensity profile identical to that of healthy tissue, but straddling the expected location of two different tissues. The next examples explore additional permutations.

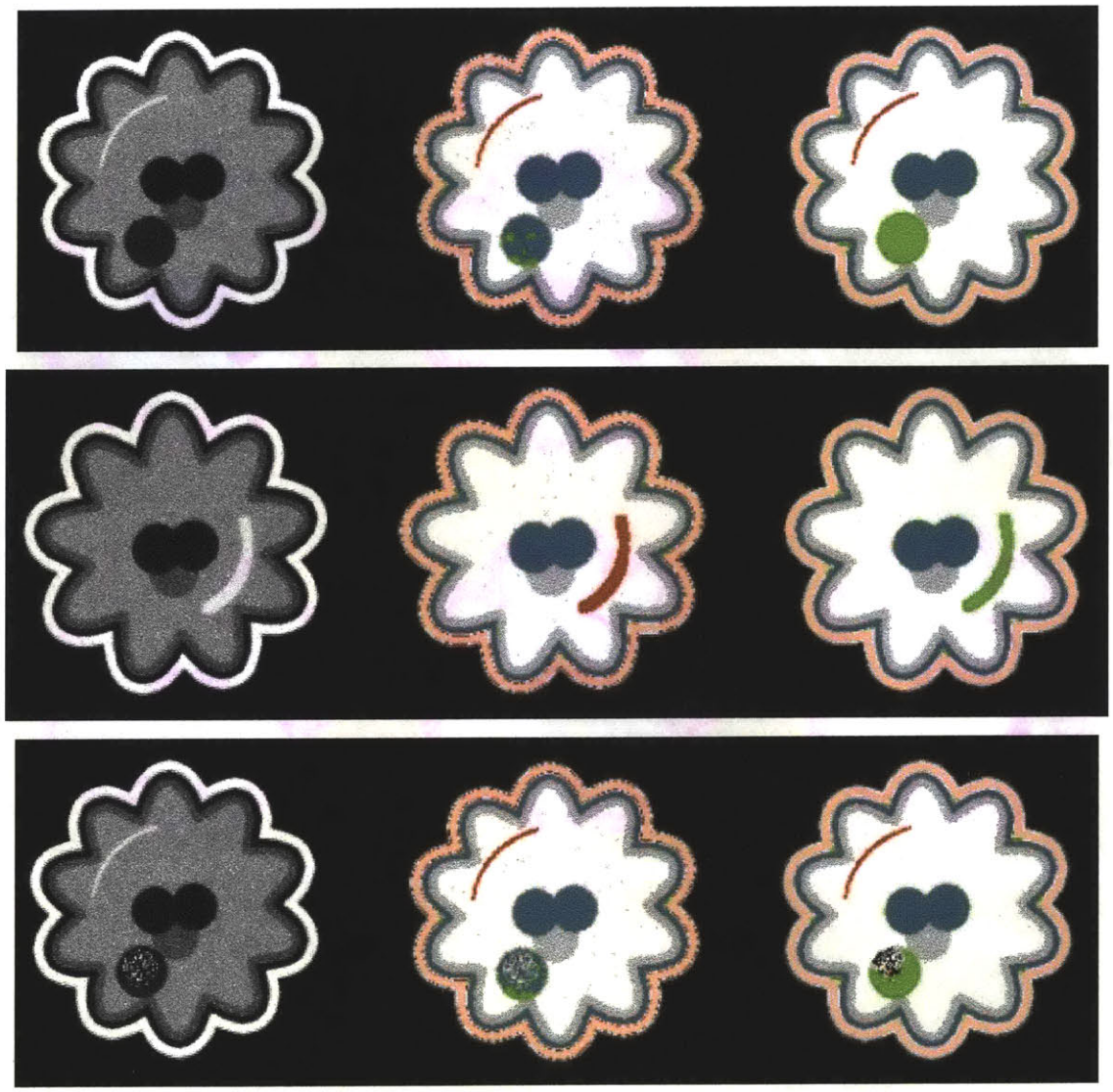

Figure 6.7. More Examples of Impact of Layer #3. From left to right are the input images, segmentation following the first outer iteration, and segmentation following the second outer iteration. Colors represent tumor (green), vessel (red), CSF (blue), white matter (white), and gray matter (gray), and skin (tan).

(TOP:) Pathology has mean intensity between that of gray matter and CSF, so the first iteration identifies some, but not all, of it as abnormal. Most of the tumor was originally labeled as CSF (blue) until corrected to green.

(MIDDLE:) Pathology has identical intensity profile as a healthy vessel, and does not extend outside of a vessel's expected location. This is a revisitation of the lighthouse anamoly case that was incorrectly segmented in Figure 4.10. Both intensity and location are insufficient to resolve ambiguity, requiring information regarding size/shape.

(BOTTOM:) The algorithm fails when a tumor's extreme heterogenaity results in a mixture of apparently normal intensities of sizes too small to measure shape. Texture recognition would be helpful in this case, as a future extension of the current framework.

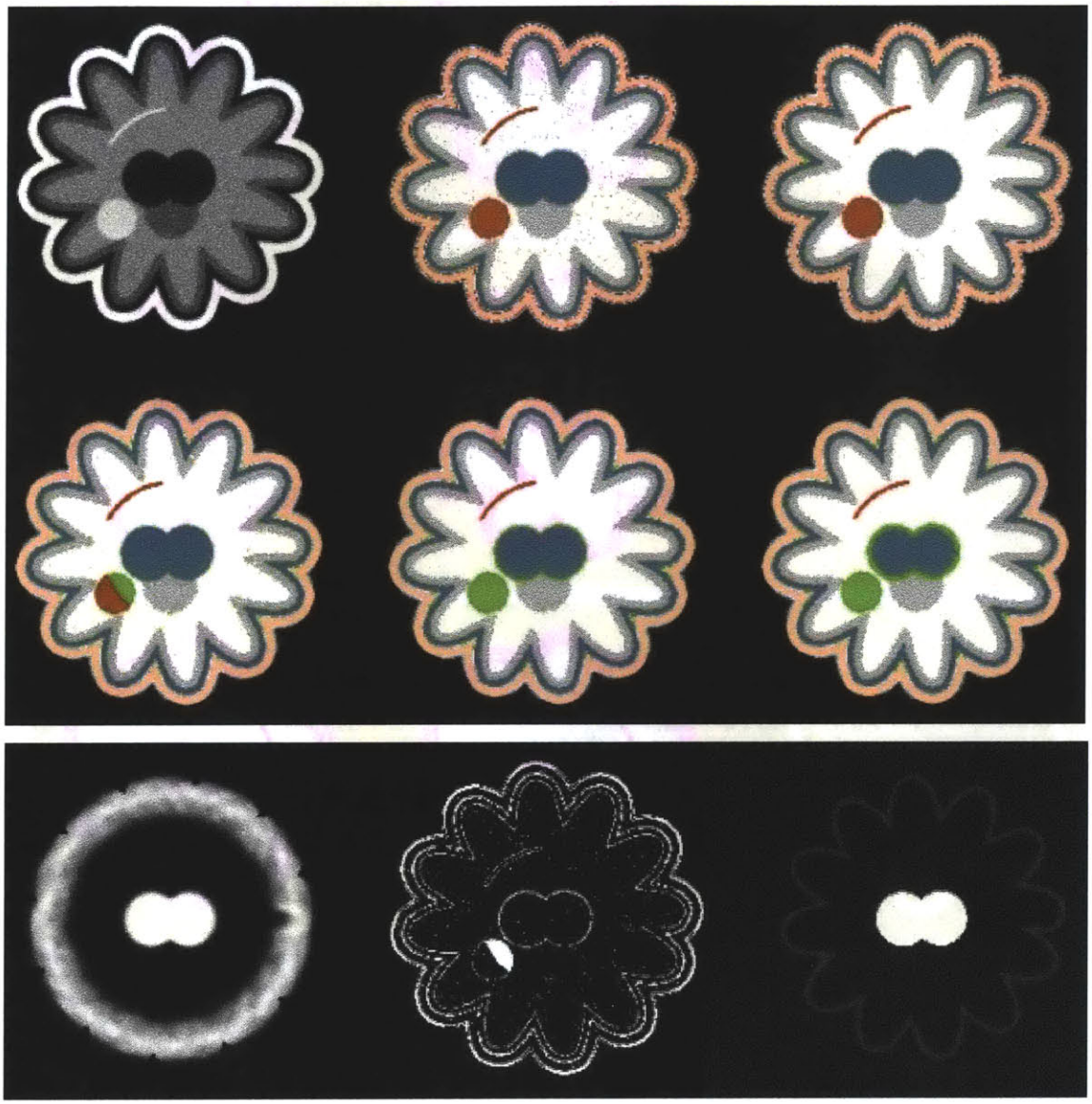

Figure 6.8. Ability to Communicate with User. The top set of images depict a sequence of intermediate results on a case with a bright tumor in addition to unusually large ventricles (blue). Observe that only the unusual portions of the ventricles were segmented as abnormal (green). To determine the reason behind this behavior, the user need only examine the various intermediate probability maps. Inspection of these in the bottom picture (intensity SVP, intensity abnormality, and shape abnormality) reveal that the unusually strong SVP overruled the spread of tumor classification into the ventricle interior. CDN's organization as a layered, Bayesian network enables the computer to respond with answers to a user's curiosity.

6.2.5 CDN without ACME

Prior to the development of ACME, we implemented a CDN where the computational nodes communicated as much knowledge to G_3 as to their neighbors [Gering02b]. For large structures – on the order of a percentage of the image size – the method has polynomial time complexity, but we include it below for completeness.

Once region-level properties are computed, the question arises of how the computation should blend the high-level information with the low-level information. The high-level information is a voxel-by-voxel representation of some region-level properties, such as distance to structure boundary. The low-level information is the classification based on intensities of individual voxels and their neighbors. Given the following premises, we conclude with an approach that satisfied our goals, and we call it the Multi-level MRF.

Premise 1	Voxels of similar low-level classifications possess various values of the high-level metrics.
Premise 2	Probability distributions can be associated with the high-level metrics.
Premise 3	High-level distributions can be used to compute high-level classifications.
Premise 4	Voxels of similar low-level classifications tend to possess similar high-level classifications.
Conclusion	High-level classifications should propagate to neighboring voxels with similar low-level classifications.

Based on this conclusion, computation in the above example would proceed as follows. Voxels toward the center of the mass would be first classified as tumor based on their unusually high distance from their structure's boundary. This tumor classification would subsequently flow outward throughout the mass over several iterations in a *probabilistic flow*. The flow is driven by our introduction of multi-layer Markov random fields, developed in the next subsection. In this way, a given voxel would change its high-level classification in the evolving presence of tumor if the attributes of lower-level layers

shared strong similarities. We now derive the multi-layer Markov random field proposed as a mechanism for propagating region-level properties.

6.2.5.1 Review of First Two Layers of CDN

We begin with a brief review of the two lowest layers to set the stage by providing a point of reference for the mathematics empowering the third layer. Recall that EM segmentation models the image intensities as visible variables, y, tissue classifications as hidden variables, w, and the bias field as governed by model parameters, b. We would like to choose the parameters that maximize the log likelihood of the data, $\log p(y, w \mid b)$, but we do not know this likelihood because w's invisibility renders $p(y, w \mid b)$ to be a random variable. Thus, although we cannot maximize it, we can *maximize* its *expectation*. This results in the following two iterative steps until convergence to a local minimum:

E-Step: Compute the expectation $\sum_{w} p(w \mid y, b) \log p(w, y \mid b)$ using the current b.

M-Step: Find new b^{t+1} to maximize the expectation, assuming $p(w \mid y, b^t)$ is correct.

We then added Layer 2 to effectively relax Layer 1's E-Step weights. The prior knowledge of spatial coherence over a configuration, w, of segmented voxels is modeled with a Gibbs distribution, $P(w)$, which takes the following form, from equation 5.7:

$$P(w) = \frac{\exp(-U(w))}{\sum_{w' \in W} \exp(-U(w'))} \tag{6.4}$$

This distribution's energy function, $U(w)$ is an Ising model generalized to the case of discrete, multi-valued labels, and we repeat it here from equation 5.10:

$$U(w) = \sum_{i \in S_1} V_1(w_i) + \sum_{i \in S} \sum_{j \in N_i} V_2(w_i, w_j) \tag{6.5}$$

This energy function is composed of clique potentials, where V_1 is the clique potential of all cliques of size 1. In other words, V_1 encodes our prior knowledge about an isolated voxel prior to viewing the image data. This prior knowledge is the tissue class prior probability, which may be either stationary, or spatially-varying. V_2 is the potential over

all cliques of size 2, and represents the tendency of two classified voxels to be neighbors. That tendency is encoded in the MxM Class Interaction Matrix, **J**, and it is computed from a segmented scan offered as training data.

To make the computation tractable, we used the mean field approximation to factorize the joint probability into a product of local conditional probabilities:

$$P(w) \approx \prod_{i \in S} P(w_i \mid \overline{w}_{N_i}) \tag{6.6}$$

Then, computation is straightforward using the local clique potentials, which we repeat here for convenience from equation 5.48. Given M possible label values, let $\mathbf{w}_i$ be an M-length binary vector of classification at the voxel indexed by *i*. Then:

$$\begin{aligned} V_1(\mathbf{w}_i) &= -\ln P(\mathbf{w}_i) \\ V_2(\mathbf{w}_i, \overline{\mathbf{w}}_j) &= -(\mathbf{w}_i^T \mathbf{J} \overline{\mathbf{w}}_j) \end{aligned} \tag{6.7}$$

6.2.5.2 Multi-Level MRF

We perform a Maximum A Posteriori (MAP) classification of the features (just radius at present) computed over Layer 2's output. Recall that the EM algorithm of Layer 1 must compute *p(w|y,b)* at each E-Step. Since the distributions over region-level properties are independent of the distributions over voxel-level properties (shape is not related to intensity or bias), *p(w|y,b,r)* can be computed with the same update equation except for an extra multiplicative term, *p(r|w)*:

$$p(w \mid y,b,r) \propto p(w \mid y,b) p(w \mid r) \tag{6.8}$$

Therefore, the posterior probabilities for the Layer 3 MAP classification are equal to the relaxed weights of Layer 2 multiplied by this new likelihood. That is, the Layer 2 weights provide the spatially varying prior for the Layer 3 MAP classification. Using superscripts to denote CDN layers, we repeat the MRF equations for Layer 2 below. There is a bar over **w** to denote that it is a vector of probabilities for the Mean Field approximation.

$$U_i^2\left(\mathbf{w}_i^2 \mid \overline{\mathbf{w}}_{N_i}^2\right) = V_1^2\left(\mathbf{w}_i^2\right) + \sum_{j \in N_i} V_2^2\left(\mathbf{w}_i^2, \overline{\mathbf{w}}_j^2\right) \tag{6.9}$$

$$V_1^2\left(\mathbf{w}_i^2\right) = -P\left(\mathbf{w}_i^2\right) \tag{6.10}$$
$$V_2^2\left(\mathbf{w}_i^2, \overline{\mathbf{w}}_j^2,\right) = -\left(\mathbf{w}_i^2\right)^T \mathbf{J}^2 \overline{\mathbf{w}}_j^2$$

Next, we desire the MAP result (corrections to Layer 2's classifications) to propagate over regions that are homogenous at Layer 2, as demonstrated in Figure 6.6. We introduce a multi-level Markov random field, and define the Gibb's energy function to encode our prior knowledge of its behavior. Compare the equations below with their Layer 2 counterparts above:

$$U_i^3\left(\mathbf{w}_i^2, \mathbf{w}_i^3 \mid \overline{\mathbf{w}}_{N_i}^2, \overline{\mathbf{w}}_{N_i}^3\right) = V_1^3\left(\mathbf{w}_i^2, \mathbf{w}_i^3\right) + \sum_{j \in N_i} V_2^3\left(\mathbf{w}_i^2, \mathbf{w}_i^3, \overline{\mathbf{w}}_j^2, \overline{\mathbf{w}}_j^3\right) \tag{6.11}$$

$$V_1^3\left(\mathbf{w}_i^2, \mathbf{w}_i^3\right) = -\left(\mathbf{w}_i^3\right)^T \overline{\mathbf{w}}_i^2 \tag{6.12}$$
$$V_2^3\left(\mathbf{w}_i^2, \mathbf{w}_i^3, \overline{\mathbf{w}}_j^2, \overline{\mathbf{w}}_j^3\right) = -\left(\mathbf{w}_i^3\right)^T \mathbf{J}^3 \overline{\mathbf{w}}_j^2$$

The *MxM* square *Similarity Matrix* (SM), $\mathbf{J}^3$ is the Layer 3 counterpart of Layer 2's *Class Interaction Matrix* (CIM), $\mathbf{J}^2$. The SM is chosen to drive voxels classified to structures with large radii to propagate over voxels associated with structures with small radii.

6.2.5.3 Results

Figure 6.7 shows results of experimenting on a toy data set where low-level classification failed to handle ambiguity, and the multi-layer MRF corrected the result.

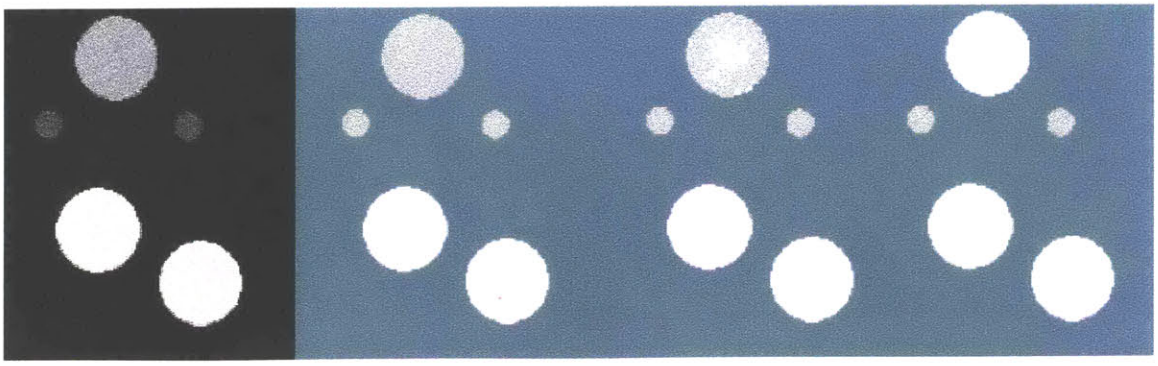

Figure 6.9. Results. The "toy" volume consists of 2 small, dark spheres and 2 large bright ones corrupted with Gaussian noise. The top, somewhat dark, and large sphere is ambiguous, and it is classified incorrectly by the lower-level layers of MAP and MRF. The 3rd layer then identifies that the center voxels are too distant from the boundary, and corrects their classification. The multi-layer MRF propagates this information across the structure because its lower-level segmentation is mostly homogenous. From left to right: Original, Result after Layer 2, Result after 15 iterations of Layer 3's multi-level MRF, Result after 50 iterations of Layer 3's multi-level MRF

6.3 CDN Layer 4: Intra-Structure Classification

While Layer #3 considered the context of a structure by itself, Layer #4 considers the context of multiple structures. Such consideration can yield two very different pieces of information. The first is whether a voxel is misclassified because it contains intensity information from not one, but multiple structures. The second is whether an entire structure is misclassified in a way that can be corrected based on its situation relative to other structures. We being with a discussion of the former, called partial volume artifacts.

6.3.1 Correcting Misclassified Voxels

Partial Volume Artifacts (PVA) arise when voxels that contain tissue belonging to more than one tissue class display an intensity value along the linear combination of the classes' distributions. While partial volume artifacts always present somewhat of an obstacle to segmentation, their effect becomes much more pronounced in our algorithm because the entire interface between structures incorrectly appears abnormal.

Figure 6.10. Autumn Artifacts. When viewed from across New Hampshire's Swift River, neighboring red and yellow leaves can easily be mistaken for orange.

6.3.1.1 Related Work

[Choi91] coined the term *mixels* to represent voxels that contain mixtures of multiple tissues. The quantity f_{il} is the fraction of the volume at the location of voxel i that consists of constituent (tissue class) l. Given M tissue classes, a mixel is an M-dimensional random variable $f_i = \{f_{i1}, f_{i2}, \cdots, f_{iM}\}$ that satisfies, at each voxel i:

$$\sum_{l \in L} f_{il} = 1 \tag{6.13}$$

Modeling each voxel as a mixel has straightforward implications for statistic analysis. For instance, the conditional probability for the observation x_i at i, given the true mixel f_i, is a Gaussian with mean $u_i = \sum_{l \in L} f_{il} \mu_l$

Since mixel constitution can be confused with image noise, [Choi91] used an MRF as a regularizer to convey that adjacent mixels are likely to have similar constituents. [Pham00a] and [Leemput01b] extended [Choi91] to favor pure mixels (homogenous voxels bordered by partial-volume mixels) either by using heuristics or by applying the MRF on the subvoxel level instead of the voxel level. [Santago95], [Jaggi98], and [Laidlaw98] took a different approach of using Bayesian classification to

match histograms by finding the mixture of materials most likely to have created the histograms. Since these methods have the drawback of discarding spatial information, [Ruan00] combined the histogram approach with MRFs. However, the search space was constrained by limiting the image model to contain only two mixture classes (CSF/GM and GM/WM) in addition to three pure classes. We note that such an approach would misclassify the voxels on the WM/CSF border of the lateral ventricles as GM. In fact, [Wang01], whose objective was to measure lateral ventricle volume, corrected for this using a scheme that performed morphological operations to identify candidate voxels for potential partial volume artifacts.

6.3.1.2 Our Approach

Given our novel approach of recognizing deviations from normalcy, we chart a different course for handling partial volume artifacts. With the exception of [Wang01], all references in the previous section pertained to unsupervised classification methods, where the statistical model parameters are determined automatically. In these approaches, the motivation for handling partial volume artifacts is to prevent the artifacts from widening the histograms of the true classes, thereby hindering the parameter estimation. Our motivation, on the other hand, is more like that of [Wang], where the artifacts are causing serious, erroneous classifications.

Our approach requires a means of resolving the ambiguity veiling whether a voxel's intensity is being influenced by pathology or partial volume artifacts. Since MRI in-plane resolution (~0.9 mm) is smaller than the size of brain structures (cortical thickness is 3-6 mm), adjacent voxels are likely to have similar constituents. Thus, we can resolve this ambiguity by referring back to our imaging model in Chapter 2 to derive a spatially varying prior on the presence of artifacts similar to [Wang]. Observe that we have been opposed to the use of morphological operations throughout this thesis, partly because of their dependence on the image lattice size. However, an appropriate application for morphological operations is when the lattice size is the very issue. Since PVA is caused by the finite lattice size, lattice-size based operations of erosion/dilation are suitable for screening candidate PVA mixels.

This screening can be performed efficiently on a global level, so we desire a G_4 function for the ACME paradigm. Since every voxel on the boundaries between distinct structures is at risk for PVA, the role of G_4 is to identify all voxels bordering structures, except those bordering substantial tumors. Define substantial tumors to be those of width greater than a single voxel. (The logic is that a PVA voxel labeled tumor by Layer #1 will not have sufficiently strong neighborhood coherence to be expanded by Layer #2.) Similar to G_3, input to G_4 is the labeling from CDN Layer #2, and the output is communicated back to CDN Layer #1 in the form of a spatially varying prior. While G_3 contributed a prior with respect to healthy tissues, G_4 offers a prior for pathology that discourages PVA candidates from consideration as abnormal. Formally, G_4 is defined as an SVP that is everywhere 1 except for 0's at voxel set PC. Given M labels and a lattice of m nodes:

$$
\begin{aligned}
&\text{Lattice:} && S = \{i \mid i \in 1..m\} \\
&\text{Healthy Tissue Labels:} && H = \{l \mid l \in \{1..M\}\} \\
&\text{Label for Tumor:} && T = \{M+1\} \\
&\text{All Labels:} && L = \{H \cup T\} \\
&\text{Neighborhood of Site i:} && N_i = \{j \mid j \in S, j \neq i, i = neighbor(j)\} \\
&\text{Neighborhood System:} && N = \{N_i \mid i \in S\} \\
&\text{Healthy Boundaries:} && HB = \{i \mid i \in S, j \in N_i, w_i \in H, w_j \in H, w_i \neq w_j\} \\
&\text{Substantial Tumor:} && ST = \{i \mid i \in S, w_i \in T\} \text{ after eroding, dilating tumor} \\
&\text{PVA Candidates:} && PC = \{i \mid i \in HB \cap \neg ST\}
\end{aligned}
\tag{6.14}
$$

We used a 3-D neighborhood system of 26 neighbors in order to involve all immediate neighbors of a given voxels 6 faces, 8 corners, and 12 edges. Figure 6.11 illustrates the intermediate and final result of handling PVA.

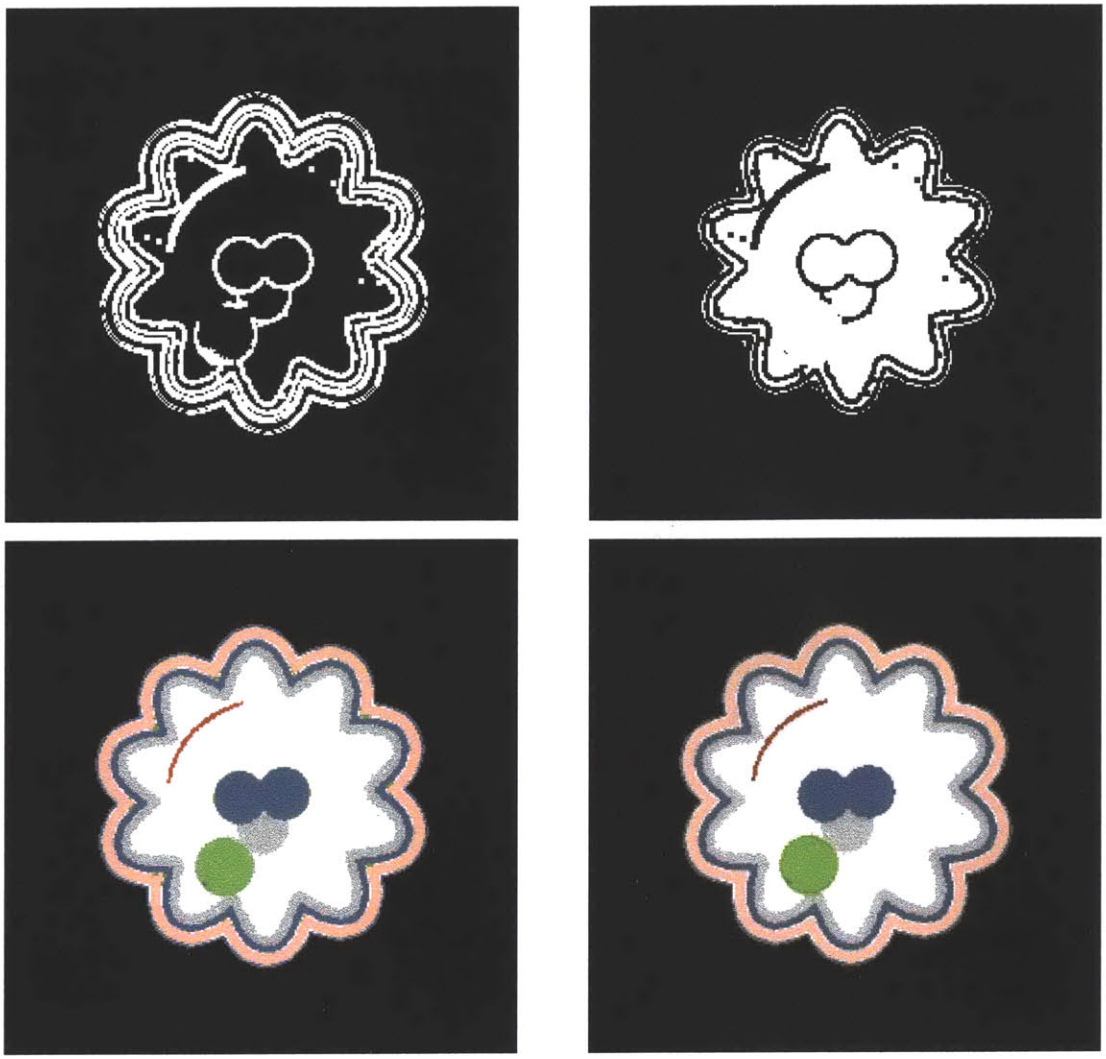

Figure 6.11. Handling of PVA. Intermediate processing steps of G4 are shown in the top row, with *HB* (boundaries of healthy tissues) on the left, and *PC* (SVP with 0's at candidate PVA locations on the right. Observe that the tumor boundary in *HB* does not appear in *PC*. This spares the tumor boundary from having errors as an artifact of correcting for PVA. This is the advantage of our algorithm over the customary, lossy operations of erosion/dilation. The bottom images represent results without (left) and with (right) PVA handling. Upon close inspection, the erroneous tumor classifications disappear from the ventricle/white matter interface and the scalp/CSF interface.

6.3.2 Correcting Misclassified Structures

Besides partial volume artifacts, another reason to consider the context of multiple structures is that an entire structure could be misclassified in a way that can be corrected based on its situation relative to other structures. We illustrate this concept with an example of edema misclassified as gray matter. Edema, or liquid diffused between cells, spreads finger-like into the white matter, while avoiding the gray matter and cortex

whose cell packing is too dense to harbor as much fluid. The extra-cellular fluid of edema and increased intra-cellular fluid of tumors can be confused when ascertaining the tumor/tissue interface. By knowing that edema always borders both white matter and tumor, we can resolve ambiguity resulting from its similar appearance to gray matter on T1-weighted MRI. We suggest this application as future work in Chapter 7.

6.4 Summary of CDN Layers #1-4

6.4.1 System Diagram

Figure 6.12 depicts the bi-directional communication between the first 4 layers of CDN. The factors involved in computing the healthy tissue posteriors are the image intensities, intensity prior, neighborhood prior, and shape prior. The factors involved in computing the tumor posterior are the probability of abnormality based on intensity, the complement of the shape prior, and the PVA prior.

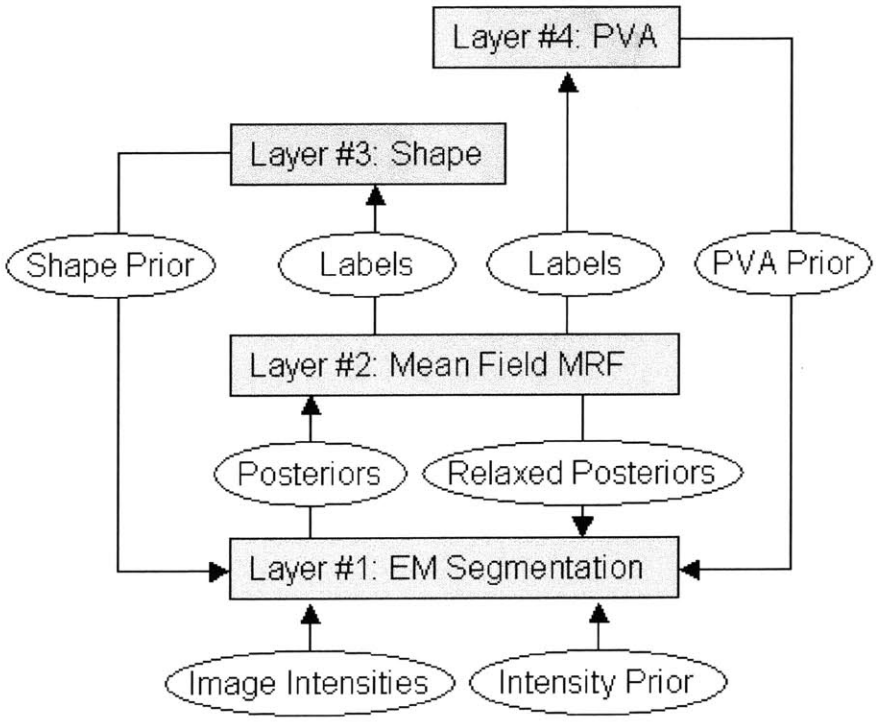

Figure 6.12. Bi-directional Information Flow. Although not drawn explicitly, the predication of Layer #4 upon Layer #3 is realized by executing Layer #3 (and then Layers #1-2 again) before Layer #4.

6.4.2 System Dynamics

The avoidance of race conditions was the motivation behind the hierarchy of predicated layers. The intent was for the predication to prevent any anomalous behavior due to unexpected critical dependence on the relative timing of events. Nonetheless, oscillations could occur when global and local forces exactly cancel, but such phenomena have not yet been observed in practice. Instead, the dominant shortcoming of the algorithm is convergence to local minima. Any input from the user is valuable in providing an initialization closer to a desirable minimum, which leads us to the 5th and final layer.

6.5 CDN Layer 5: Supervisory Classification

CDN Layer #5 differs from the four lower layers in that it adds context derived not from the image, but the user.

6.5.1 Intelligent Interaction

As described in Chapter 3, one of our goals was to produce an algorithmic framework that facilitates intelligent interaction with the user throughout the segmentation process. A segmentation should not be just presented, but be responsive. For example, suppose the user wishes to suggest, "No computer, that's not gray-matter, that's edema." If a human segmenter were told this, he or she would re-label not just the one voxel touched by the user, but all of the voxels whose classification should logically change in response. Obviously, this would include all neighboring voxels with the same properties as the first. But additionally, the presence of edema may have other ramifications. Since edema always borders tumor, some voxels whose classification had been borderline between tumor and gray-matter, may now be corrected with the new information that resolves the ambiguity. Thus, we seek a framework where a user's assertion (at any time within the segmentation process) of a single voxel's classification would have logical repercussions throughout the entire image.

6.5.2 The Role of the Supervisor

There are three reasons for segmentation systems to feature intelligent interaction with human users. The first one, described above, is corrective.

The second is prescriptive. Depending both on the nature of a patient's ailment, and the stage of the treatment process, a medical professional may be looking for different information from the segmentation. What exactly should be segmented as tumor? Is it just the enhancing portion, or also non-enhancing areas? Necrotic areas and edema may be unimportant at first, but the user may then decide to isolate them upon seeing the first segmentation.

The third reason is that different users have dissimilar definitions of how the segmentation boundaries should be drawn. One origin of the challenge for segmentation posed by inter-operator variability is user preference. Certain physicians prefer over-segmenting (larger than normal) of tumors, while others prefer under-segmenting. Figure 6.13 demonstrates the extreme differences displayed by the four experts applied to the tumorbase. We suggest more development of this application as future work in Chapter 7.

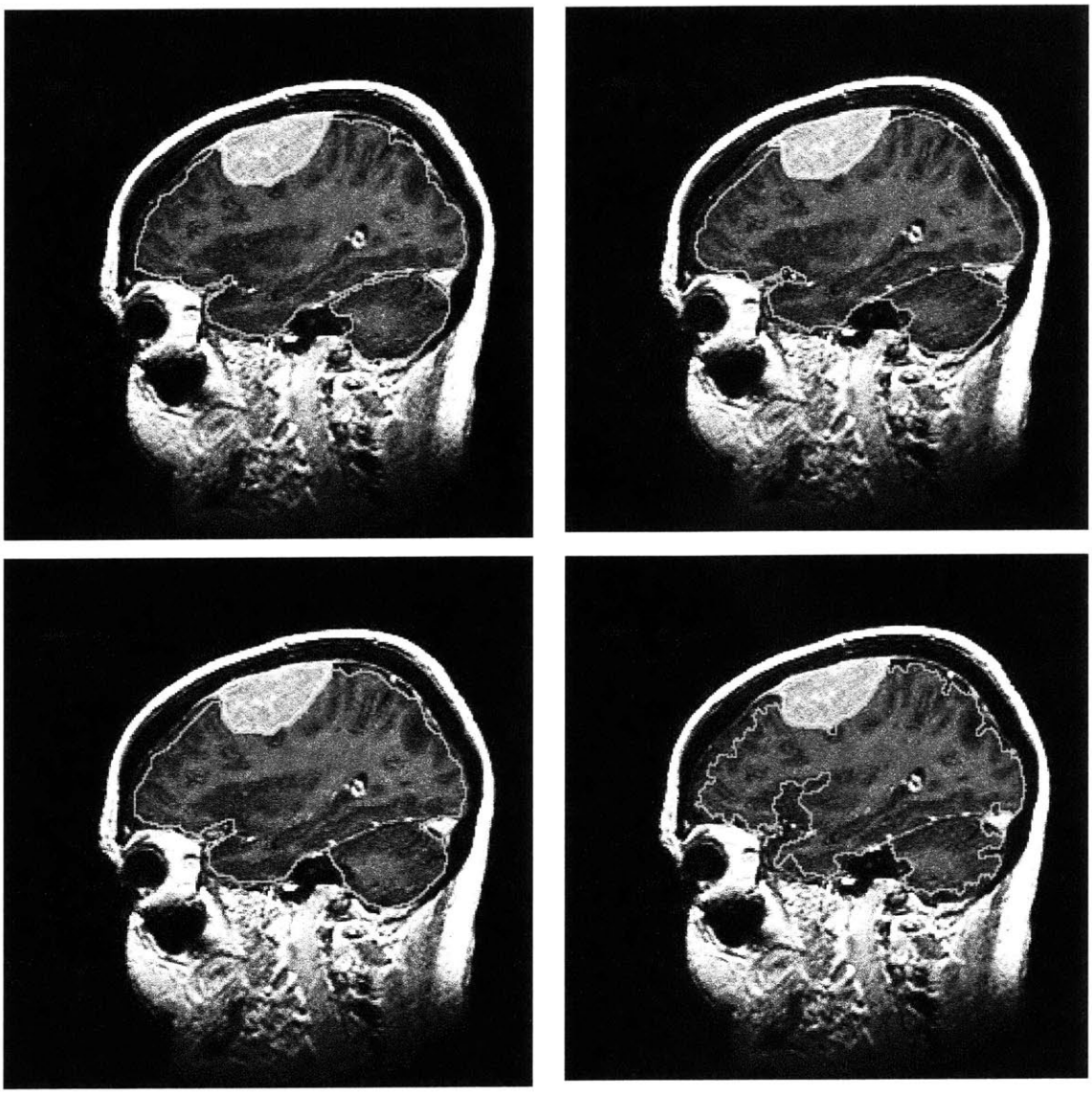

Figure 6.13. Intra-Operator Manual segmentations performed by the four different experts vary greatly, but systematically.

6.6 Results on Real Data

For analyzing algorithmic performance on real data, we return to the Tumorbase used in the experiments with Diagonalized NNPM. To produce training data, we augmented a publicly available anatomy atlas [BWHSPL], which is a healthy brain manually segmented into scores of structures. As shown in Figure 6.14, we reduced the number of structures to white matter, gray matter, and CSF for initial experiments. Future experiments can readily increase the number of modeled structures. For example, sub-

cortical gray matter such as the thalamus (indicated with an arrow) feature intensity and maximum-distance-to-boundary profiles situated between those of cortical gray matter and white matter. It would be straightforward to add these structures to our *model space* in the same manner that we employed anisotropic Markov random fields in Chapter 5.

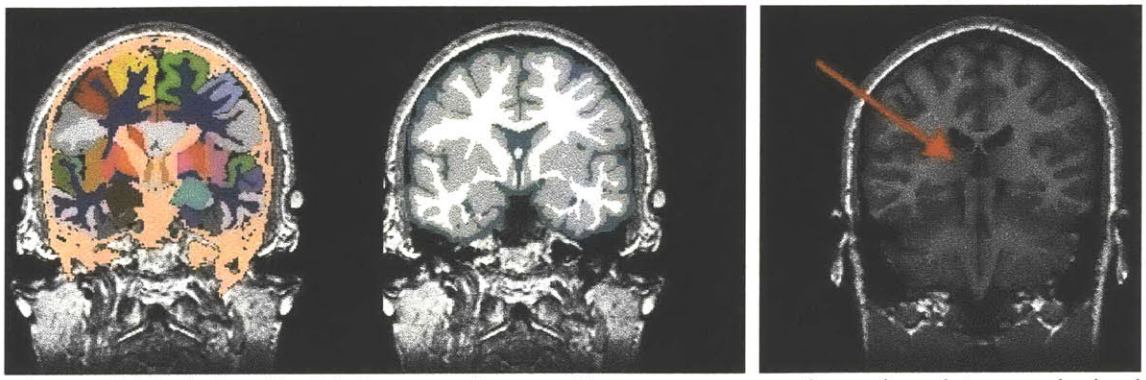

Figure 6.14. Atlas (Left:) Scores of manually structures can be reduced to our desired size of model space. (Right:) The thalamus features intensity and maximum-distance-to-boundary profiles situated between those of cortical gray matter and white matter, but is considered gray matter in many related works, including our initial experiments.

Table 6.5. Stationary Priors computed from the atlas in the center of Figure 6.14. Vessels, not included in the atlas, were added manually.

Tissue Class	Stationary Probability
White matter	0.28
Gray matter	0.50
CSF	0.21
Vessel	0.01

For spatially varying priors, we used the atlas of Figure 4.6. The atlas is rigidly registered to patient specific scans automatically[2] by maximizing mutual information [Wells96a].

We experimented with both automatic, unsupervised segmentation, and supervised segmentation that requires a few seconds of the user's time to draw a crude line on each structure of interest in order to collect sample prototype points. In all cases, the tissue

[2] Some manual assistance is required when the atlas and patient data sets are not both whole-head scans.

class variances are taken *a priori* from the training data, but the tissue class means are adapted to individual patient scans. The reason, in the supervised case, is that a small set of prototype points are an insufficient sampling to produce an accurate variance measurement. The reason, in the unsupervised case, is that the Gaussian model parameters are updated during the M-Step of EM to form a *Generalized EM* as derived in section 5.2.1. Allowing variances to adapt influences the algorithm to converge more slowly, and more likely to an undesirable local minimum. To prevent tumor intensities from adversely affecting the unsupervised clustering, the voxel contributions were weighted by their probability of pathology in the exact same manner as they are weighted in computation of the bias field, described in Section 4.2.2. The best results, by a wide margin, were achieved using both: a few seconds of user initialization and unsupervised clustering within generalized EM. In all experiments, we used tissue class standard deviations of 6, except 25 for vessels, apparently due to being a product of contrast injection. The following iteration schedule was selected empirically by allowing the iterations to proceed at each level until convergence (nearly all voxels ceased changing value).

Table 6.6. Bi-directional Communication between CDN Layers. Layer #1 passes its result to Layer #2, which passes its result to Layers #3-4, which return their results to Layer #1 in the form of spatially varying priors to. Layer #3 executes during the first outer iteration, and Layer #4 waits until the second in order to benefit from Layer #3's contribution.

Outer Iteration #	EM Iterations	MRF Iterations	Higher Layer
1	5	3	#3
2	3	12	#4

6.6.1 Results using Stationary Intensity Prior

Figures 4.15-4.17 display the convergence of the algorithm pictorially.

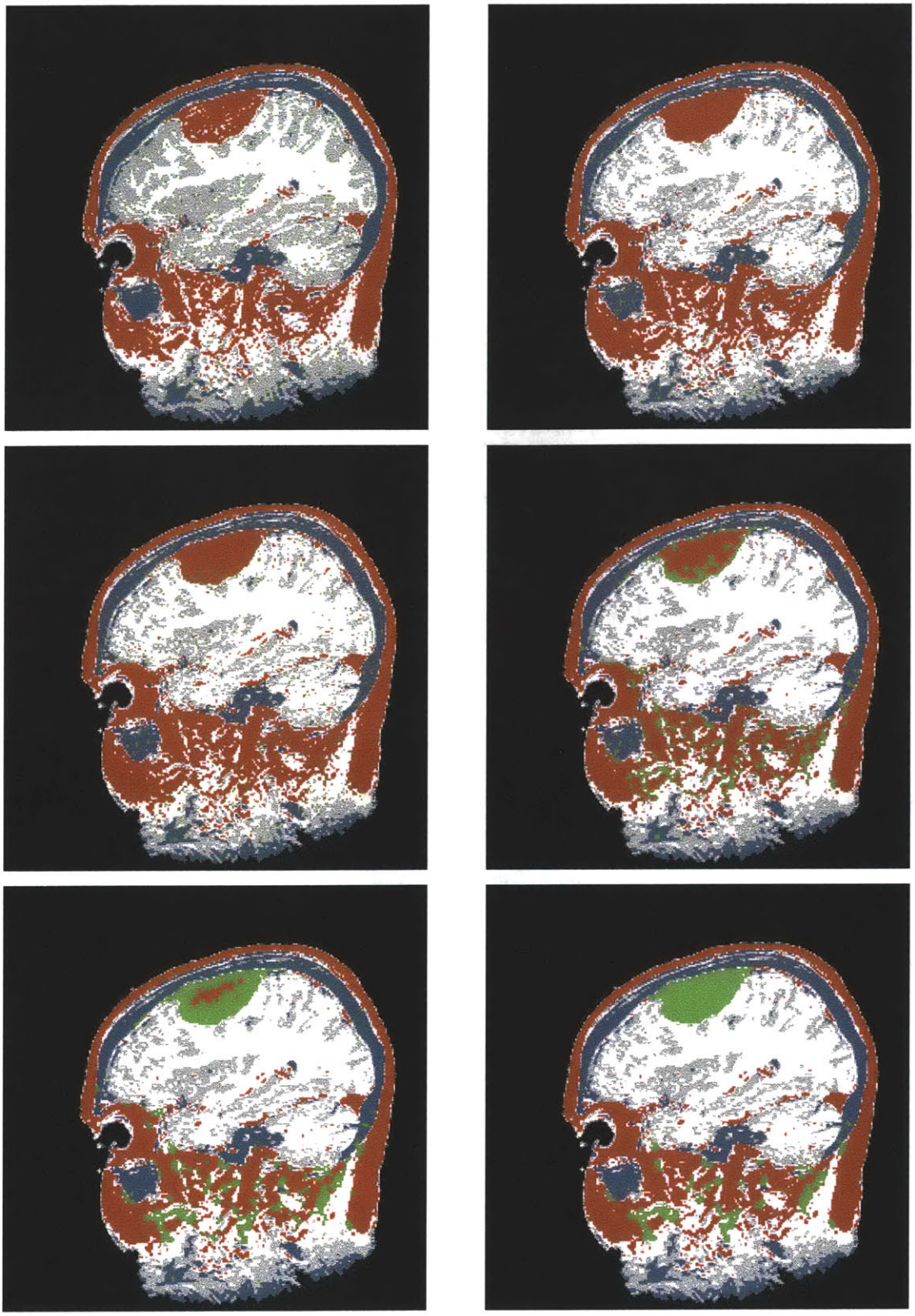

Figure 6.15. Intermediate Results. The top row shows the sequence of results during the first outer iteration, and the next 2 rows display the second. Observe how the spurious fragments of abnormality disappear except in the neck, which is irrelevant.

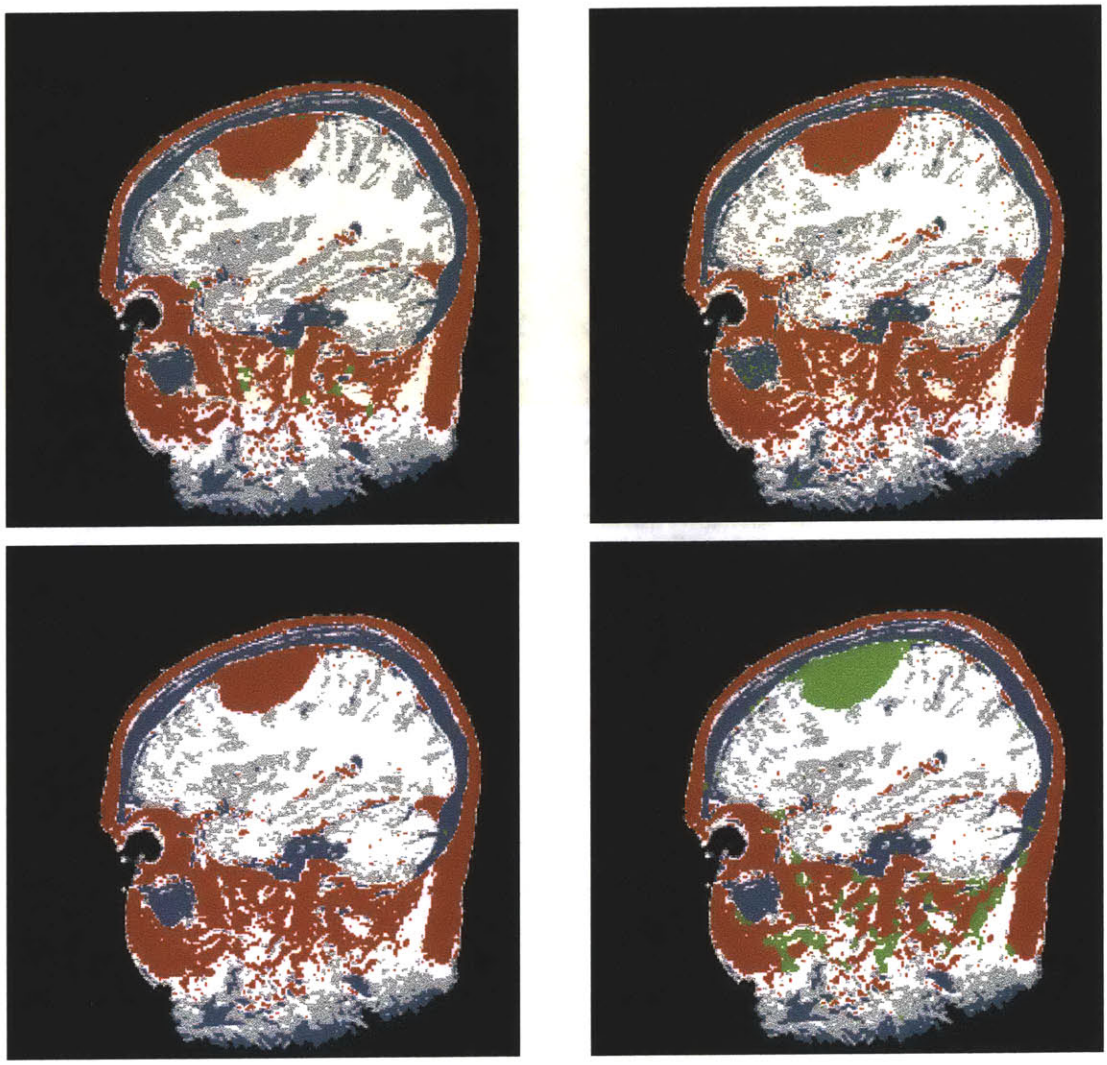

Figure 6.16. Dependence on Each Layer. This figure displays the final segmentation that results when a certain layer is absent from the framework. From top left to bottom right, are results missing EM, Layer #2, Layer #3, and Layer #4. Without EM, the parameters are not allowed to converge to a suitable explaination of the image. Without Layer #2, the probability of abnormality is unable to propogate across a structure. Without Layer #3, the abnormal shape is never recognized. And Without Layer #4, spurious tumor fragments remain as a partial volume artifact.

As a reminder of the color scheme: tumor (green), vessel (red), CSF (blue), white matter (white), and gray matter (gray), and skin (tan).

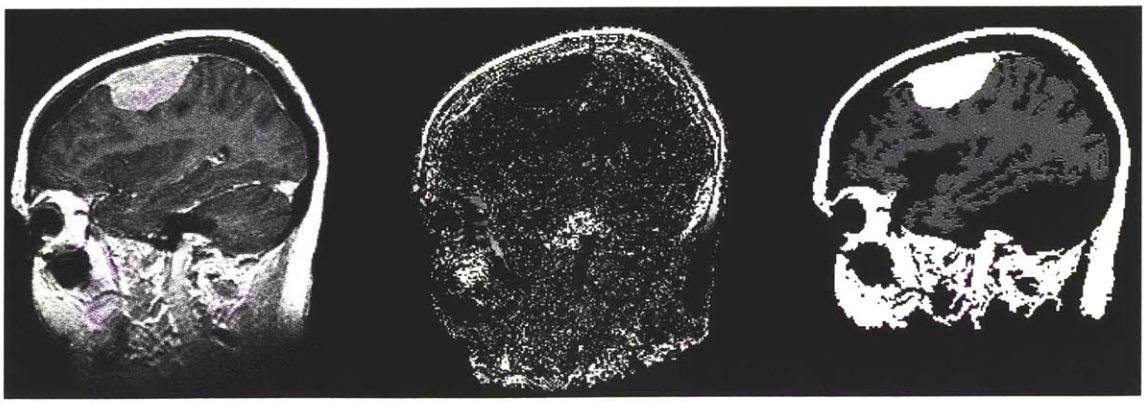

Figure 6.17. Abnormality Maps. From left to right are shown the input image, probability of abnormality based on Layer #1 and probability of abnormality based on Layer #2. Observe how the intensity information alone was nearly useless.

6.6.2 Results using Spatially Varying Intensity Prior

The complete atlas is composed of spatially varying probability maps for each healthy tissue class, in addition to a "brain mask" that restricts computation to occur within the approximate boundary of the ICC. Since the tumor is not represented in the probability maps, the probability maps are not applied during the first outer iteration. However, the brain mask is always applied – partly to speed the computation, but mostly to prevent structures outside the brain (which were not included in training) from interfering with the algorithm's convergence. Figures 6.18-19 illustrate the results on the hyper-intense tumor as well as a hypo-intense one.

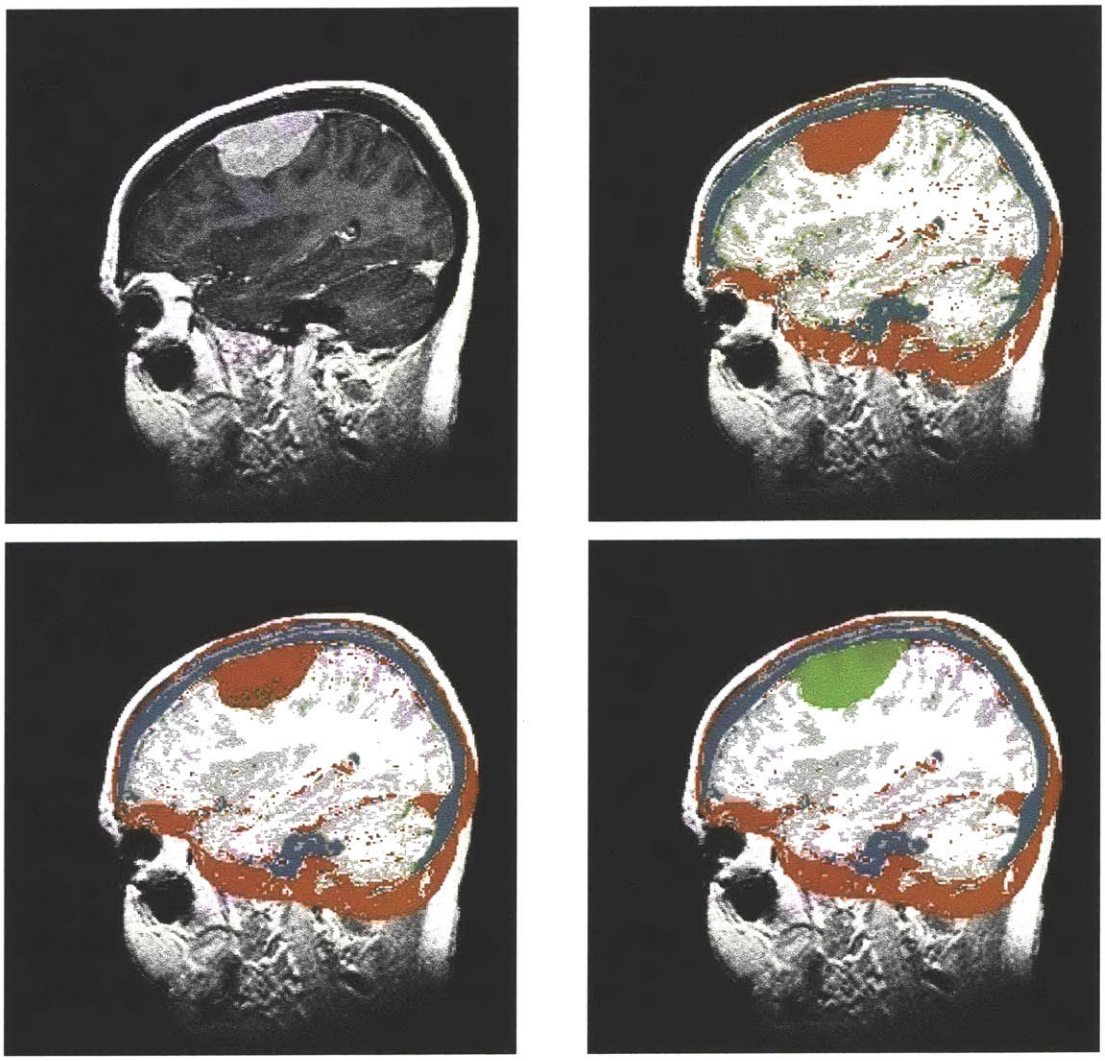

Figure 6.18. Results Using an Atlas. From top to bottom, left to right, is the sequence of results during convergence. Compared with the previous figure, the brain mask of the atlas prevents neck structures from corrupting the model parameter estimation. The probability maps of the atlas improve discernment of the interface between healthy structures, especially white matter and gray matter. This, in turn, produces better parameter estimation, which results in better tumor recognition.

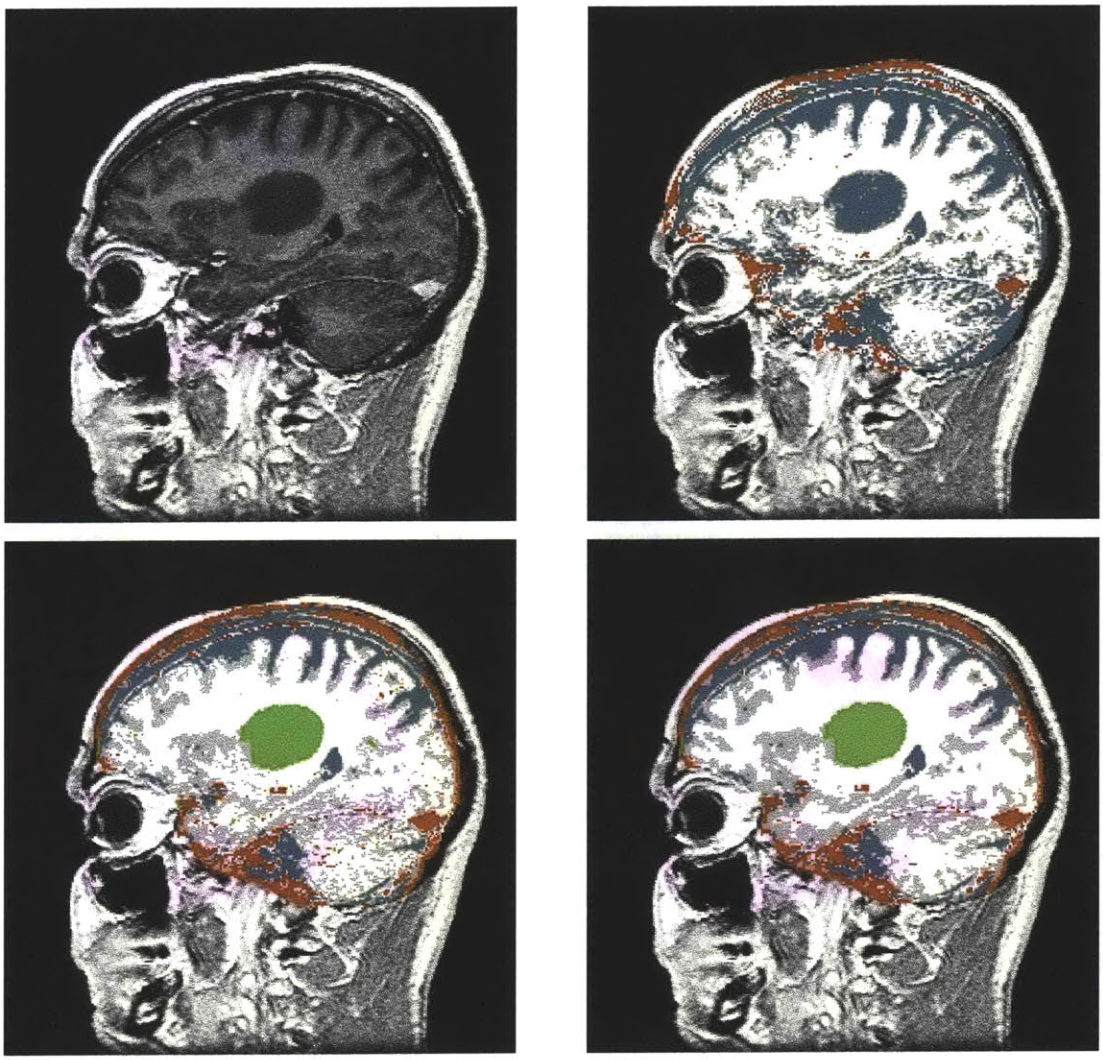

Figure 6.19. Hypo-Intense Tumor. Since the algorithm has no knowledge of tumors, it applies unchanged to both hyper-intense and hypo-intense tumors. As an interesting note, the process of convergence for this case was quite different from the other. In the previous figure, the few most abnormal voxels first recognized their identity as tumor, and neighborhood coherence propogated this information throughout the structure of voxels sharing similarity with respect to other properties. However, in the current figure, the shape prior alone was sufficient to identify most voxels as belonging to tumor. Hence, an observer watching the convergence notices the tumor "pop out" rather than evolve progressively.

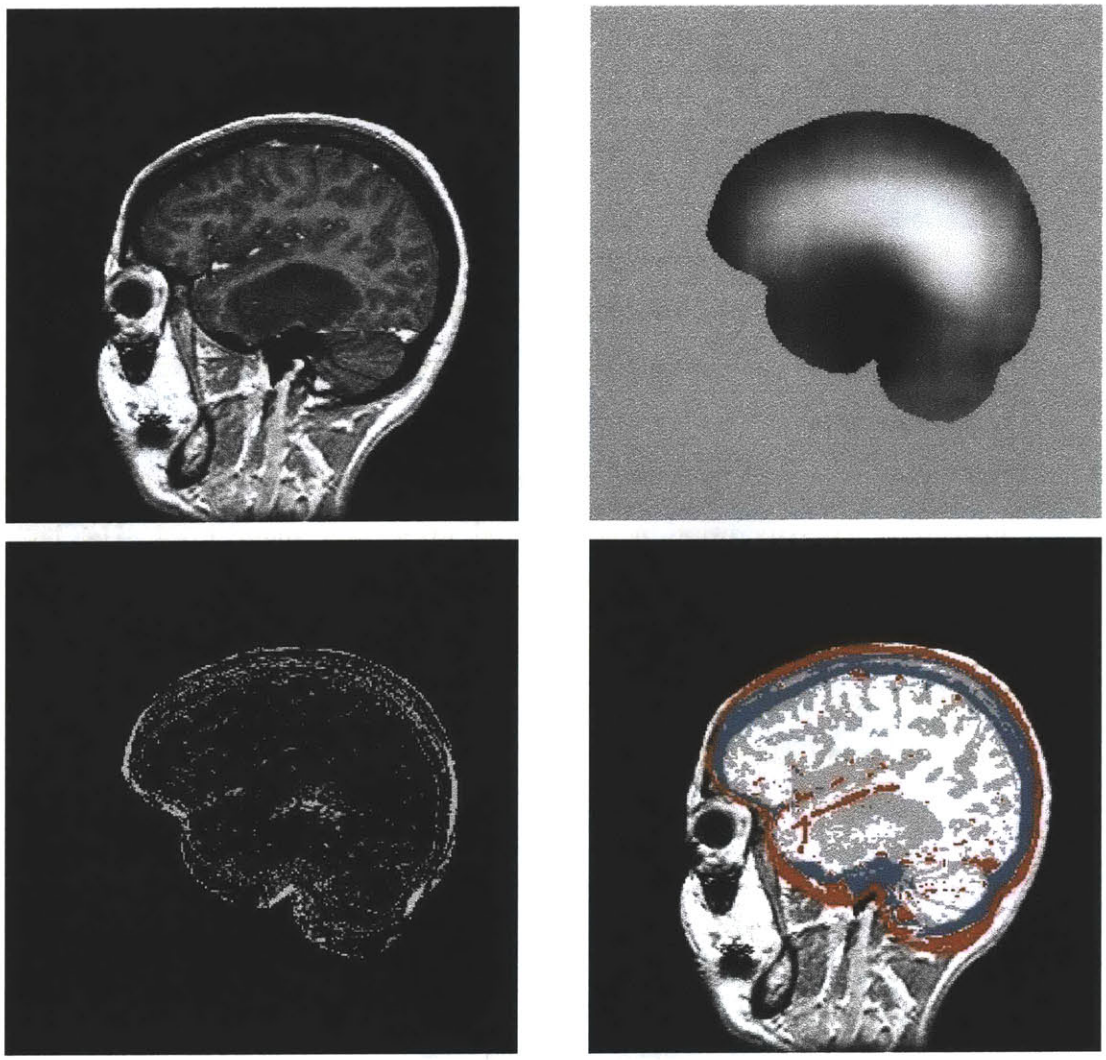

Figure 6.20. Algorithm Failure. This is a case where the algorithm failed to recognize the tumor in the input image (top left). The combination of the tumor's vast size and gradually-varying intensity distribution caused the bias field to overcompensate (top right). After bias correction, few tumor voxels had a significant probability of abnormality based on intensity (bottom left). The resulting tumor segmentation deviated little in shape from healthy structures (compare with the region above the tumor in the bottom right image). Future work can overcome this challenge with more robust bias estimation, such as a bias field constrained to be typical of head coils. Shape descriptors can also be made more robust using the higher dimesional model space of Figure 6.14.

6.7 Chapter Summary

We presented the three high-level layers of CDN for incorporating context on a broad scale. First, we incorporated intra-structure properties by computing shape descriptors over the results of the first 2 layers. Second, we incorporated inter-shape properties by examining relationships between structures computed by the first 3 layers. Together, all

layers of the CDN compute a probability map for pathology, from which one can delineate tumor boundaries.

To summarize the important principles asserted in this chapter:

6.1 The ACME segmenter incorporates context using an algorithm of linear time complexity by restricting information flow between computational nodes.

6.2 High-level layers can communicate to low-level layers via a Bayesian prior and an outer iteration.

6.3 The factors involved in computing the healthy tissue posteriors are the image intensities, intensity prior, neighborhood prior, and shape prior. The factors involved in computing the tumor posterior are the probability of abnormality based on intensity, the complement of the shape prior, and the PVA prior.

6.4 Morphological operations are appropriate when the image discretization is the very issue being addressed.

Chapter 7

Conclusion

7.1 Contribution Summary

The contributions of this thesis are two-fold. First, we proposed segmenting large brain tumors by training exclusively on healthy brains to recognize deviations from normalcy. Second, we designed a framework for a Contextual Dependency Network (CDN) that incorporates multiple levels of predicated context. This framework extends EM-based segmentation with region-level properties, and it allows information to flow bi-directionally between layers using either ACME or a multi-level MRF. Experimental results demonstrated our framework to be superior to nearest neighbor pattern recognition. We also improved NNPM with our diagonalization method that makes an effort to isolate micro- and macro texture by monotonically increasing window size with decreasing resolution.

The simple instantiation of the framework presented herein requires more sophisticated components to achieve clinically usable results. Regardless, the results are encouraging given the goal of this thesis, which is to solve the *recognition* problem for brain tumors. Existing methods have largely focused on boundary delineation, leaving the recognition task for humans. Together, this thesis and these methods could form an end-to-end solution for automatic recognition and delineation of brain tumors.

The specific findings of this thesis can be summarized with the principles tabulated in each of the chapter summaries. A brief review of these summaries is listed below:

Ch. 3 For general applicability, tumor segmentation systems should recognize deviations from normalcy, rather than identifying known features of tumors. They must answer the following two questions:

1.) What is normal?

2.) How is abnormality measured?

Ch. 4 Each voxel's contribution to the EM-based bias estimation is weighted by its typicality in order to produce an estimation that is robust to pathology. A function for computing a probability of pathology is based on integrating the area under the tails of Gaussian distributions, and is thus shifted from the origin, exponentially rising, and asymptotic.

Ch. 5 The Simple Smoothing, ICM, and Mean Field approximations use progressively "softer" functions and function parameters – moving from discrete mathematics to probabilities. Pathology is included in CDN Layer #2 by relaxing the weights computed by normalizing the combination of the posterior probabilities and the probability of pathology. Bi-directional communication between layers #1 and #2 can be achieved with 2-3 outer iterations.

Ch. 6 The ACME segmenter incorporates broad context using an algorithm of linear time complexity by restricting information flow between computational nodes. High-level layers communicate with low-level layers via a Bayesian prior and iteration.

Our approach of recognizing deviations from normalcy, rather than focusing on detecting specific features of certain pathology, holds promise for becoming more generally applicable in the broad, and rapidly growing, field of computer-aided medical image analysis. Toward this end, the next section presents several avenues of research for improving the ability of computers to assist patients along the road to recovery.

7.2 Future Directions of Research

7.2.1 Correcting Misclassified Structures

Sections 6.3.2 suggested future work in correcting misclassified structures with edema as an example target. CDN Layer #4 can be extended to leverage the information that

edema always borders both white matter and tumor, and this fact can be used to resolve ambiguity resulting from its similar appearance to gray matter on T1-weighted MRI.

Furthermore, we believe that the CDN framework is well-suited for producing intelligent human-computer interaction. Human input can be modeled as another G function in the ACME model so that human input is propagated throughout the other CDN layers. See Section 6.4 for more details.

7.2.2 More Sophisticated Shape Descriptors

Our framework used distance-to-boundary as a basic shape descriptor that readily facilitated measurements of normality. Other simple shape descriptors can take the form of curvature measurements or coefficients for combining a series of basis functions. The medical computer vision field is rapidly developing progressively better models of anatomic shape. Future developments in topological atlases and shape variations will be well suited for recognizing deviations from normalcy. In particular, we mentioned data dimensionality reduction schemes, such as PCA and nonlinear variants, in Section 2.2. For example, such a scheme could model the variability of the shape of white matter. Subsequently, cortical gray matter could be modeled as a sheet of certain thickness enveloping the outer surface of the white matter.

Given our model of recognizing deviations from normalcy, it is important to note that PCA is not able to answer how well new data are fit by the model in a non-Gaussian, probabilistic sense. Instead, the only criterion available is the squared distance of a given image from its projection into eigen-space. [Roweis98] has addressed this problem with an EM-based approach.

7.2.3 Non-rigid Atlas Registration

Our Bayesian framework incorporated spatially-varying statistical priors via rigid-registration with an atlas. Section 3.5 detailed several alternative approaches for richer implementations, including extending [Pohl02] so that the warping involved in the registration process would not be hindered by the presence of pathology.

7.2.4 Alternatives to MF-Optimized MRFs for Inter-Layer Communication

Section 6.2.4 derived a conclusion from several premises that could be satisfied using our multi-level MRF developed later in that section. Nonetheless, a multi-level MRF is not necessarily the optimal solution for meeting those requirements. Belief propagation networks [Belhumeur96, Weiss97, Freeman00, Yedidia02] could be designed to facilitate inter-layer communication: blending neighborhood-, region-, and global-level properties.

7.2.5 Alternative Metrics for Deviation from Normalcy

Our framework tended to fit Gaussian models to voxel- and region-level properties. The motivation for this was the convenience lent by normal distributions for expressing definitions of normalcy and measurements of abnormality. Further research can explore alternatives to Gaussian models in these instances.

Generally, alternative metrics of normality need to be explored. For a texture-based approach, consider [DeBonet97, DeBonet98]. For incorporating frequency information, consider a wavelet-based approach as performed with mammography [Laine94]. In comparison to a windowed Fourier transform which has a fixed resolution in the spatial and frequency domain, the resolution of the wavelet transform varies with the scale parameter, decomposing an image into a set of frequency channels.

7.2.6 Exhaustive Implementation of Multi-scale NNPM

Chapter 3 presented experiments run using a multi-scale implementation of NNPM. Further research should replicate these experiments using the full possible range of all scales and extents. PCA can be used to reduce the dimensionality of each patch for more efficient computation, and more convenient fitting of probability distributions to the occurrence of each possible patch. Although we used a measure of RMS error to characterize abnormality, a probabilistic approach can be taken given an extensive training set.

Furthermore, a full set of training data (300 cases) can be used instead of just a couple slices from the healthy hemisphere, as in our example.

Additionally, Section 3.4.2 suggested using non-rectangular windows. In essence, these are rectangular windows with a portion masked out to remove it from consideration.

Appendix

8.1 EM Segmentation

8.1.1 EM Segmentation: ML Derivation

[Wells96b] derived the EM segmentation algorithm from the standpoint of a MAP estimator of the bias field. We present here a slightly different derivation by deriving EM segmentation directly from [Dempter77]'s definition of EM based on ML estimation. Additionally, our derivation uses our imaging model from Chapter 2 to explain the validity of the various assumptions.

Define the following notation:

y	the observed log-transformed image intensities
b	bias field (additive to log-transformed data)
w	the tissue classification
L	set of all possible tissue labels, l
μ_l	mean of tissue class l
σ_l	standard deviation of tissue class l
i	index into voxel locations

Begin by writing the expectation that we wish to maximize:

$$\arg\max_{b'} Q(b' \mid b) \tag{8.1}$$

$$\arg\max_{b} \underset{w}{E}[\log p(y, w \mid b)] \tag{8.2}$$

Apply the definition of conditional probability:

$$\arg\max_{b} \underset{w}{E}[\log(p(y \mid w,b)p(w))] \tag{8.3}$$

Decouple the problem using the logarithm:

$$\arg\max_{b} \underset{w}{E}[\log p(y \mid w,b) + \log p(w)] \tag{8.4}$$

Assume from our imaging model in Chapter 2 that the bias field and tissue classes are statistically independent. While this is not completely true in practice, it is a viable approximation for mathematical tractability. Since we will maximize with respect to b, the $p(w)$ term can be dropped.

$$\arg\max_{b} \underset{w}{E}[\log p(y \mid w,b)] \tag{8.5}$$

Next, assume the statistical independence of voxel intensities. We noted in discussion of Chapter 2's imaging model that this was not completely true in practice, so we will relax this assumption later using Markov random fields.

$$\arg\max_{b} \underset{w}{E}\left[\log\left(\prod_i p(y_i \mid w_i,b_i)\right)\right] \tag{8.6}$$

Decouple using the logarithm:

$$\arg\max_{b} \underset{w}{E}\left[\sum_i \log p(y_i \mid w_i,b_i)\right] \tag{8.7}$$

Apply the fact that for linear functions, f, $f(E) = E(f)$:

$$\arg\max_{b} \sum_i \underset{w_i}{E}[\log p(y_i \mid w_i,b_i)] \tag{8.8}$$

Apply the measurement model. The probability of observing a particular image intensity, given knowledge of the tissue class and the bias field, is given by a Gaussian distribution centered at the biased mean intensity for the class:

$$\arg\max_{b} \sum_i \underset{w_i}{E}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma_{w_i}^2}}\exp\left(-\frac{(y_i - b_i - \mu_{w_i})^2}{2\sigma_{w_i}^2}\right)\right)\right] \tag{8.9}$$

Decouple using the logarithm:

$$\arg\max_{b} \sum_{i} E_{w_i}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma_{w_i}^2}} \right) - \frac{(y_i - b_i - \mu_{w_i})^2}{2\sigma_{w_i}^2} \right] \tag{8.10}$$

Drop the term not dependent on b:

$$\arg\min_{b} \sum_{i} E_{w_i}\left[\frac{(y_i - b_i - \mu_{w_i})^2}{2\sigma_{w_i}^2} \right] \tag{8.11}$$

To find the minimum, apply the zero-gradient condition by differentiating with respect to each component of b:

$$\frac{\partial}{\partial b_i} \sum_{i} E_{w_i}\left[\frac{1}{2\sigma_{w_i}^2} \left(\left(y_i - \mu_{w_i} \right) - b_i \right)^2 \right] = 0 \quad , \forall i \tag{8.12}$$

Consider that only the ith component of the summation depends on b_i. Also, move the differentiation inside the summation, and expand the quadratic:

$$E_{w_i}\left[\frac{\partial}{\partial b_i} \frac{1}{2\sigma_{w_i}^2} \left(\left(y_i - \mu_{w_i} \right)^2 - 2b_i \left(y_i - \mu_{w_i} \right) + b_i^2 \right) \right] = 0 \quad , \forall i \tag{8.13}$$

Apply the derivative:

$$E_{w_i}\left[\frac{b_i}{\sigma_{w_i}^2} - \frac{y_i - \mu_{w_i}}{\sigma_{w_i}^2} \right] = 0 \quad , \forall i \tag{8.14}$$

The expectation of a linear function is a linear function of expectations:

$$E_{w_i}\left[\frac{b_i}{\sigma_{w_i}^2} \right] = E_{w_i}\left[\frac{y_i - \mu_{w_i}}{\sigma_{w_i}^2} \right] \quad , \forall i \tag{8.15}$$

The bias field is independent of tissue class, so it can be pulled out of the expectation over the probabilities of tissue classes:

$$b_i = \frac{\underset{w_i}{E}\left[\frac{y_i - \mu_{w_i}}{\sigma^2_{w_i}}\right]}{\underset{w_i}{E}\left[\frac{1}{\sigma^2_{w_i}}\right]} \quad ,\forall i \tag{8.16}$$

We will revisit equation 8.16, but we'll simplify it for now:

$$b_i = \underset{w_i}{E}\left[y_i - \mu_{w_i}\right] \quad ,\forall i \tag{8.17}$$

Equation 8.17 states that the bias field is the expected value of the difference between the actual and predicted intensities. To conclude the derivation of the computation to be performed during the M-Step, we express the expectation:

$$b_i = \sum_{w_i} W_{w,i}(y_i - \mu_{w_i}) \quad ,\forall i \tag{8.18}$$

where

$$W_{w,i} = p(w_i \mid y_i, b_i) \tag{8.19}$$

The weights $W_{w,i}$ used to compute the weighted average are the probabilities of the hidden variables given the visible data and the current belief for the bias. As noted at the end of Section 4.1, the objective of the E-Step is merely to compute these weights. Apply Bayes' Theorem:

$$W_{w,i} = \frac{p(y_i, b_i \mid w_i)p(w_i)}{\sum_{l \in L} p(y_i, b_i \mid w_i = l)} \tag{8.20}$$

Apply the definition of conditional probability:

$$W_{w,i} = \frac{p(y_i \mid b_i, w_i)p(b_i \mid w_i)p(w_i)}{\sum_{l \in L} p(y_i \mid b_i, l)p(b_i \mid l)p(l)} \tag{8.21}$$

Since the bias field is independent of tissue class, $p(b \mid w) = p(b)$. For the same reason, $p(b)$ can be pulled out of the summation over w:

$$W_{w,i} = \frac{p(y_i \mid b_i, w_i)p(w_i)}{\sum_{l \in L} p(y_i \mid b_i, l)p(l)} \tag{8.22}$$

The remaining factors in equation 8.22 are known quantities:

$$p(y_i \mid b_i, w_i) = \frac{1}{\sqrt{2\pi\sigma_{w_i}^2}} \exp\left(-\frac{(y_i - b_i - \mu_{w_i})^2}{2\sigma_{w_i}^2} \right) \tag{8.23}$$

$p(w_i) =$ prior probability of tissue class. This is a stationary prior now, but we will use a spatially varying prior later.

To summarize, the EM algorithm performs the following iterations at each voxel location. Conceptually, the E-Step computes the weighting associated with each tissue class, and the M-Step computes the bias field as the weighted residual intensities:

E-Step:
$$W_{w,i} \leftarrow \frac{p(y_i \mid b_i, w_i)p(w_i)}{\sum_{l \in L} p(y_i \mid b_i, w_i = l)p(w_i = l)} \tag{8.24}$$

M-Step:
$$\beta_i \leftarrow \sum_{w_i} W_{w,i}(y_i - \mu_{w_i}) \tag{8.25}$$

We would now like to revisit equation 8.16 for computing the bias field. We performed the above derivation by using EM for a maximum-likelihood approach with no prior knowledge of the bias field. But in fact, we do know that the bias field is slowly varying, and we could apply this knowledge with a low-pass filter to attenuate the high-frequency components. We could impose this constraint by applying such a filter, F, to equation 8.17. Alternatively, we could apply F to both the numerator and denominator of equation 8.16 in order to remove any DC gain intrinsic to the filter. In our implementation, F is a 3-D, isotropic, boxcar filter with a radius approximately 1/10 the image radius. To summarize, replace equation 8.16 with:

M-Step:

$$b_i = \frac{F\left[\sum_{w_i} W_{w,i}\left(\frac{Y-\mu_{w_i}}{\sigma^2_{w_i}}\right)\right]_i}{F\left[\sum_{w_i} W_{w,i}\left(\frac{1}{\sigma^2_{w_i}}\right)\right]_i} \tag{8.26}$$

8.1.2 EM Segmentation: MAP Derivation

Observe that equation 8.26 is a filtered *Weighted Mean Residual* image divided by a filtered *Weighted Inverse Variance* image. This is the identical result as equation 22 in [Wells96b], but Wells used prior knowledge of the bias field from the beginning of the derivations. This would be equivalent to us computing EM based on MAP instead of ML, which involves replacing equation 8.02:

ML:

$$\arg\max_b E_w[\log p(y,w \mid b)] \tag{8.27}$$

MAP:

$$\arg\max_b E_w[\log p(b \mid y,w)] \tag{8.28}$$

Apply Bayes' rule to equation 8.27:

$$\arg\max_b E_w\left[\log \frac{p(y,w \mid b)p(b)}{p(y,w)}\right] \tag{8.29}$$

Decouple using the logarithm, and drop p(y,w) because it does not vary with β:

$$\arg\max_b E_w[\log p(y,w \mid b) + p(b)] \tag{8.30}$$

Thus, the only difference between MAP (equation 8.28) and ML (equation 8.02) is the *p(b)* term that captures prior knowledge of the nature of the bias field. From here, the derivations would proceed almost identically to what we have shown for the ML case. In the end, handling the *p(b)* term proves intractable to compute exactly, so [Wells96b] proposed an approximation identical to equation 8.26.

8.1.3 EM Segmentation: Rejection Class

In this thesis, we are applying EM segmentation to images of abnormal tissue not explained by our models of healthy tissue classes. The EM algorithm will attempt to fit

unhealthy tissue to a class for healthy tissue by adjusting the bias field. For breast segmentation, [Guilllemaud97] proposed using a rejection class to collect intensities that are not a reasonable fit to an established tissue class. Unlike the tissue classes modeled by Gaussian distributions, the rejection class has a uniform distribution with a probability just high enough to be greater than the tails of Gaussians distant from their means. To preserve the bias field's integrity, we wish to only calculate the bias where we know the tissue's classification. Then the bias field computed over known tissues will diffuse into the regions of uncertainty. The equation for computation of the Mean Residual during the M-Step changes to:

$$R_i = \sum_{\{w-wreject\}} W_{w,i} \left(\frac{y_i - \mu_{w_i}}{\sigma^2_{w_i}} \right) \tag{8.31}$$

Generally, we would like to weight our computation of the bias field by our confidence in knowledge of the tissue class. We will use this idea instead of a rejection class, and we will demonstrate results of this later in the chapter.

Bibliography

[BWHSPL] Surgical Planning Lab. http://splweb.bwh.harvard.edu:8000.

[Basri98] R. Basri, D. Roth, D. Jacobs. "Clustering Appearances of 3D Objects". In: *Computer Vision and Pattern Recognition (CVPR)*. Santa Barbara, CA: 1998; 414-420.

[Belhumeur96] P.N. Belhumeur. "A Computational Theory for Binocular Stereopsis". In: D.C. Knill, W. Richards, eds. *Perception as Bayesian Inference*. Cambridge University Press, 1996; 323-363.

[Besag74] J. Besag. "Spatial Interaction and the Statistical Analysis of Lattice Systems". *Journal of the Royal Statistical Society, Series B* 1974; 36:192-236.

[Besag86] J. Besag. "On the Statistical Analysis of Dirty Pictures". *Journal of the Royal Statistical Society, Series B* 1986; 48:259-302.

[Black97] P.M. Black, T. Moriarty, E. Alexander III, P. Stieg, E.J. Woodard, P.L. Gleason, C.H. Martin, R. Kikinis, R. Schwartz, F.A.Jolesz. "The Development and Implementation of Intraoperative MRI and its Neurosurgical Applications". *Neurosurgery* October 1997; 41:831-842.

[Borgefors86] G. Borgefors. "Distance Transforms in Digital Images". *Computer Vision, Graphics, and Image Processing* 1986; 34:344-371.

[Bouman91] C. Bouman. "Multiple Resolution Segmentation of Textured Images". *IEEE Trans. on Pattern Analysis and Machine Intelligence* February 1991; 13:99-113.

[Bregler95] C. Bregler, S.M. Omoundro. "Nonlinear Image Interpolation using Manifold Learning". *Advances in Neural Information Processing Systems* 1995; 7:973-980.

[BusinessWeek02] http://www.businessweek.com. "Focusing on Picture-Perfect Diagnoses". *Businesss Week* October 15 2002;

[Capelle00] A.S. Capelle, O. Alata, C. Fernandez-Maloigne, J.C. Ferrie. "Unsupervised Segmentation for Automatic Detection of brain Tumors in MRI". In: *IEEE International Conference on Image Processing (ICIP)*. Vancouver, BC, Canada: 2000; 613-616.

[Chandler87] D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford

University Press, 1987.

[Chatfield80] C. Chatfield, A.J. Collins. *Introduction to Multivariate Analysis.* Chapman & Hall, 1980.

[Choi91] H.S. Choi, D.R. Haynor, Y. Kim. "Partial Volume Tissue Classification of Multichannel Magnetic Resonance Images -- A Mixel Model". *IEEE Trans. Med. Imag.* Sept 1991; 10:395-407.

[Clark98] M.C. Clark, L.O. Hall, D.B. Goldgof, R. Velthuizen, F.R. Murtagh, M.S. Silbiger. "Automatic Tumor Segmentation Using Knowledge-Based Techniques". *IEEE Transactions on Medical Imaging* April 1998; 17:238-251.

[Clarke95] L.P. Clarke, R.P. Velthuizen, M.A. Camacho, J.J. Heine, M. Vaidyanathan, L.O. Hall, R.W.Thatcher, M.L. Silbiger. "Review of MRI Segmentation: Methods and Applications". *Magn Reson Imaging* 1995; 13:343-368.

[Cline87] H.E. Cline, C.L. Dumoulin, H.R. Hart, W.E. Lorensen, S. Ludke. "3D Reconstruction of the Brain from Magnetic Resonance Images Using a Connectivity Algorithm". *Magn Res Imag* 1987; 5:345-352.

[Cline90] H.E. Cline, W.E. Lorensen, R. Kikinis, F.A. Jolesz. "Three-Dimensional Segmentation of MR Images of the Head Using Probability and Connectivity". *Journal of Computer Assisted Tomography* November/December 1990; 14:1037-1045.

[Cover91] T.M. Cover, J.A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 1991.

[Cox01] T.F. Cox, M.A.A. Cox. *Multidimensional Scaling.* Chapman & Hall, 2000.

[DeBonet97] J.S. DeBonet. "Novel Statistical Multiresoltion Techniques for Image Synthesis, Discrimination, and Recognition". Masters Thesis, Massachusetts Institute of Technology, 1997.

[DeBonet98] J.S. Debonet, P. Viola. "Texture Recognition Using a Nonparametric Multi-scale Statistical Model". In: *Computer Vision and Pattern Recognition (CVPR).* Santa Barbara, CA: 1998; .

[Dempster77] A.P. Dempster, N.M. Laird, D.B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal Royal Statistical Society* 1977; 39:1-38.

[Dongfeng91] L. Dongfeng. "A Technique of Double-Resonant Quadrature Birdcage Coils". *Magn Reson Med* 1991; 19:180-185.

[Duda01] R.O. Duda, P.E. Hart, D.G. Stork. *Pattern Classification.* John Wiley & Sons, 2001.

[Dumanli00] H. Dumanli, J. Fielding, D. Gering, R. Kikinis. "Assessment of the Normal Female Cervix with MR imaging: a Comparison of Segmentation Techniques and Two Geometric Formulas". *Academic Radiology* July 2000; 7:502-505.

[Elfadel93] I.M. Elfadel. "From Fields to Networks". Ph.D. Thesis, Massachusetts Institute of Technology, 1993.

[Evans93] A.C. Evans, D.L. Collins, S.R. Mills, E.D. Brown, R.L. Kelly, T.M. Peters. "3D Statistical Neuroatomical Models from 305 MRI Volumes". In: *Proc. IEEE Nuclear Science Symp. Medical Imaging Conf.*. IEEE, 1993; 1813-1817.

[Falcao00] A.X. Falcao, J.K. Udupa, F.K. Miyazawa. "An Ultra-fast User-steered Image Segmentation Paradigm: Live Wire on the Fly". *IEEE Trans Med Imag* 2000; 19:55-61.

[Falcao98] A.X. Falcao, J.K. Udupa, S. Samarasekera, S. Sharma. "User-Steered Image Segmentation Paradigms: Live Wire and Live Lane". *Graphical Models and Image Processing* 1998; 60:233-260.

[Fielding00] J. Fielding, H. Dumanli, A. Schreyer, S. Okuda, D. Gering, R. Kikinis, F. Jolesz. "MR Based Three-Dimensional Modeling of the Normal Female Pelvic Floor in Women: Quantification of Muscle Mass". *American Journal of Roentgenology* March 2000; 174:657-660.

[Fischl02] B. Fischl, D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A.V.D. Kouwe, R. Killian, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, A.M. Dale. "Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain". *Neuron* January 2002; 33:341-355.

[Freeman00] W.T. Freeman, E.C. Pasztor, O.T. Carmichael. "Learning Low-Level Vision". *International Journal of Computer Vision* 2000; 40:25-47.

[Geman84] S. Geman, D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence* November 1984; 6:721-741.

[Gerig92] G. Gerig, O. Kübler, R. Kikinis, F.A. Jolesz. "Nonlinear Anisotropic Filtering of MRI Data". *IEEE Trans Med Imaging* 1992; 2:221-232.

[Gering01] D.T. Gering, A. Nabavi, R. Kikinis, N. Hata, L.J. Odonnell, W. Eric L. Grimson, F.A. Jolesz, P. Black, W. Wells III. "An Integrated Visualization System for Surgical Planning and Guidance Using Image Fusion and an Open MR". *Journal of Magnetic Resonance Imaging* June 2001; 13:967-975.

[Gering02a] Linear and Nonlinear Data Dimsionality Reduction. http://www.ai.mit.edu/people/gering/areaexam/.

[Gering02b] D.T. Gering, W.E.L. Grimson, R. Kikinis. "Recognizing Deviations from Normalcy for Brain Tumor Segmentation". In: T. Dohi, R. Kikinis, eds. *Medical Image Computing and Computer-Assisted Intervention*. Tokyo, Japan: Springer, 2002; 388-395.

[Gering99a] D.T. Gering, A. Nabavi, R. Kikinis, W.E.L. Grimson, N. Hata, P.

Everett, F.A. Jolesz, W.M. Wells III. "An Integrated Visualization System for Surgical Planning and Guidance Using Image Fusion and Interventional Imaging". In: C. Taylor, A. Colchester, eds. *Second International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cambridge, UK: Springer-Verlag, 1999; 809-819.

[Gering99b] D.T. Gering. "A System for Surgical Planning and Guidance using Image Fusion and Interventional MR". Masters Thesis, Massachusetts Institute of Technology, 1999.

[Giger00] M.L. Giger. "Computer-aided Diagnosis of Breast Lesions in Medical Images". *Computing in Science & Engineering* Sept-Oct 2000; 2:39-45.

[Ginneken02] B.V. Ginneken, B.M.H. Romeny, M.A. Viergever. "Computer-Aided Diagnosis in Chest Radiography: A Survey". *IEEE Trans Med Imaging* December 2001; 20:1228-1241.

[Grimson99] W.E.L. Grimson, R. Kikinis, F.A. Jolesz, P.M. Black. "Image-Guided Surgery". *Scientific American* June 1999; 280:62-69.

[Guillemaud97] R. Guillemaud, M. Brady. "Estimating the Bias Field of MR Images". *IEEE Transactions on Medical Imaging* June 1997; 16:238-251.

[Guttmann99] C.R.G. Guttmann, R. Kikinis, M.C. Anderson, M. Jakab, S.K. Warfield, R.J. Killiany, H.L. Weiner, F.A. Jolesz. "Quantitative Follow-up of Patients with Multiple Sclerosis using MRI: Reproducibility". *Magn Reson Imaging* 1999; 9:123-132.

[Haacke99] E.M. Haacke, R.W. Brown, M.R. Thompson, R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. US patent 6192263, 1999.

[Held97] K. Held, E.R. Kops, B.J. Krause, W.M. Wells III, R. Kikinis, H.-W. Muller-Gartner. "Markov Random Field Segmentation of Brain MR Images". *IEEE Transactions on Medical Imaging* December 1997; 16:878-886.

[Henkelman85] R.M. Henkelman. "Measurement of Signal Intensities in the Presence of Noise in MR Images". *Medical Physics* Mar/Apr 1985; 12:232-233.

[Hinton95] G.E. Hinton, M. Revow, P. Dayan. "Recognizing Handwritten Digits Using Mixtures of Linear Models". *Advances in Neural Information Processing Systems* 1995; 7:1015-1022.

[Jaakkola00] Tutorial on Variational Approximation Methods. http://www.ai.mit.edu/people/tommi/papers.html.

[Jaggi98] C. Jaggi, S. Ruan, D. Bloyet. "Mixture Modeling Applied to the Partial Volume Effect in MRI Data". In: *IEEE Engineering in Medicine and Biology Society*. 1998; 693-695.

[Jain95] R. Jain, R. Kasturi, B.G. Schunck. *Machine Vision*. McGraw Hill,

1995.

[Joe99] B.N. Joe, M.B. Fukui, C.C. Meltzer, Q. Huang, R.S. Day, P.J. Greer, M.E. Bozik. "Brain Tumor Volume Measurement: Comparison of Manual and Semiautomated Methods". *Radiology* 1999; 212:811-816.

[Jordan98] M.I. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul. "An Introduction to Variational Methods for Graphical Models". In: M.I. Jordan, ed. *Learning in Graphical Models*. Kluwar Academic Publishers, 1998; .

[Kapur99] T. Kapur. "Model Based Three Dimensional Medical Image Segmentation". Ph.D. Thesis, Massachusetts Institute of Technology, 1999.

[Karssemeijer95] N. Karssemeijer. "Detection of Stellate Distortions in Mammograms using Scale-Space Operators". In: *Proc. Information Processing in Medical Imaging*. 1995; 335-346.

[Kaus01] M.R. Kaus, S.K. Warfield, A. Nabavi, P.M. Black, F.A. Jolesz, R. Kikinis. "Automated Segmentation of MR Images of Brain Tumors". *Radiology* 2001; 218:586-591.

[Kichenassamy95] S. Kichenassamy, A. Kumar, P. Olver, A. Tannenbaum, A. Yezzi. "Gradient Flows and Geometric Active Contour Models". In: *International Conference on Computer Vision*. Boston, USA: 1995; 810-815.

[Kikinis96] R. Kikinis, M.E. Shenton, D.V. Iosifescu, R.W. McCarley, P. Saiviroonport, H.H. Hokama, A. Robatino, D. Metcalf, C.G. Wible, C.M. Portas, R.M. Donnino, F.A. Jolesz. "A Digitial Brain Atlas for Surgical Planning, Model-driven Segmentation, and Teaching". *IEEE Transactions on Visualization and Computer Graphics* September 1996; 2:232-241.

[Koenderink84] J. Koenderink. "The Structure of Images". *Biological Cybernetics* 1984; 50:363-370.

[Laidlaw98] D.H. Laidlaw, K.W. Fleischer, A.H. Barr. "Partial-Volume Bayesian Classification of Material Mixtures in MR Volume Data Using Voxel Histograms". *IEEE Trans. Med. Imag.* Feb 1998; 17:74-86.

[Laine94] A.F. Laine, S. Schuler, J. Fan, W. Huda. "Mammographic Feature Enhancement by Multiscale Analysis". *IEEE Trans Med Imaging* December 1994; 13:725-740.

[Langan92] D.A. Langan, K.J. Molnar, J.W. Modestino, J. Zhang. "Use of the Mean-field Approximation in an EM-based Approach to Unsupervised Stochastic Model-based Image Segmentation". In: *ICASSP*. 1992; 57-60.

[Lange84] K. Lange, R. Carson. "EM Reconstruction Algorithms for Emission and Transmission Tomography". *J. Comput. Asst. Tomo.* 1984; 8:306-316.

[Leemput01a] K.V. Leemput, F. Maes, D. Vandermeulen, P. Suetens. "Automated Segmentation of Multiple Sclerosis Lesions by Model Outlier Detection". *IEEE Transactions on Medical Imaging* August 2001; 20:677-688.

[Leemput01b] K.V. Leemput, F. Maes, D. Vandermeulen, P. Suetens. "A Statistical Framework for Partial Volume Segmentation". In: W.J. Niessen, M.A. Viergever, eds. *MICCAI 2001: Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention*. Utrecht, The Netherlands: Springer, 2001; 204-212.

[Leemput99a] K.V. Leemput, F. Maes, D. Vandermeulen, P. Suetens. "Automated Model-Based Bias Field Correction of MR Images of the Brain". *IEEE Transactions on Medical Imaging* October 1999; 18:885-896.

[Leemput99b] K.V. Leemput, F. Maes, D. Vandermeulen, P. Suetens. "Automated Model-Based Tissue Classification of MR Images of the Brain". *IEEE Transactions on Medical Imaging* October 1999; 18:897-908.

[Leuttgen93] M.R. Luettgen. "Multiscale Representations of Markov Random Fields". *IEEE Trans. on Signal Processing* December 1993; 41:3377-3396.

[Leventon00] M.E. Leventon. "Statistical Models for Medical Image Analysis". Ph.D. Thesis, Massachusetts Institute of Technology, 2000.

[Li01] S.Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 2001.

[Lorensen87] W.E. Lorensen, H.E. Cline. "Marching Cube: A High Resolution 3-D Surface Construction Algorithm". *Computer Graphics* 1987; 21:163-169.

[Marr80] D. Marr, E. Hildreth. "Theory of Edge Detection". *Proceedings of the Royal Society of London* 1980; B207:187-217.

[Marr82] D. Marr. *Vision*. Freeman, 1982.

[Miller02] E.G. Miller. "Learning from One Example in Machine Vision by Sharing Probability Densities". Ph.D. Thesis, Massachusetts Institute of Technology, 2002.

[Moon02] N. Moon, E. Bullitt, K.V. Leemput, G. Gerig. "Automatic Brain and Tumor Segmentation". In: T. Dohi, R. Kikinis, eds. *Medical Image Computing and Computer-Assisted Intervention*. Tokyo, Japan: Springer, 2002; 372-379.

[Neal98] R.M. Neal, G.E. Hinton. "A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants". In: M.I. Jordan, ed. *Learning in Graphical Models*. Kluwar Academic Publishers, 1998; 355-367.

[ODonnell01] L. O'Donnell, C.F. Westin, W.E.L. Grimson, J.R. Alzola, M.E. Shenton, R. Kikinis. "Phase-Based User-Steered Image Segmentation". In: W.J. Niessen, M.A. Viergever, eds. *MICCAI*

2001: Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention. Utrecht, The Netherlands: Springer, 2001; 1022-1030.

[Olabarriaga01] S.D. Olabarriaga, A.W. Smeulders. "Interaction in the Segmentation of Medical Images: a Survey". *Medical Image Analysis* June 2001; 5:127-142.

[Papoulis91] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.

[Parisi88] G. Parisi. *Statistical Field Theory*. Adison-Wesley, 1988.

[Pham00a] D.L. Pham, J.L. Prince. "Unsupervised Partial Volume Estimation in Single-Channel Image Data". In: *Mathematical Methods in Biomedical Image Analysis (MMBIA)*. Hilton Head Island, South Carolina: 2000; 170-177.

[Pham00b] D.L. Pham, C. Xu, J.L. Prince. "A Survey of Current Methods in Medical Image Segmentation". *Annual Review of Biomedical Engineering* 2000; 2:

[Poggio85] T. Poggio, V. Torre, C. Koch. "Computational Vision and Regularization Theory". *Nature* 1985; 317:314-319.

[Pohl02] K.M Pohl, W.M. Wells, A. Guimond, K. Kasai, M.E. Shenton, R. Kikinis, W.E.L. Grimson, S.K. Warfield. "Incoperating Non-Rigid Registration into Expectation Maximization Algorithm to Segment MR Images". In: T. Dohi, R. Kikinis, eds. *Medical Image Computing and Computer-Assisted Intervention*. Tokyo, Japan: Springer, 2002; 564-572.

[Rexilius01] J. Rexilius, S.K. Warfield, C.R.G. Guttman, X. Wei, R. Benson, L. Wolfson, M. Shenton, H. Handels, R. Kikinis. "A Novel Nonrigid Registration Algorithm and Applications". In: W.J. Niessen, M.A. Viergever, eds. *MICCAI 2001: Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention*. Utrecht, The Netherlands: Springer, 2001; 923-931.

[Rice95] J.A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.

[Rosenfeld71] A. Rosenfeld, M. Thurston. "Edge and Curve Detection for Visual Scene Analysis". *IEEE Transactions on Computers* 1971; 20:562-569.

[Roweis00] S.T. Roweis, L.K. Saul. "Nonlinear Dimensionality Reduction by Locally Linear Embedding". *Science* December 2000; 290:2323-2326.

[Roweis98] S. Roweis. "EM Algorithms for PCA and SPCA". *Advances in Neural Information Processing Systems* 1998; 10:

[Ruan00] S. Ruan, C. Jaggi, J. Xue, J. Fadili, D. Bloyet. "Brain Tissue Classfication of Magnetic Resonance Images Using Partial Volume

Modeling". *IEEE Trans. Medical Imaging* Dec 2000; 4:1179-1187.

[SPM] Statistical Parametric Mapping. http://www.ifl.ion.ucl.ac.uk/spm/.

[Saito94] T. Saito, J.I. Toriwaki. "New Algorithms for Euclidean Distance Transformation of an n-Dimensional Digitized Picture with Applications". *Pattern Recognition* 1994; 27:1551-1565.

[Santago95] P. Santago, H.D. Gage. "Statistical Models of Partial Volume Effect". *IEEE Trans. Image Processing* Nov 1995; 4:1531-1540.

[Schenk95] J.F. Schenk, F.A. Jolesz, P.B. Roemer, others. "Superconducting open-configuration MR imaging system for image-guided therapy". *Radiology* 1995; 195:805-814.

[Schroeder92] W. Schroeder, J. Zarge, W. Lorensen. "Decimation of Triangle Meshes". *Computer Graphics* 1992; 26:65-78.

[Sears93] W. Sears, M. Sears. *The Baby Book.* Little, Brown and Company, 1993.

[Sethian99] J.A. Sethian. *Level Set Methods and Fast Marching Methods.* Cambridge University Press, 1999.

[Simmons94] A. Simmons, P.S. Tofts, G.J. Barker, S.R. Arridge. "Sources of Intensity Nonuniformity in Spin Echos Images at 1.5T". *Magn Reson Med* 1994; 32:121-128.

[Simmons96] A. Simmons, S.R. Addridge, G.J. Barker, S.C.R. Williams. "Simulation of MRI Cluster Plots and Application to Neurological Segmentation". *Magnetic Resonanance Imaging* 1996; 14:73-92.

[Sipser97] M. Sipser. *Introduction to the Theory of Computation.* PWS Publishing, 1997.

[Sled98] J.G. Sled, G.B. Pike. "Understanding Intensity Nonuniformity in MRI". In: W.M. Wells III, A. Colchester, S. Delp, eds. *First International Conference on Medical Image Computing and Computer-Assisted Intervention.* Boston: Springer-Verlag, 1998; .

[Stansfield80] J.L. Stansfield. *Conclusions from the Commodity Expert Project.* MIT AI Memo #601. http://www.ai.mit.edu/research/publications/, 1980.

[Tenenbaum00] J.B. Tenenbaum, V.d.Silva, J.C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction". *Science* December 2000; 290:2319-2323.

[Thirion98] J.-P. Thirion. "Image Matching as a Diffusion Process: an Analogy with Maxwell's Demons". *Medical Image Analysis* 1998; 2:243-260.

[Turk91] M. Turk, A. Pentland. "Eigenfaces for Recognition". *Journal of Cognitive Neuroscience* 1991; 3:71-86.

[Vannier85] M.W. Vannier, R.L. Butterfield, D. Jordan, et al.. "Multispectral Analysis of Magnetic Resonance Imaging". *Radiology* 1985; 154:221-224.

[Wang01] D. Wang, D.M. Doddrell. "A Segmentation-based and Partial-volume-compensated Method for an Accurate Measurement of Lateral Ventricular Volums on T1-Weighted Magnetic Resonance Images". *Magn Reson Imaging* 2001; 19:267-272.

[Warfield00] S.K. Warfield, M. Kaus, F.A. Jolesz, R. Kikinis. "Adaptive, Template Moderated, Spatially Varying Statistical Classification". *Medical Image Analysis* October 2000; 4:43-55.

[Warfield01] S.K. Warfield, J. Rexilius, P.S. Huppi, T.E. Inder, E.G. Miller, W.M. Wells III, G.P. Zientara, F.A. Jolesz, R. Kikinis. "A Binary Entropy Measure to Assess Nonrigid Registration Algorithms". In: W.J. Niessen, M.A. Viergever, eds. *MICCAI 2001: Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention*. Utrecht, The Netherlands: Springer, 2001; 266-274.

[Warfield95] S. Warfield, J. Dengler, J. Zaers, C.R.G. Guttmann, W.M. Wells III, G.J. Ettinger, J. Hiller, R. Kikinis. "Automatic Identification of Grey Matter Structures from MRI to Improve theSegmentation of White Matter Lesions". *Journal of Image Guided Surgery* 1995; 6:326-338.

[Warfield98b] S.K. Warfield, F.A. Jolesz, R. Kikinis. "A High Performance Computing Approach to the Registration of Medical Imaging Data". *Parallel Computing* Sept 1998; 24:1345--1368.

[Weiss97] Y. Weiss. "Interpreting Images by Propagating Bayesian Beliefs". *Advances in Neural Information Processing Systems* 1997; 9:908-915.

[Wells96a] W.M. Wells III, P.A. Viola, H. Atsumi, S. Nakajima, R. Kikinis. "Multi-ModalVolume Registration by Maximization of Mutual Information". *Medical Image Analysis* 1996; 1:35-51.

[Wells96b] W.M. Wells III, R. Kikinis, W.E.L. Grimson, F.A. Jolesz. "Adaptive Segmentation of MRI Data". *IEEE Trans Med Imaging* 1996;

[Winston92] P.H. Winston. *Artificial Intelligence*. Addison-Wesley, 1992.

[Witkin83] A. Witkin. "Scale Space Filtering". In: *Proc. International Joint Conference on Artificial Intelligence*. Karlsruhe: 1983; .

[Yedidia02] Understanding Belief Propagation and its Generalizations. Technical Report 2000-26, MERL, Mitsubishi Electric Research Labs. http://www.merl.com.

[Yezzi97] A. Yezzi, S. Kichenassaym, A. Kumar, P. Olver, A. Tannenbaum. "A Geometric Snake Model for Segmentation of Medical Imagery". *IEEE Transactions on Medical Imaging* April 1997; 16:199-209.

[Youmans96] J.R. Youmans. *Neurological Surgery: A Comprehensive Reference Guide to the Diagnosis and Management of Neurosurgical Problems*. Saunders, 1996.

[Yuille86] A.L. Yuille, T.A. Poggio. "Scaling Theorems for Zero Crossings". *IEEE Transactions on Pattern Analysis and Machine Intelligence*

January 1986; 8:15-25.

[Zhang92] J. Zhang. "The Mean Field Theory in EM Procedures for Markov Random Fields". *IEEE Transactions on Image Processing* 1992; 40:2570-2583.

[Zhu97] Y. Zhu, H. Yan. "Computerized Tumor Boundary Detection Using a Hopfield Neural Network". *IEEE Transactions on Medical Imaging* February 1997; 16:55-67.