# Confidence-Based Segmentation of MR Imagery Using Region and Boundary Information with Nonlinear Scale-Space and Fast Marching Level Sets

Jerod Weinman, George Bissias, Edward Riseman,
Allen Hanson, Joseph Horowitz†
Dept. of Computer Science †Dept. of Mathematics and Statistics
University of Massachusetts-Amherst
Amherst, MA 01003

April 2003
UM-CS-2003-017

**Abstract**

Automatic segmentation of stroke lesions in magnetic resonance imagery is a difficult problem because anatomical knowledge is required for the most accurate decisions. Without such knowledge, classification rules seem inconsistent. We propose a hybrid boundary and region based segmentation model built upon nonlinear scale-space and geometric active contours that captures the various segmentation rules necessary to segment lesions. After a user selects a point within damaged tissue and another point within healthy tissue, the image is examined at several levels of detail. At each such scale, the lesion is segmented several times by varying a parameter that models the range of criteria for boundaries between healthy and damaged tissue. These segmentations are collected, and the relative frequency of tissue being labeled lesion is regarded as a measure of confidence in the classification of the tissue as damaged. Experiments compare volumes and segmentations of lesions given by physicians to those given by the automatic method. Performance upper bounds are established by matching automatic segmentation parameters (scale, threshold, and/or confidence) for each image with each physician's hand segmentation. These results may be compared with results that fix parameters for a particular physician's segmentation or all physicians' segmentations. Sensitivity to parameter values and initializations are tested as well. With little initialization, the model achieves zero error on average with a standard deviation near clinically useful bounds. A modest amount of additional input gives zero error on each image.

# Contents

# List of Algorithms



# List of Figures



# List of Tables

# 1 Introduction

## 1.1 Background

Magnetic resonance imagery (MRI) is noisy by nature, and much research has been done on the properties of the noise, as well as methods for overcoming it [3, 14, 27]. Segmenting MRIs in general is a challenge for this reason as well as others that may depend on the physiology of what is being imaged. There is a wealth of literature on MRI segmentation, much of it geared toward specific applications [23, 39]. Ischemic stroke lesions in MR imagery are difficult to segment for their own reasons, but especially shape complexity and ambiguity. Creating and evaluating automatic methods for segmenting lesions is difficult because ground-truth is elusive and stroke lesions are widely varied.

The concept of a ground-truth segmentation is elusive. There is much variability in manual segmentation, and it occurs on three levels. Different doctors may give substantially different segmentations of the same lesion. The same doctor on different occasions might produce different segmentations. Additionally, when evaluating an MRI with several slices, the same doctor may not exhibit a consistent strategy throughout all slices. Ambiguity in a lesion boundary is one cause of the variance, while some other lesions are so convoluted that accurate manual segmentation is prohibitively tedious. Even segmenting the most basic lesions is a taxing manual process, so a large scale study is infeasible and a large window of results in a small study should be expected.

The second difficulty is that lesions vary quite widely. Much of the finesse in diagnosing tissue as affected by stroke is due to anatomical knowledge, which can be difficult to incorporate in automatic methods. Some imaging methods show distinct tissue types differently, requiring an allowance for multiple intensities within a lesion. In other cases, doctors know that stroke never stops at cortical boundaries, so even the faintest signal in regions neighboring the primary affected area are likely to be labeled stroke, whereas in other lesions a similar drop in intensity might signal a boundary with healthy tissue or a normal imaging artifact. The result is that simple rules relying on intensity patterns and changes therein are inconsistent across lesions, even on the small dataset explored in this work.

Given the hurdles before us, what is the best we can do? The simplest solution is to find the consistent strategy that gives the least error. This works reasonably well on average, but is unsatisfying on lesions where there is ambiguity or subtle anatomical knowledge is required. Alternatively, one might trade the overhead of image registration and attempt to tie in more medical domain knowledge in hopes of eliminating the error. We argue, however, that with minimal additional user interaction, these subtleties can be hypothesized automatically and confirmed manually.

The intensity nuances found in stroke lesions are well-modeled by a computer vision theory called scale-space, which facilitates the inspection of an image at varying levels of detail. For segmenting both complex and regular shapes quite easily, we employ geometric deformable models, which require a single click to initialize and inflate like balloons that fill to lesion boundaries.

The paper is organized as follows: Section 2 discusses the nature of MR image slices and a method for normalizing intensity across slices; Section 3 gives an overview and

technical review of nonlinear diffusion scale-space, how it relates to MR imagery, and some implementation details; Section 4 covers the fundamentals of level sets and the fast marching method for geometric contours as well as a function for integrating region and boundary information for robust segmentation; Section 5 describes the segmentation model incorporating scale-space and active contours, how these are integrated into a segmentation confidence, introduces a new metric for segmentation variance, and gives results on several experiments. The remainder of this section discusses the nature of stroke lesions in more detail, the results of previous work, and our new approach.

## 1.2 Problem Description

Ischemic stroke can be a critical medical condition. Treatment for stroke is often a time-sensitive issue, and quick, consistent computer segmentations should be useful [23]. Given an MRI of a stroke victim, the primary task is to separate voxels (volume elements) deemed clinically to be lesion from those voxels which are not. The primary difficulty is that the segmentation occurs over an indirect observation, rather than directly from the generating process (i.e. tissue). Of course, the noisiness of MRIs is one issue that needs to be handled, but others include tissue mixtures, low contrast boundaries, multi-modal lesion intensity, complex shape, and anatomical knowledge.

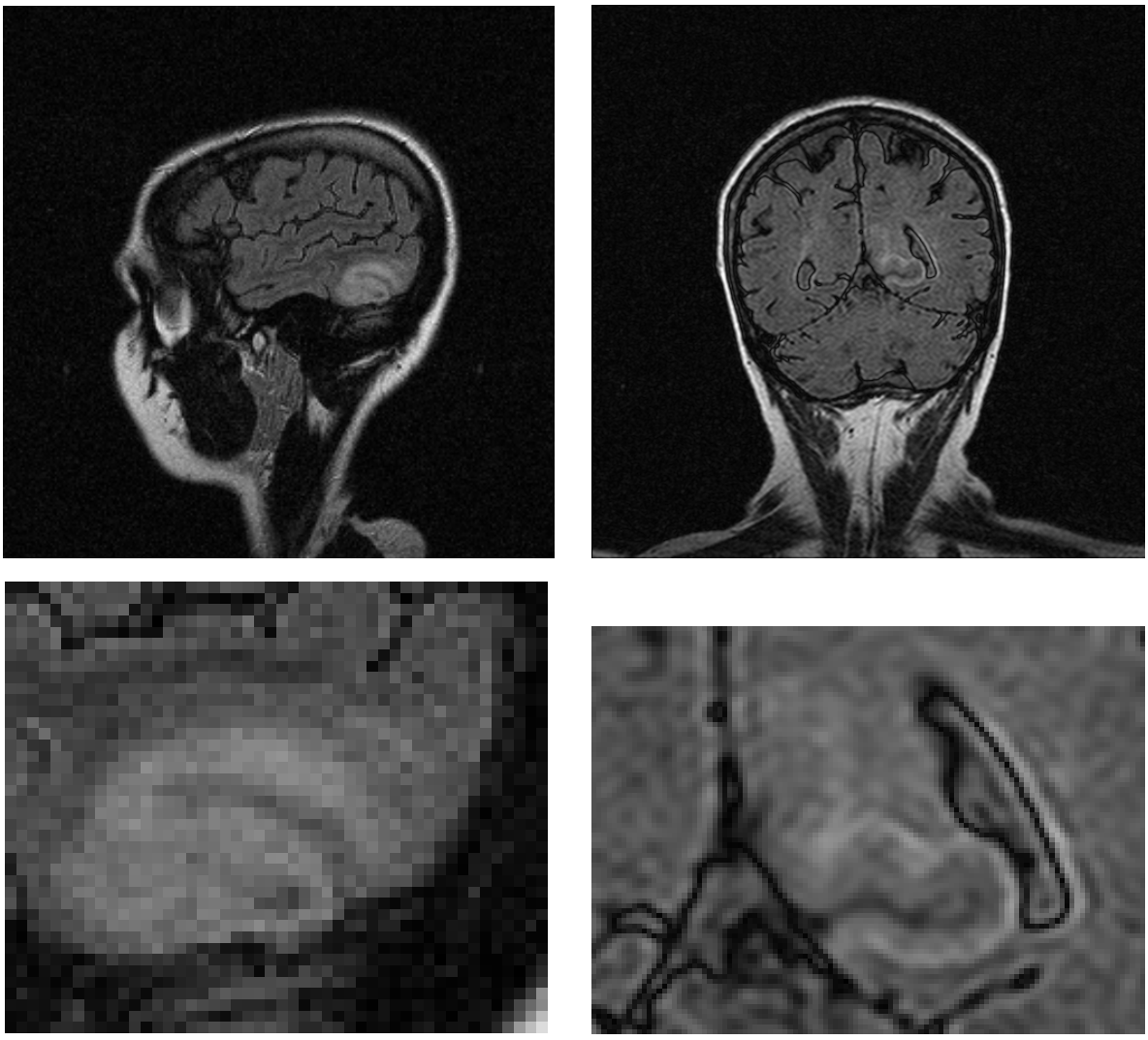
### 1.2.1 Normal and Abnormal Tissue

Because voxels are discrete observations, their intensities are a spatial average of responses. The volume a voxel covers might include more than one tissue type, which gives rise to terms like “mixels,” which views voxels as mixtures of tissues, rather than the observation from a single class [6]. However, when manually segmenting lesions, physicians refer to areas as being “normal” and “abnormal,” which hints at a slightly different segmentation philosophy. They must compare regions around the infarct (the area of dead tissue) to other regions of the same scan because tissue near folds and ventricles may brighten, which is normal.

This comparison can involve questions of inclusion or exclusion. Figure 1(a) shows a sagittal slice where the tissue is bright near the folds of the brain, and slightly darker toward the “center.” While the lesion in the lower right portion of the brain has a relatively clear left boundary, as the lesion extends to the right, the tissue at the right edge of the brain must be compared to tissues in other similar regions, not just to the darkest of gray regions in the center of the folds. After consideration, a physician who segmented this image labeled the upper-right corner as an abnormal region of tissue affected by stroke.

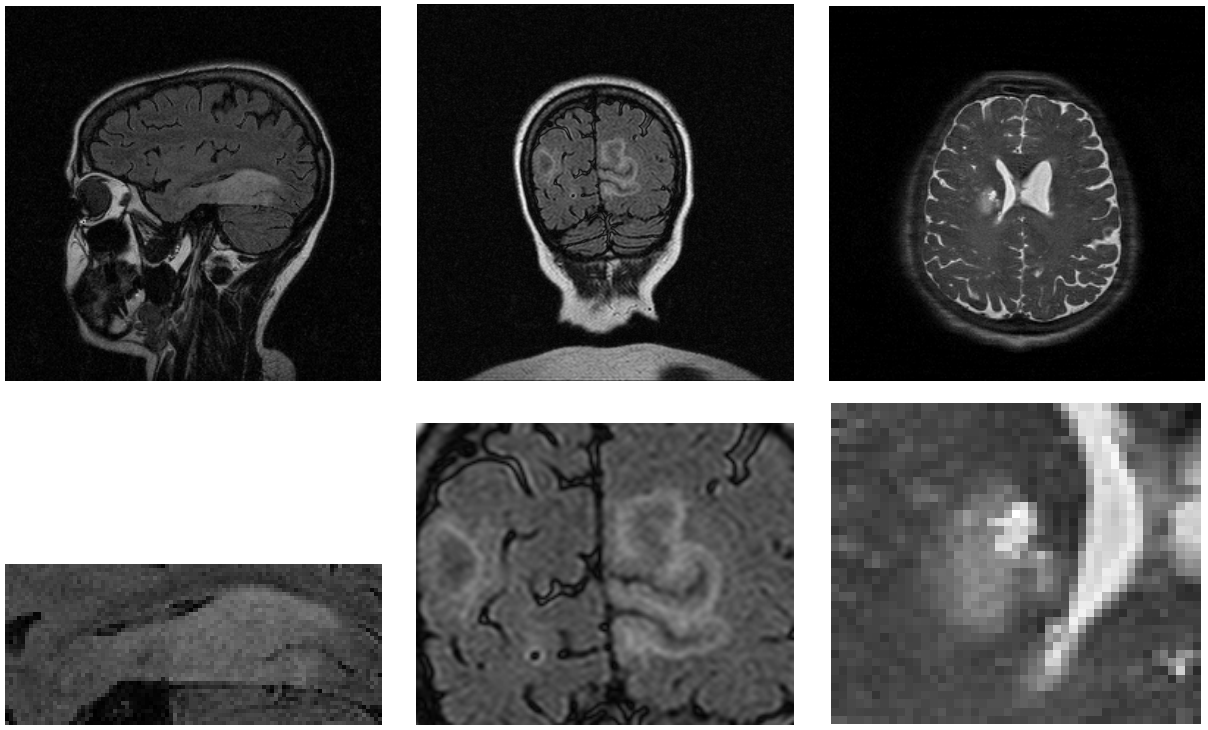
Figure 1(b) illustrates a lesion that borders part of the ventricle in a coronal sequence. The tissue surrounding both the left and right ventricles is bright in appearance. Since the region immediately inside both ventricles is normally bright, the physician excludes that tissue from the segmentation of lesion immediately left of the right ventricle. gray

Finally, different doctors may simply have entirely different ideas for what constitutes normal and abnormal. We received two segmentations of one image by different doctors that differ in volume by over sixty percent.



(a) (b)

Figure 1: Comparing normal and abnormal regions. (a) Probable inclusion, $T_1$ sequence; (b) Probable exclusion, FLAIR sequence.



(a) (b) (c)

Figure 2: Stroke lesions: (a) Ambiguous boundary, $T_1$ sequence; (b) Complex shape, FLAIR sequence; (c) Multi-modal intensity, $T_2$ sequence.

### 1.2.2 Ambiguous Boundaries

While some lesions have boundaries that are very obvious under some MRI modalities, others exhibit intensities that slowly drift toward that of healthy brain tissue. Physiologically, this may be due to a gradual change in the healthiness of the tissue. This ambiguity often makes precise segmentations difficult, or even non-existent. Therefore, some variance is to be expected from both manual human and automatic segmentations.

The left boundary of the lesion in Figure 2(a) is difficult to pinpoint; there does not appear to be a definite delineation between bright lesion and gray brain matter. While physicians attempt to use consistent criteria when segmenting the slices of one MRI, these may change from one time to the next, giving rise to intra-observer segmentation variability. Furthermore, some physicians may be more conservative than others in classifying abnormalities. For instance, when manually segmenting a slice similar to 2(a), one physician admitted the possibility of varying the segmentation by up to a centimeter in one dimension.

### 1.2.3 Complex Shape

Figure 2(b) shows a slice from a stroke lesion with a highly irregular shape. This is because the lesion envelops several folds of the brain containing cerebro-spinal fluid (CSF). While

these areas are not lesion in the most technical sense, physicians often include them in manual segmentations, both because stroke can induce swelling, which reduces the "empty" volume in these areas, but also because the complex shape makes their exclusion difficult. However, exclusion of CSF is a valid segmentation policy [15]. The main advantage of an automatic system pointed to by physicians would be consistency.

Interpreting CSF in a scan falls roughly into one of three cases. In the first, there is clear and definite separation between CSF and brain tissue. In the second, there is only moderate separation between the regions—black CSF areas are averaged in the imaging process with the brighter edema of stroke, resulting in a kind of gray. Finally, in a third case there is no separation, where stroke edema has completely squeezed out CSF from between the folds, leaving perhaps only the barest trace of it in the image. The segmentations required in the first and last cases are quite clear: definite exclusion and inclusion, respectively. It is the middle, in-between case that is the most ambiguous. Here, the consistency of a computer segmentation will be advantageous, because the strategy may vary from physician to physician.

### 1.2.4 Variable Tissue Response

Lesions and their surrounding areas also tend to contain regions of varying brightness. Determining which regions ought to be considered lesion is a challenge different from ambiguous borders. The response of different tissue types is different under various MRI modalities. For instance, the edema in a $T_2$ image shows up much more brightly in subcortical white matter than in the cortex. Thus, images may contain abnormal regions that are brighter than healthy tissue, yet darker than the brightest lesion areas. Figure 2(c) illustrates such a case.

Very often, physiological and anatomical constraints aid physicians in determining what is abnormal. For instance, physicians know that stroke does not respect cortical boundaries. Therefore, when even the faintest response is registered in a region, they are likely to extend a segmentation to cover it because they know that the infarct most likely extends to a different boundary, such as the extreme right of the lesion in Figure 1(a).

## 1.3 Relation to Previous Work

Our previous work used a parametric active contour model [17]. In two dimensions, this may be thought of either as a curve that actively seeks a local energy minimum, or as a curve subject to "forces" that pull it into particular directions. The same thing happens when the force of gravity pulls a ball down a ramp; the ball is just minimizing its potential energy. The energy of (or force upon) an active contour is defined by both the internal properties of the curve and external properties of the image. Ordinarily, the primary external force upon the contour is gradient magnitude, so that the contour is drawn to edges.

The previous segmentation process began with the user drawing a rough contour around the lesion in some user chosen base slice. The active contour then minimized energy based on the usual gradient forces, presumably to lesion boundaries. A separate similar (in the geometric sense) contour was then automatically initialized outside the original contour and

allowed to deform. The resulting two regions within the outer contour provided initialization for a course-to-fine statistical packing method requiring samples from the lesion and non-lesion voxel populations. The results of a segmentation due to statistical packing were then used as an additional force upon the contour [23, 38, 39]. In this sense, the work was a hybrid of both gradient (boundary) and intensity (region) based segmentation methods. Yet the method was not without its limitations.

The active contour model was represented as a stack of 2D contours rather than a true 3D surface, which in practice is a parametric mesh. The contour in each slice was deformed under forces derived from the image, the statistical segmentation, and the contours in neighboring slices (although no two contours were completely independent of one another, since the forces can propagate from slice to slice). This approach suffered from several limitations. First and foremost, the 2D stack of contours restricts the topology of a segmentation. While a stroke lesion may be a contiguous volume in 3D, a single slice through it may render the lesion as multiple disconnected blobs in 2D. Additionally, a 2D slice may contain a "hole" within a lesion, which may or may not be a hole in 3D as well. For example, an object and a slice through it may have different Euler characteristics. However, the explicit nature of parametric methods does not allow them to change topology without the addition of complicated schemes. Moreover, extending parametric contours to be fully three-dimensional requires significant computational overhead [8]. All these issues reflect poorly on the appropriateness of both parametric contours and the constrained 2D contours to imitate a full 3D model. In short, lesions change topology from slice to slice, which is difficult for the parametric contour stack to handle. Topology can also differ from lesion to lesion, preventing the use of a constant number of multiple contours for slice segmentation, a practice common to heart ventricle segmentation.

In a standalone setting, the statistical packing method operated effectively in 3D, which is advantageous. However, it required a large amount of user input to define a region of interest. Moreover, the statistical tests involved assumed a unimodal distribution on lesion intensity, which is not always the case (cf. §1.2.4).

One of the reasons the previous work was so successful in achieving volumes comparable to those of hand segmentations is that it captured the segmentation mechanics (and limitations therein) of the manual process. Doctors segment lesions by examining them on a slice by slice basis, often glancing at neighboring slices for cues, which is precisely the way the stack of active contours is built and deformed. Furthermore, because segmentation can be difficult and tedious, doctors tend to draw regularized curves surrounding affected areas, rather than producing detailed complex curves that might include strictly tissue, rather than CSF (§1.2.3). While our previous work dovetailed nicely with the manual segmentations provided by physicians, an approach that allows for full 3D consideration of MR imagery and its possible intricate structures should prove advantageous.

## 1.4 Overview of Our Approach

As suggested in §1.1 and detailed in §1.2, different lesions exhibit different intensity patterns and thus often require different rules for segmentation. We do not propose the use of prior anatomical knowledge, but to approximate it with rules that can vary what constitutes

a significantly bright intensity and a significantly large change in intensity. This characterization allows us to consider varying strengths of boundaries, such as contrasting the oft less-significant cortical boundaries with those that are stronger. Simultaneously, such rules will also allow for considering that different lesions are often characterized by different brightnesses.

We therefore propose to segment the MRI at several scales and with varying intensity "thresholds" for a different type of active contour. In this two-dimensional parameter space, the frequency of a voxel being labeled lesion by the algorithm will be regarded as a measure of confidence in the classification of that voxel. The more often a voxel is included, the more certain the algorithm, without prior anatomical knowledge, can be about the results. To see why this is the case, we may reason about the two parameters independently. When the scale is increased, more and more details of the lesion fade away. As protrusions blend into background, only the strongest, most stable parts of the lesion remain and are included in segmentations. As small swaths of CSF merge with the lesion, holes within begin to be included in segmentations. The result is that these regions are included (excluded) only at the lowest scales and the only most persistent areas of the lesion (background) are included (excluded) in *every* segmentation. Orthogonal to scale, changing an intensity "threshold" will also change segmentation results. When the threshold is set just above the intensity of the background, the classification criterion is quite liberal and many voxels are included. As the threshold rises, segmentations become more conservative, and only the brightest voxels, those of which we can be most certain, are included. Again, the result is that the brightest voxels most likely to be lesion are included in nearly all segmentations, and those less likely are included in fewer. Voxels included less frequently may correspond to neighboring structures, spurious image properties, or areas of lesion that have not affected tissue as strongly.

When each voxel has an inclusion frequency, the frequency isocontours of the resulting segmentations might be used in one of many ways. If a more automatic segmentation is desired, a certain frequency threshold could be specified, say 50%, giving a contour corresponding to voxels that were labeled lesion in at least half the segmentations. Alternatively, trained technologists could manually choose the frequency contour they feel best specifies the lesion boundaries.

Such frequency isocontours are a shortcut for the scale selection problem. Large changes in image entropy have been proposed as boundaries between stable scales [36, 37]. While we do not make use of this method, it might be an interesting avenue for further research. Instead, since the frequency of a voxel's inclusion is proportional to the number of scales at which it is included, the isocontours may be thought of as segmentations at empirically derived pseudo-scales.

The extensions discussed in this paper overcome the restrictions mentioned in the previous section. They allow for a selection of a segmentation from various scales and thresholds, which captures the variability of conservative and liberal segmentations often given by physicians.

The rest of the paper is organized as follows: Section 2 deals with the nature of MR imagery, the nature of stroke lesions, and the problem of their segmentation in such imagery. Scale-space and diffusion are discussed in Section 3, while Section 4 covers geometric de-

formable models such as level set and fast marching methods. Experiments and results are presented in Section 5, and the conclusions appear in Section 6.

# 2 MR Images

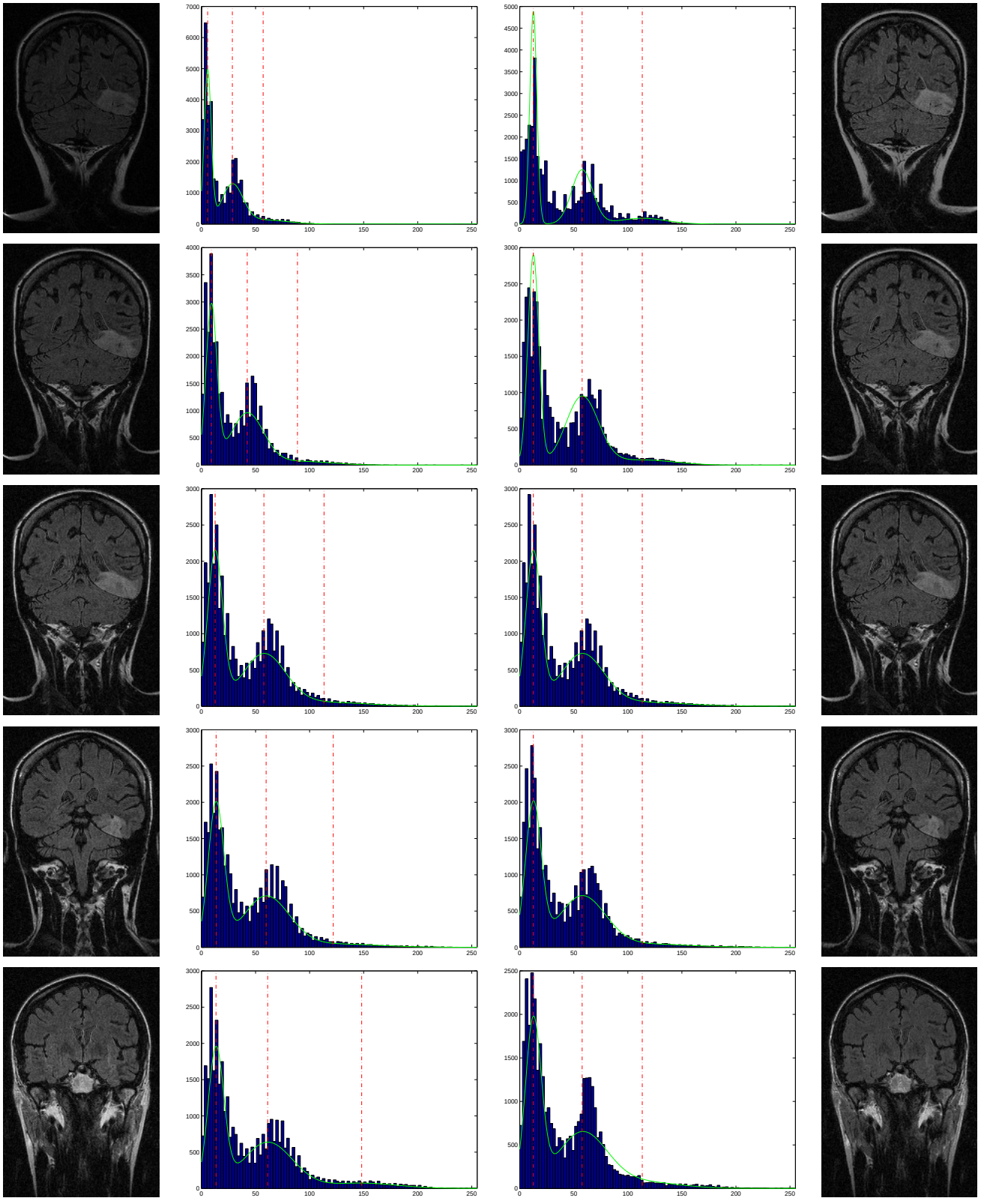
A 3D MRI is acquired as a sequence of 2D images, called slices. Unfortunately, some MR images have slices that exhibit intensity levels vastly different from neighboring slices. This is a result of the imaging process rather than what is being imaged. These intensity variations cause problems for segmentation methods that require a consistent 3D image.

Since MR images are captured in a slice-by-slice fashion, our goal is to bring the intensity values from the same population into closer alignment across slices. While there are several regions from which intensity populations are drawn, we concentrate on the three most plainly visible: black background, gray brain tissue, and white skull. The overall population image intensities may be thought of as samples from a mixture of these different distributions.

While the noise present in MRIs is Rician [14], for simplicity we assume it is additive Gaussian because the maximum-likelihood equations are straightforward and give reasonable results. Maximum-likelihood methods do exist for the Rician distribution and may be substituted in the future [34]. We use the expectation-maximization algorithm (EM) [10] to recover the mean, standard deviation, and mixing parameters for these three populations. The white pixels of the skull are actually a small percentage of the pixels and can be broadly distributed as well. As a result, EM occasionally finds modes that do not correspond exactly to the populations we are interested in because it is not a global maximization method. Care must be taken to ensure that the distribution parameters returned by EM are the "right" ones. While we have not developed a method to handle this automatically, it should not be difficult to introduce a constraint on the reasonableness of the result. Indeed, once one knows what a reasonable result is, it may be used as an initial condition for the algorithm. This is the approach we have taken and in practice few incorrect modes were returned by EM.

The distributions' means returned by EM are used as anchor points for stretching the histogram of each slice. One representative slice is chosen whose anchor points will be used as the basis for all other slices. With the minimum and maximum intensity values (i.e. 0 and 255) included as the extreme anchor points, the histograms for all slices are stretched so that the anchor points align with those of the representative slice. We use a linear adjustment with $\gamma = 1$, although perhaps some bias toward gray matter from the background may be advantageous. See Figure 3 for an example, and note that the top two images have much less contrast than the bottom three.

Modifying the images is not necessarily medically inappropriate. As physicians manually segment slices of an MR image, they themselves adjust the brightness and contrast of each slice, playing with the values and watching how the image changes before settling on values and segmenting.



(a) (b) (c) (d)

Figure 3: Intensity normalization with row three as the representative slice. (a) Original slices; (b) Slice histogram with fitted mixture models and means; (c) Stretched histograms with aligned means; (d) Resulting images.

# 3 Diffusion

Diffusion is a description of a physical process with its foundations well outside image processing. Imagine a room in which some of the air is cold and some of the air is hot. Our intuition should agree with the laws of physics, which say that by mixing together, the cold air wants to warm up and the hot air wants to cool off. The differences in air temperature cause heat to flow in manner such that the room temperature will eventually be absolutely uniform. But what if a person on one side of the room wanted it to remain warm, and someone on the other side wanted it to remain cold? They would likely erect an impediment to air flow that allows their respective sides to converge on an average temperature of their individual liking, preventing hot from mixing with cold and keeping both satisfied.

This very idea is applied to image processing. In a noisy image, there are both bright and dark areas, and we would like each of these to be relatively uniform. One way is to open the gates for intensity to flow just as air temperature does. As brightnesses blend, the visual effect is blurring. While this has the positive effect of smoothing out noise, the unconstrained flow will also mix very bright and very dark areas. When our goal is image segmentation, different regions of the image are of interest, so we'd like to keep the bright areas bright and the dark areas dark, but still remove the noise. This can be accomplished by impeding the flow of brightness across region boundaries.

As we allow intensity to flow through the image, details tend to erode. So long as the impediment that constrains intensity flow never stops it altogether, all regions will eventually blend to a uniform intensity. Imagine a hot cup of coffee on a desk inside the only heated room of an otherwise well-insulated house. While the mug keeps the coffee warm, unfortunately, no one has developed the perfect cup that keeps coffee hot indefinitely, so heat escapes little by little from the mug into the room, until everything within is the same temperature. On a larger scale, the four walls of the room slow the transfer of heat into the chilly house. But, in much the same way, the room does not stay warm forever, and eventually the whole house is the same temperature throughout, from the cold mug of coffee to the far end of the hall.

While cold coffee may not be enticing, the same effect applied to images is quite attractive. Some details of an medical image are unimportant, such as noise. Other details, such as an edema caused by stroke are highly pertinent. The problem we face is deciding ultimately how much detail is needed to correctly identify the region of tissue damaged by stroke. What is a significant change in brightness? What is a significant region size? Since there is a considerable amount of variation in lesions, the answers to these questions are not always the same. Therefore, we seek a framework that allows us to evaluate an image when the answers to questions such as these are different. The scale-space theory provides just such a framework. Consider the lesion in Figure 6. Is the difference in intensity between the area below the left-hand finger and the rest of the background significant? Is the black stripe of CSF through the lower-right corner of the lesion large enough to worry about? This section reviews the model that describes the flow of image intensity, how it is directed, and how it is constrained.

## 3.1 Theory

Scale-space theory has become more fully developed as a method of examining an image and its features at multiple scales. While proposed in the west by Witkin [52], there is evidence of formal scale-space theory decades earlier in Japan [16]. In its simplest, uncommitted form, the single-parameter family of images can be defined by convolution with a Gaussian kernel of increasing width, which is a solution to the linear diffusion equation [21]. Other scale-spaces are created by generalizing the diffusion equation [30]. These methods allow us to disambiguate some object boundaries and consider stroke lesions at different scales, in essence modeling the variability in segmentation styles and the interpretation of various tissue types damaged by stroke.

### 3.1.1 Physical Background

In a classic paper, Koenderink discusses the structure of images, especially scale and resolution, and how a family of images can be generated from a single parameter [20]. Koenderink concluded that the heat conduction equation governs the so-called deep structure of the image. Later work expounded upon the axioms, such as causality and extrema non-enhancement, that govern the scale-space of an image [2].

Weickert characterizes the physical background of diffusion as follows [46, 47]. Diffusion is based on a physical process that creates an equilibrium between differences in concentrations while conserving some quantity, such as mass or energy. In MR imagery, this quantity is image intensity. A concentration gradient causes a flux that strives to compensate for the gradient. This is captured in *Fick's Law*,

$$j = -D \cdot \nabla u,$$

where a diffusion tensor $D$, a positive definite symmetric matrix, characterizes the relationship between a gradient $\nabla u$ and the flux $j$ it induces.

The *continuity equation* reflects the preservation of mass by diffusion,

$$\partial_t u = -\text{div}\left(j\right),$$

where $t$ represents time, and divergence is given by $\nabla \cdot j$. The *diffusion equation* is the result of substituting Fick's law into the continuity equation:

$$\partial_t u = \text{div}\left(D \cdot \nabla u\right), \tag{1}$$

thus unifying the conservation of mass, temperature, intensity, etc., with the description of the quantity's flux.

It is important to distinguish several special cases of (1). When $j$ and $\nabla u$ are parallel, the equilibrium is called *isotropic*, and the diffusion tensor may be replaced by a positive scalar-valued *diffusivity*, $g$. In the general case when the flux $j$ and gradient $\nabla u$ are not parallel, the system is called *anisotropic*. When the diffusion tensor is constant over the entire image, the diffusion is said to be *homogeneous*, while space-dependent filtering is known as *inhomogeneous*. Another distinction can be made when the diffusion tensor or scalar diffusivity is a function of the evolving image, which leads to a nonlinear equation.

Nonlinear diffusion in the context of image processing gives the following model. Let a $d$-dimensional image domain be $\Omega := (0, a_1) \times \ldots \times (0, a_d)$ with boundary $\Gamma := \partial\Omega$. The filtered version $u(x,t) : \Omega \times [0,\infty) \to \mathbb{R}^d$ of an image $f(x) \in L^\infty(\Omega)$ is the solution of a diffusion equation (1) with $f$ as the initial condition and reflecting boundary conditions [42]:

$$\partial_t u = \operatorname{div}(D(\nabla u)\nabla u) \quad \text{on} \quad \Omega \times (0,\infty) \tag{2}$$

$$u(x,0) = f(x) \quad \text{on} \quad \Omega \tag{3}$$

$$\langle D(\nabla u)\nabla u, n\rangle = 0 \quad \text{on} \quad \Gamma \times (0,\infty) \tag{4}$$

where $n$ denotes the outer normal and $\langle .,. \rangle$ the inner product on $\mathbb{R}^d$. Thus (2) describes the flow of heat through a house or intensity through an image as it is controlled by $D(\nabla u)$, (3) describes the starting temperature of everything in the house or the initial image, and (4) says the house is perfectly insulated or that no intensity ever leaves an image.

Nonlinear diffusion was applied to MRI data shortly after its introduction [12]. Recent research points out the tacit assumption of Gaussian noise in standard filtering methods and the bias accompanying that assumption [35]. Our research has not yet made use of developments without the assumption.

### 3.1.2 Linear and Nonlinear Diffusion

When the diffusion is isotropic, the diffusion tensor is replaced by a positive scalar,

$$\partial_t u = \operatorname{div}(g\nabla u). \tag{5}$$

Thus, $g$ influences the magnitude of the diffusion in the direction of the gradient. In the special case where $g = c$ is a constant, (5) reduces to

$$\partial_t u = \operatorname{div}(c\nabla u) = \nabla \cdot (c\nabla u) \tag{6}$$

$$= c\nabla^2 u, \tag{7}$$

where $\nabla^2$ is the Laplacian operator. As it turns out, this is equivalent to the common Gaussian blur. The solution to the isotropic linear diffusion equation in (7) with $c = 1$ and conditions (3) and (4) is the convolution of $f$ with a Gaussian of standard deviation $\sqrt{2t}$ [22].

While isotropic homogeneous linear diffusion—Gaussian blurring—meets the criteria for a scale-space, there are both advantages and disadvantages to using it. Koenderink noted first and foremost that it makes analysis very simple (especially since a solution to the diffusion equation may be found for it), but he also remarked that isotropy, homogeneity, and linearity are by no means necessary. Furthermore, the constant, uniform diffusion gives no preference to any region of the image for any reason—everything is blurred equally. In some cases this may be an advantage, but it quickly destroys features we are interested in when attempting to segment images, especially edges and perceivable boundaries.

Two of the axioms for scale-spaces are causality (structures at coarser scales should be a result of structures from finer scales, rather than the scale-space operator) and non-enhancement of extrema (extrema monotonically converge). It is well-documented that

while Gaussian blurring obeys these axioms, the location of structures, such as extrema, are dislocated; features drift in scale-space. Thus, when an interesting feature (e.g. an edge) is found at a particular scale, its "true" location must be traced back through scale space to the finer scale, a correspondence problem that is complicated and expensive [7].

Here, we are seeking to localize object boundaries and will want to use a semantically meaningful scale-space. Perona and Malik introduced nonlinear diffusion as a way to control the process in a manner that better preserves object boundaries [30].

Perona and Malik proposed to drastically slow the rate of diffusion at object boundaries. However, we do not know where exactly object boundaries are, or our problem would already be solved. Fortunately, the gradient of the image serves as a good indicator of object boundaries. Hence, the magnitude of the gradient, which may also be thought of edge strength, is used to control the diffusivity. When feedback is introduced to the system so that diffusion is directed not by the initial image, but by the evolving image, the resulting general nonlinear isotropic diffusion equation is

$$\partial_t u = \operatorname{div}\left(g\left(|\nabla u|^2\right)\nabla u\right). \tag{8}$$

The function $g : \mathbb{R} \to (0, 1]$ controls the amount of diffusion at any point. If our goal is to remove noise but preserve significant boundaries, then $g$ should encourage blurring within regions and discourage blurring between regions. Since the image gradient magnitude is our approximate indicator of boundaries, a threshold $\lambda$ indicating how strong the gradient must be before it is considered significant should be established. Diffusivities such as

$$g\left(|\nabla u|^2\right) = \frac{1}{1 + |\nabla u|^2 / \lambda^2} \tag{9}$$

and

$$g\left(|\nabla u|^2\right) = \exp\left(-|\nabla u|^2 / \lambda^2\right) \tag{10}$$

were initially proposed [30, 31] (Figure 4(a)). The result on the filtered images will be to remove noise and small features (where $|\nabla u| < \lambda$) while leaving edges pronounced and their locations preserved.

Weickert further developed the scale-space theory of anisotropic nonlinear diffusion in [42, 43]. Additional properties of Perona and Malik's pioneering work are covered in [18, 19, 49].

### 3.1.3 Edge Enhancement

By slowing the diffusion at strong gradients, we prevent boundaries from quickly eroding. But by selecting $g$ carefully, the process can actually *enhance* edges as well as preserve them. This is accomplished by having the slope of the edge grow over time—the boundary thus becomes sharper and more distinct. When a stroke lesion in an MR image has a weak or ambiguous boundary, edge enhancement can be used to distinguish regions more precisely. Areas of tissue unaffected by stroke are blended to nearly uniform intensity via the noise removal properties described in the previous section. This section describes how lesion boundaries can be made more prominent.

The one-dimensional version of (5) is

$$\partial_t u = \partial_x \Phi\left(\partial_x u\right) = \Phi'\left(\partial_x u\right) \partial_{xx} u \tag{11}$$

where $\Phi\left(\partial_x u\right) = g\left(\partial_x u^2\right) \partial_x u$ is the flux. We want to know how the slope of an edge, $\partial_x u$ changes with time, which is described by $\partial_t\left(\partial_x u\right)$. If we reverse the order of differentiation and substitute (11), then

$$\begin{aligned} \partial_t\left(\partial_x u\right) &= \partial_x\left(\partial_t u\right) = \partial_x\left(\partial_x \Phi\left(\partial_x u\right)\right) \\ &= \Phi'' \partial_{xx} u^2 + \Phi' \partial_{xxx} u. \end{aligned}$$

Without loss of generality, suppose the edge ramps up, so $\partial_x u > 0$. The edge's inflection point $\partial_{xx} u = 0$ corresponds to the location of maximum slope, so $\partial_{xxx} u \ll 0$. Around this inflection, the change in the edge slope with time $\partial_t\left(\partial_x u\right)$ has sign opposite to $\Phi'$. Therefore, when the derivative of the flux is positive, $\Phi'\left(\partial_x u\right) > 0$, the slope of the edge is increasing. Similarly, the slope of the edge decreases over time when $\Phi'\left(\partial_x u\right) < 0$ [30]. Therefore, edge enhancement happens precisely when the edge slope is increasing with time.

The negative flux derivative giving rise to edge-enhancement causes backward diffusion via the inverse heat equation, a problem well-known to be ill-posed [19]. While discretization has certain stabilizing effects, Catté et al. formalize a solution's existence, uniqueness, and regularity by replacing the diffusivity $g\left(\left|\nabla u\right|^2\right)$ with $g\left(\left|\nabla u_\sigma\right|^2\right)$ where $u_\sigma := G_\sigma * u$, $G_\sigma$ is the Gaussian with standard deviation $\sigma$, and $*$ is the convolution operator [5]. Thus, the gradient is replaced by its estimate, and the equation (8) will diffuse only if the gradient is *estimated* to be small.

The need for regularization comes from the fact that the model allows for backward diffusion. When the flux is monotonically increasing, the diffusion always runs forward in time. While this ensures well-posedness properties, it does not allow for edge-enhancement [46]. Backward diffusion sharpens and enhances edges; to facilitate this behavior, the flux function must be decreasing (have a negative derivative) somewhere. Diffusion with nonmonotonic flux functions is ill-posed in general, but the regularization described above gives well-posedness results.

Weickert has proposed the additional diffusivity function

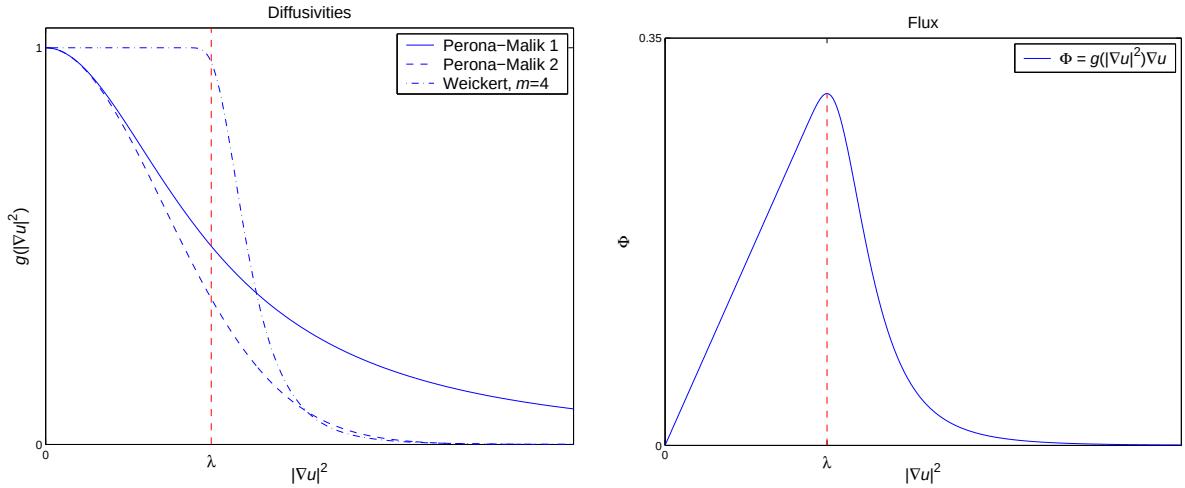
$$g\left(\left|\nabla u\right|^2\right) = 1 - \exp\left(-\frac{C_m}{\left(\left|\nabla u\right|^2 / \lambda^2\right)^m}\right), \tag{12}$$

for $m \in (0, \infty)$, where $C_m$ is a constant described below (Figure 4(a)). Like (9) and (10), the expression contains a contrast parameter $\lambda$ that will allow us to specify where blurring should occur (forward diffusion) and where edge sharpening should occur (backward diffusion).

The flux of the continuity-cum-diffusion equation (8) is

$$\Phi\left(\nabla u\right) = g\left(\left|\nabla u\right|^2\right) \nabla u.$$

As shown above, for the diffusion process to smooth unwanted features and enhance desired edges, the flux function (in a simple one-dimensional system) must satisfy $\Phi'\left(\nabla u\right) \geq 0$




(a) (b)

Figure 4: (a) Diffusivity functions; (b) Flux using equation (12).

for $|\nabla u| \leq \lambda$ and $\Phi'(\nabla u) < 0$ for $|\nabla u| > \lambda$ [30, 47]. In (12), the constant $C_m$ should be chosen accordingly. If we assume that an inflection point in the flux function occurs at $|\nabla u| = \lambda$, so that the derivative $\Phi'$ takes on the appropriate positive and negative values, we may solve for $C_m$ by setting the derivative equal to zero when $|\nabla u| = \lambda$. The result turns out to be

$$C_m = -W_{-1}\left(-\frac{1}{2m}\exp\left(-\frac{1}{2m}\right)\right) - \frac{1}{2m}$$

where $W_k(x)$ is the Lambert W function for the $k$'th branch.[1] Figure 4 illustrates the diffusivity functions (9), (10), and (12), each with the same value for $\lambda$. The Perona-Malik diffusivities (9) and (10) are very rough approximations to the type of control that may be desired. The Weickert diffusivity blurs at the maximum rate almost until the contrast parameter is reached (Figure 4). Thus in (12), $m$ may be thought of as a decay parameter—as $m$ increases, the diffusivity drops off more rapidly, and it approaches a step function.

As demonstrated in Figure 2(a), lesions often have very ambiguous boundaries. The edge-enhancement allowed by diffusion with a non-monotonic flux function, such as (12), gives rise to more easily detectable object boundaries. The result is an image containing regions of essentially piecewise constant intensity. Thus, monotonic ramps become stair-cased plateaus. For some applications this is an undesirable result, and some interesting modifications have been explored; complex diffusion processes are proposed in [13], and diffusion over the gradient of the image, rather than the image itself, is discussed in [25].

Because we are interested in detecting boundaries, it is also interesting to examine *where* these plateaus will arise. Backward diffusion occurs when the derivative of the flux is negative, $\Phi'(\nabla u) < 0$, which we have established occurs when $|\nabla u| > \lambda$. Boundaries between plateaus will become evident at the points where backward diffusion is greatest, or when the change in flux has an extremum. Thus, we seek the extrema of $\Phi'(\nabla u)$ in

[1] The Lambert W function is the inverse of the function $f(W) = We^W$. See [9].

| Diffusivity $g\left(\left\|\nabla u\right\|^2\right)$ | Maximal backward diffusion $\psi$ |
|---|---|
| $\frac{1}{1+\left\|\nabla u\right\|^2/\lambda^2}$ | $\sqrt{3}\lambda$ |
| $\exp\left(-\left\|\nabla u\right\|^2/\lambda^2\right)$ | $\frac{\sqrt{6}}{2}\lambda$ |
| $1-\exp\left(-\frac{C_m}{\left(\left\|\nabla u\right\|^2/\lambda^2\right)^m}\right)$ | $\left(\frac{C_m}{1-\frac{1}{2m}}\right)^{\frac{1}{2m}}\lambda$ |

Table 1: Maximal backward diffusion for various diffusivity functions.

regions of backward diffusion, or

$$\psi = \left\{\left|\nabla u\right| : \Phi''\left(\nabla u\right) = 0 \text{ and } \left|\nabla u\right| > \lambda\right\}$$

Table 1 gives these points of maximal backward diffusion in terms of the contrast parameter for some diffusivity functions. For diffusivities (9) and (10), there is but one point, a constant multiple of the contrast parameter $\lambda$. However, for diffusivity (12), the point in $\psi$ is also a function of $m$. If we let

$$A_m := \left(\frac{C_m}{1-\frac{1}{2m}}\right)^{\frac{1}{2m}}$$

so that $\psi = \left\{A_m\lambda\right\}$, it may be proven that

$$\begin{aligned}\lim_{m\to 0^+} A_m &= \infty \\ \lim_{m\to\infty} A_m &= 1.\end{aligned}$$

Therefore, this diffusivity allows explicit control over the gradient strength that is required for maximal edge enhancement, relative to the contrast parameter. For example, the standard value $m = 4$ gives $A_4\lambda \approx 1.1812\lambda < \frac{\sqrt{6}}{2}\lambda < \sqrt{3}\lambda$.

The implications of the relationship between $\psi$ and $\lambda$ are illustrated in Figure 5. The Gauss-like function in 5(a) is filtered using diffusivity (12) with $m = 4$. The absolute value of the gradient is shown in 5(b). Backward diffusion occurs here in regions above the bottom-most line, when $\left|\nabla u\right| > \lambda$. Figure 5(d) shows the flux and its derivative. Backward diffusion is maximal at the extrema of $\Phi'$, which are the points where $\left|\nabla u\right|$ crosses $\psi$ in 5(b). Figure 5(c) shows the filtered function at $t = 1$, with extrema of $\Phi'\left(\nabla u\left(x;\, t=1\right)\right)$ marked. Naturally, as $u$ changes with the diffusion process, so do $\nabla u$, $\Phi\left(\nabla u\right)$, and the locations of the extrema. Therefore, the stability of the locations $\left|\nabla u\right| = \psi$ through the diffusion process will determine the strength of the plateaus, or even their number. The significance of the relationship between $\psi$ and $\lambda$ should now be apparent. The contrast parameter is intended to specify what magnitude of gradient constitutes a significant edge. However, when $\psi \gg \lambda$, the points of maximal edge enhancement are likely far from the points that begin to meet the minimum criterion. As $\psi \to \lambda$, which we can accomplish through $A_m$ and $m \to \infty$, the plateaus—the very edges we use to segment an image—can be pushed as close as desired to the locations exhibiting a significant edge. Additional analysis can be found in [28, 51]

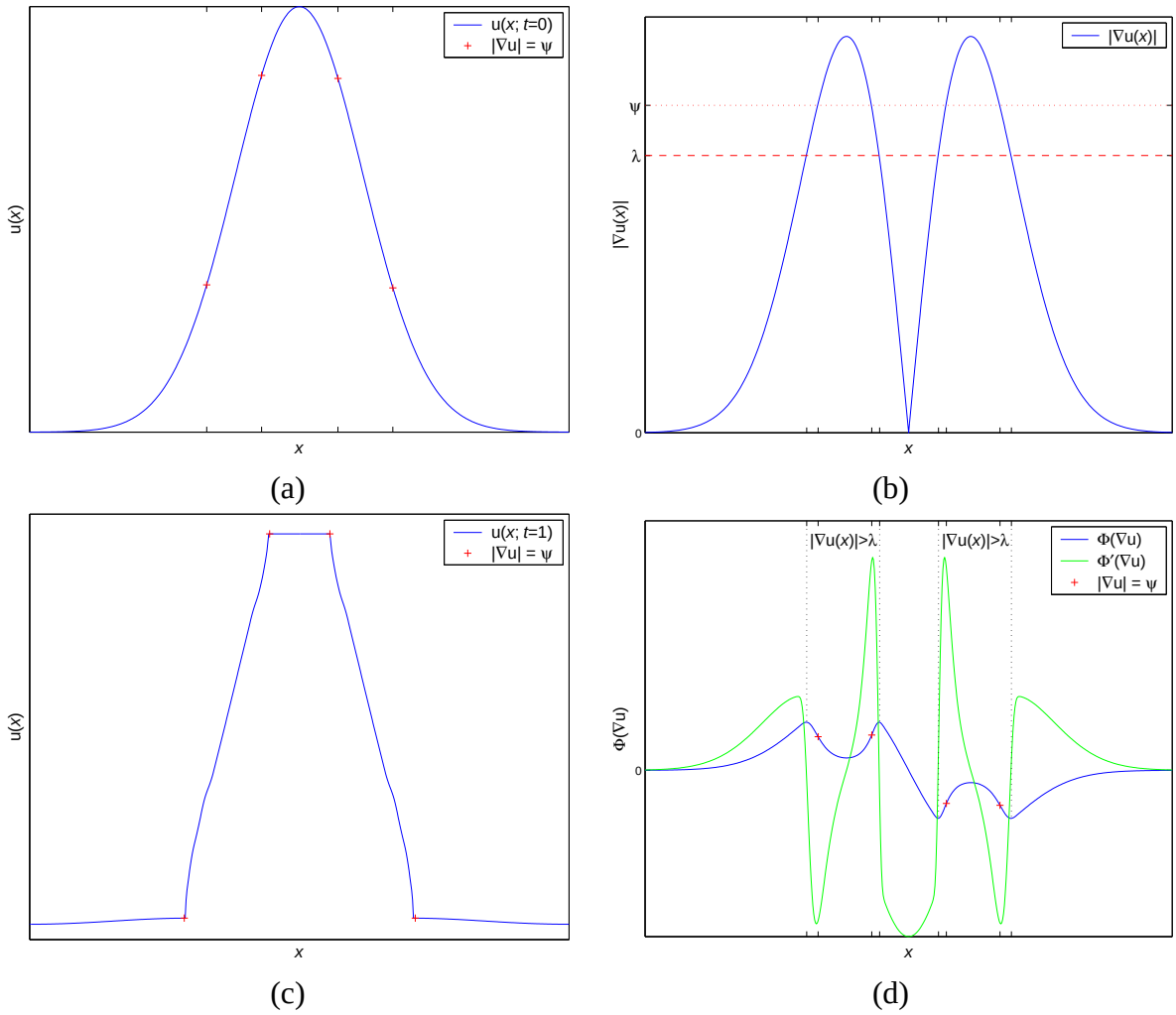



Figure 5: Edge enhancement and creation. (a) Original function $u\left(x;t=0\right)=f\left(x\right)$ with maximal backward diffusion locations; (b) $\left|\nabla u\right|$ with contrast parameter $\lambda$ and maximal backward diffusion intersections; (c) Resulting staircased image $u\left(x;t=1\right)$; (d) Flux function $\Phi\left(\nabla u\right)$ with derivative $\Phi'\left(\nabla u\right)$ and regions of backward diffusion with maximal points marked.

### 3.1.4 Isotropic and Anisotropic Diffusion

When the diffusion is isotropic, blurring nearly stops at locations with a large gradient. Thus, while this process stops diffusion across the edge (along the direction of the gradient), it does not allow smoothing along the edge, leaving edges noisy. For many image denoising and enhancement applications, this is a problem. The use of a tensor to control both the magnitude and direction of diffusion gives an extra element of control. Indeed, the diffusion can be directed purely across the gradient, so as to enhance the coherence of images, such as fingerprints [48].

Unfortunately, anisotropic diffusion can present problems when applied to objects with structures as small as those found in some stroke lesions. Edge-enhancing anisotropic nonlinear diffusion via the construction of a diffusion tensor is formulated by Weickert as follows [44]. In a $d$-dimensional system, let $v_1$, $v_i$, $i = 2 \dots d$ be the orthonormal system of eigenvectors of the diffusion tensor $D$, such that $v_1 \| \nabla u_\sigma$ and $v_i \perp \nabla u_\sigma$. The corresponding eigenvalues $\lambda_1$ and $\lambda_i$ then control the amount of smoothing across and along the edge, respectively. Weickert proposes to use a maximal blur along edges by setting these eigenvalues to be

$$\lambda_1 := g\left(|\nabla u_\sigma|^2\right) \tag{13}$$
$$\lambda_i := 1,\ i = 2, \dots, d. \tag{14}$$

Once again, $g$ is a function controlling diffusivity, but now it is only in the direction of the gradient. All other directions have maximal diffusion. This has the effect of rounding smaller structures and highly curved objects. While the effect is less adverse for larger images, it may be problematic for our application. Stroke lesions often contain delicate protrusions of abnormal tissue into healthy tissue, and these become rapidly blurred under maximal anisotropic diffusion.

One alternative would be to use a reduced diffusivity along edges, e.g. $\lambda_i = \kappa \in (0, 1)$. We may not set $\lambda_2 = \kappa = 0$ because $D$ must be positive definite. This model works poorly in practice for two reasons. First, if we set $\kappa$ near one, the rounding and blurring problems of maximal diffusivity manifest themselves. Alternatively, if $\kappa$ is set near zero, then smooth regions will experience a large diffusivity in the direction of the gradient, but only a small diffusivity in all other directions, severely reducing the advantageous blurring properties.

A slightly improved model we experimented with is:

$$\lambda_1 := g\left(|\nabla u_\sigma|^2\right) \tag{15}$$
$$\lambda_i := \max\left\{\kappa, \lambda_1\right\},\ i = 2 \dots d \tag{16}$$

where $\kappa \in (0, 1]$. Obviously, when $\kappa = 1$, the model is equivalent to Weickert's. As $\kappa \to 0$, the amount of blurring along edges decreases, however, we are assured that diffusion perpendicular to the gradient is never less than that parallel to the gradient. This means that even with a conservative, small value for $\kappa$, a smooth region will give rise to a large diffusivity in both directions, but strong gradients will give rise to a sharply reduced diffusivity across edges and a moderate diffusivity along them. The diffusivity function must always be positive and setting $\kappa = 0$ would have the effect of setting $\lambda_2 = \lambda_1$, which would yield isotropic diffusion. We may now control the amount of diffusion along edges, and

setting $\kappa \ll 1$ helps preserve the finer lesion structures and reduce the rounding effect of anisotropic diffusion. Unfortunately, the directional diffusivity (16) violates the assumptions required for the proof of scale-space properties of nonlinear anisotropic diffusion, namely that $D$ is Lipschitz continuous and $\lambda_i \in C^\infty\left([0,\infty);(0,1]\right)$ [42].

### 3.1.5 Robust Estimation

Black et al. have proven that isotropic diffusion is equivalent to a robust estimation procedure that estimates a piecewise smooth image. They illustrated that diffusion is the gradient descent of an estimation problem with a robust error norm [4]. Where as diffusion with diffusivity functions that never drop to zero ensures that the image approaches uniformity, the results on robust estimation allow for converting different error norms into diffusivity functions that eventually halt, always preserving significant edges, just like the perfect heat-trapping coffee mug.

In earlier work on this problem, a portion of the segmentation was due to hypothesis testing, which was part of the so-called packing method [23, 39]. To determine whether a set of pixels was completely within a lesion, the hypothesis test involved comparing the mean intensity of the pixels under consideration to the mean intensity of already accepted lesion pixels. This assumed that the complete lesion comes from a unimodal (Gaussian) distribution, which is not necessarily the case (cf. §2). The approach we take is to use nonlinear diffusion to restore the noisy image to piecewise constant regions. Since diffusion is equivalent to robust estimation, the resulting image is built upon solid theoretical foundation, just as the packing method is, but allows for lesions with more than one mode.

## 3.2 Implementation

The transition from continuous theory to discrete implementation is documented in [43, 47]. In the fully discrete (time and space) model, the $d$-dimensional image of size $N_1 \times \ldots \times N_d$ with $N = \prod_{l=1}^{d} N_l$ total pixels is modeled as a vector $f \in \mathbb{R}^N$ with components $f_i$, $i = 1 \ldots N$. The index $i$ represents some pixel $x_i$ and $f_i$ is the intensity at the pixel. Time is discrete with $t_k := k\tau$, where $k \in \mathbb{N}_0$ and $\tau$ is the step size. The approximation of $u\left(x_i, t_k\right)$ is denoted $u_i^k$. The discrete model is then summarized by

$$u^0 = f \tag{17}$$

$$u^{k+1} = Q\left(u^k\right) u^k, \, \forall k \in \mathbb{N}_0, \tag{18}$$

where $Q$ is a matrix describing the update of $u$ from step $k$ to $k+1$. The value for $u^{k+1}$ may be arrived at by an explicit, semi-implicit, or even fully implicit scheme. More precisely, the explicit scheme is a forward Euler explicit numerical scheme. To see this, we begin to transform the continuous equation (2) into its discrete counterpart:

$$\partial_t u = \sum_{i,j=1}^{d} \partial_{x_i}\left(D_{ij}\left(\left|\nabla u_\sigma\right|^2\right)\partial_{x_j} u\right). \tag{19}$$

For example, in two dimensions, we have

$$\nabla \cdot (D\nabla u) = \partial_{x_1}\left(D_{11}\partial_{x_1} u + D_{12}\partial_{x_2} u\right) + \partial_{x_2}\left(D_{21}\partial_{x_1} u + D_{22}\partial_{x_2} u\right).$$

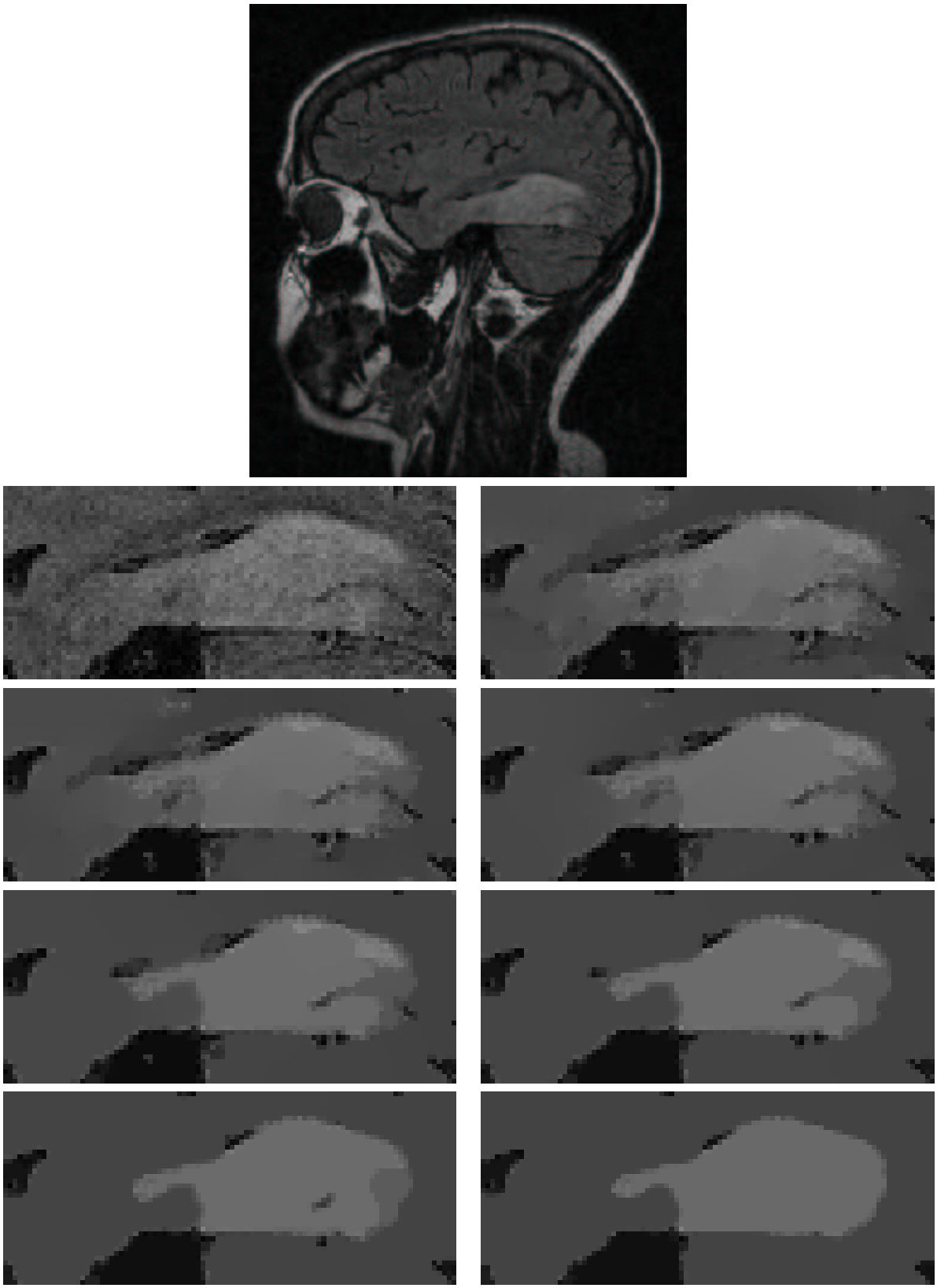


Figure 6: Evolution of a lesion in an MRI slice (image 3P) under 3D isotropic diffusion $(m = 4,\ \sigma = 1,\ \lambda = 2.55,\ \tau = 2.5)$. Left to right, top to bottom: $t = 0,\ 37,\ 81,\ 181,\ 602,\ 2441,\ 5432,\ 26903$.

In the isotropic case, there are no mixed derivatives, so we may simplify (19) to a single-index summation

$$\partial_t u = \sum_{l=1}^{d} \partial_{x_l} \left( g\left( \left| \nabla u_\sigma \right|^2 \right) \partial_{x_l} u \right). \tag{20}$$

### 3.2.1 Explicit Schemes

We can approximate the effect of the partial derivative operator $\partial_{x_l}$ on some function $f$ at each location $i$ with the central difference $\frac{1}{h_l} \sum_{j \in \mathcal{N}_l(i)} (f_j - f_i)$, where $\mathcal{N}_l(i)$ is the set of neighbors to pixel $i$ along direction $l$ (border pixels have but one neighbor), and $h_l$ is the resolution along the $l$ dimension. Using this as the approximation to the spatial derivative and a backward difference to approximate the time derivative, a simple discretization of (20) takes the form

$$\frac{u_i^{k+1} - u_i^k}{\tau} = \sum_{l=1}^{d} \sum_{j \in \mathcal{N}_l(i)} \frac{g_j^k + g_i^k}{2h_l^2} \left( u_j^k - u_i^k \right),$$

where $g_i$ is an approximation to $g\left( \left| \nabla u_\sigma \right|^2 \right)$ at that pixel. This can be written in vector-matrix notation as [45]

$$\frac{u^{k+1} - u^k}{\tau} = \sum_{l=1}^{d} A_l\left(u^k\right) u^k. \tag{21}$$

Isolating $u^{k+1}$ gives

$$u^{k+1} = \left( I + \tau \sum_{l=1}^{d} A_l\left(u^k\right) \right) u^k. \tag{22}$$

This scheme is explicit because it allows for calculating $u^{k+1}$ from $u^k$ directly and without any matrix inversions. Thus the explicit scheme for the model (18) is

$$Q\left(u^k\right) = \left( I + \tau \sum_{l=1}^{d} A_l\left(u^k\right) \right).$$

Since no matrix inversions are necessary, each step may be performed quickly. However, in order to guarantee stability, the step size must be small [50]:

$$\tau \leq \left( \sum_{l=1}^{d} \frac{2}{h_l^2} \right)^{-1}.$$

A typical MRI has voxel dimensions $1 \times 1 \times 2.5$ mm. To perform explicit diffusion in three dimensions with this method would require a step size no larger than about 0.23.

A slightly different (perhaps even more straightforward) approach to an explicit scheme in two dimensions is detailed in [41]. However, the approach of this section is preparation for a review of the more efficient additive operator splitting schemes.

### 3.2.2 Additive Operator Schemes

Because it requires many iterations, the step size of the explicit scheme is highly prohibitive in practice. However, we may consider the slightly more complicated scheme

$$\frac{u^{k+1}-u^k}{\tau} = \sum_{l=1}^{d} A_l\left(u^k\right) u^{k+1} \tag{23}$$

which gives the iteration step

$$u^{k+1} = \left(I - \tau \sum_{l=1}^{d} A_l\left(u^k\right)\right)^{-1} u^k. \tag{24}$$

This scheme does not give $u^{k+1}$ directly, rather, it requires a matrix inversion and is thus semi-implicit. Unfortunately, the matrix $B := I - \tau \sum_{l=1}^{d} A_l\left(u^k\right)$ does not lend itself to efficient inversion [50]. Fortunately, this limitation may be overcome by modifying the semi-implicit scheme, resulting in the additive operator splitting (AOS)

$$u^{k+1} = \frac{1}{d}\sum_{l=1}^{d}\left(I - d\tau A_l\left(u^k\right)\right)^{-1} u^k. \tag{25}$$

This discretization has several benefits. First, the first-order Taylor expansion of (25) is the same for the explicit scheme (22) and the semi-implicit scheme (24), so they are consistent approximations to the continuous equation. Second, unlike multiplicative splittings, all dimensions are treated equally. Third, it meets discrete scale space properties, such as convergence to the mean intensity as a steady state and the extremum principle (causality). Most importantly, it is efficient. While it requires about twice the effort of a typical explicit scheme, this modification of a semi-implicit scheme guarantees stability with no limit on step size. The only remaining consideration is a trade-off between accuracy and speed.

With $B_l\left(u^k\right) := I - d\tau A_l\left(u^k\right)$, the iterative step in (18) has

$$Q\left(u^k\right) := \frac{1}{d}\sum_{l=1}^{d} B_l^{-1}.$$

The discrete scale-space properties of this scheme are proven in [45]. If we let $w_l := B_l^{-1}u^k$, we see that we need to solve equations of the form $B_l w_l = u^k$ so that we may calculate (25) as

$$u^{k+1} = \frac{1}{d}\sum_{l=1}^{d} w_l. \tag{26}$$

Each diffusion direction[2] $l$ has a matrix $A\left(u^k\right) = \left(a_{ij}\left(u^k\right)\right)$, with

$$a_{ij}\left(u^k\right) := \begin{cases} \frac{g_i^k + g_j^k}{2h_l^2} & j \in \mathcal{N}_l(i) \\ -\sum_{n\in\mathcal{N}_l(i)} \frac{g_i^k + g_n^k}{2h_l^2} & j = i \\ 0 & \text{otherwise.} \end{cases}$$

[2]We temporarily abstain from specifying the diffusion direction with notation such as $a_{l_{ij}}\left(u^k\right)$ to avoid notational baggage.

Therefore the entries of $B\left(u^k\right) = \left(b_{ij}\left(u^k\right)\right)$ are

$$b_{ij}\left(u^k\right) := \begin{cases} \gamma_{ij} & j \in \mathcal{N}_l(i) \\ \alpha_i & j = i \\ 0 & \text{otherwise.} \end{cases}$$

where

$$\alpha_i := 1 + d\tau \sum_{n \in \mathcal{N}_l(i)} \frac{g_i^k + g_n^k}{2h_l^2} \tag{27}$$

$$\gamma_{ij} := -d\tau \frac{g_i^k + g_j^k}{2h_l^2} \tag{28}$$

The tridiagonal and diagonally dominant matrix $B$ takes the form

$$B\left(u^k\right) = \begin{bmatrix} \alpha_1 & \gamma_{12} & & & \\ \gamma_{21} & \alpha_2 & \gamma_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{N-2\,N-1} & \alpha_{N-1} & \gamma_{N-1\,N} \\ & & & \gamma_{N\,N-1} & \alpha_N \end{bmatrix},$$

where the unmarked entries are all 0. But we observe from (28) that $\gamma_{ij} = \gamma_{ji}$, so may we reduce storage overhead by using a single index for $\gamma$:

$$B\left(u^k\right) = \begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ \gamma_1 & \alpha_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{N-2} & \alpha_{N-1} & \gamma_{N-1} \\ & & & \gamma_N & \alpha_N \end{bmatrix}. \tag{29}$$

The most efficient way of solving an equation such as $B_l w_l = u^k$ with system matrix (29) is the Thomas algorithm. It involves three basic steps: an LR decomposition, a forward substitution, and a backward substitution. These details are covered very well in [50].

Some further implementation details are worth mentioning. It is not necessary to convert an image stored as $d$-dimensional arrays to a vector. Consider a regular two-dimensional image having $N_1$ rows and $N_2$ columns. Rather than constructing a single vector $u \in \mathbb{R}^{N_1 N_2}$ with a new pixel ordering, diffusion along the row dimension would actually be implemented as $N_2$ diffusion processes—one for each column. For a three-dimensional MR image, there would be a diffusion process for each column *and* slice. MATLAB's built-in component-wise operations are a simple way to do this. One only needs to explicitly perform diffusion along each dimension (rows, columns, slices), allowing MATLAB to implicitly handle the fact that there are many parallel diffusion processes occurring. Proceeding in this manner, the $N/N_l$ equations for each dimension $l$ are solved in one fell swoop (see Appendix).

Unfortunately, the potential of AOS for performing anisotropic diffusion is limited. The number of directional diffusivities jumps from $d$ in the isotropic case to $d^2$ in the anisotropic case. The simplest discretization of (19) is

$$\frac{u^{k+1}-u^k}{\tau} = \sum_{i,j=1}^{d} A_{ij}\left(u^k\right) u^k.$$

Once again, $A_{ij}$ represents a central difference approximation to $\partial_{x_i}\left(D_{ij}\left(\left|\nabla u_\sigma\right|^2\right)\partial_{x_j} u\right)$. The explicit scheme for this discretization is given by

$$u^{k+1} = \left(I + \tau \sum_{i,j=1}^{d} A_{ij}\left(u^k\right)\right) u^k.$$

As mentioned before, this scheme requires small time steps to remain stable. Additionally, an AOS scheme such as

$$u^{k+1} = \frac{1}{d^2}\left(I - d^2\tau \sum_{i,j=1}^{d} A_{ij}\left(u^k\right)\right)^{-1} u^k$$

is problematic because the directional diffusivities $A_{ij}$ must be non-negative [26, 47]. Therefore, an AOS-stabilized scheme

$$u^{k+1} = \frac{1}{d}\sum_{l=1}^{d}\left(I - d\tau A_{ll}\left(u^k\right)\right)^{-1}\left(I + \tau\sum_{i=1}^{d}\sum_{j\neq i} A_{ij}\left(u^k\right)\right) u^k$$

was proposed [48]. Since they are potentially negative, the off-diagonal diffusivities are calculated via an explicit scheme, giving $v^k = \left(I + \tau\sum_{i=1}^{d}\sum_{j\neq i} A_{ij}\left(u^k\right)\right) u^k$. The remaining non-negative matrices $B_l\left(u^k\right) := I - d\tau A_{ll}\left(u^k\right)$ that describe the diffusion caused by diagonal entries of the diffusion tensor each give a system $B_l w_l = v^k$. The result $u^{k+1}$ is then the same as (26) and is also solved for with the Thomas algorithm. A step size of $\tau = 2$ is commonly used.

# 4 Deformable Models

In §3 we described a procedure that allows one to examine MR images at different levels of scale. We now approach the problem of segmenting such images. The Level Set Method (LSM), a common technique used for image segmentation, was developed by Osher and Sethian in 1988 [29] and was subsequently refined to a more efficient implementation, the Fast Marching Method (FMM) [1]. We use FMM for image segmentation. Unlike so-called "snakes," which use an explicit, parametric representation of the contour [17], FMM uses an implicit representation where the contour deformation is based on geometric measures that allow it to segment highly irregular structures. Such structures are common in lesions and are often missed in segmentations performed by snakes due to a smoothing caused by regularization.

## 4.1 Theory

### 4.1.1 Parametric Deformable Models

Previous work used parametric deformable models, also known as snakes, to segment lesions. An initial contour $\upsilon$ was drawn roughly around the lesion and allowed to deform to a local energy minimum that closely matches the lesion boundary. The contour is represented explicitly as a parametric curve defined in the 2D image plane. The energy of curve $\upsilon$ parameterized by $s \in [0, 1]$ is given by

$$E_{total} = \int E_{internal} + \int E_{external} \tag{30}$$

where

$$E_{internal} = \alpha \left| \frac{d\upsilon}{ds} \right|^2 + \beta \left| \frac{d\upsilon^2}{d^2 s} \right|^2 \tag{31}$$

and $E_{external}$ is a function of the image gradient and an additional statistical "force" [23]. By manipulating the constants $\alpha$ and $\beta$ in $E_{internal}$, one can control the elasticity and rigidity of the curve, respectively. The contour $\upsilon$ is represented numerically as a spline, so it must remain simply connected to avoid costly re-parameterizations. Toward this end, $\alpha$ and $\beta$ are given relatively large values to penalize irregularity. Unfortunately, regularity constraints keep the contour from filling the intricate lesion structures mentioned in §1.2.3.

### 4.1.2 The Level Set Method

In contrast with snakes, LSM does not represent the segmenting contour explicitly. Instead, it is embedded within a higher dimensional function, $\Phi(t)$. Images of arbitrary dimension may be segmented by LSM, so we consider a higher dimensional analog to the contour: a surface or front. Assume that we are given an initial front either inside or outside the lesion in question. If $\mathbb{R}^d$ is set of points in the image hyper-plane, let $\gamma(t) \subseteq \mathbb{R}^d$ be a family of surfaces in the image found by evolving the initial front $\gamma(0)$ with speed $F$ in the direction outward normal to $\Phi$. Like the external energy of snakes, $F$ can be a function of global, local, or independent properties. LSM defines $\Phi$ as a surface so that for each point, $\mathbf{x} \in \mathbb{R}^d$, $\Phi(\mathbf{x}; t)$ is the distance to the evolving front $\gamma(t)$. Hence, for any given time $t$

$$\Phi(\gamma(t); t) = 0, \tag{32}$$

and

$$\Phi(\mathbf{x}; t) = c,$$

where $c$ is the signed distance from $\mathbf{x}$ to the closest point on the front $\gamma(t)$ (points inside the front are negative). Equation (32) tells us that the segmenting surface can be found at any time $t$ by identifying the zero level set of the function $\Phi(\gamma(t); t)$.

Given a speed function $F(\mathbf{x})$ and initial front $\gamma(0)$, the next step is to devise an update rule, or evolution equation, for $\Phi$ . We follow the formulation used in [24]. Consider the

path $\mathbf{x}(t), t \in [0, \infty)$ of a particle on the segmenting surface $\mathbf{x}(t) \in \gamma(t)$. It moves in the direction of the surface outward normal with a prescribed speed:

$$\frac{\partial \mathbf{x}}{\partial t}(t) = F(\mathbf{x}(t)) \frac{\nabla \Phi}{|\nabla \Phi|}.$$

Since the point $\mathbf{x}$ is always on $\gamma$ and $\Phi(\gamma(t); t) = 0$ for all $t \geq 0$, we know that

$$\Phi(\mathbf{x}(t), t) = 0.$$

By the chain rule,

$$\frac{\partial \Phi}{\partial t} + \sum_{i=1}^{d} \frac{\partial \Phi}{\partial x_i} \frac{\partial x_i}{\partial t} = 0 \tag{33}$$

where,

$$\begin{aligned}
\sum_{i=1}^{d} \frac{\partial \Phi}{\partial x_i} \frac{\partial x_i}{\partial t} &= \left\langle \frac{\partial \Phi}{\partial x_1}, \frac{\partial \Phi}{\partial x_2}, \ldots, \frac{\partial \Phi}{\partial x_d} \right\rangle \cdot \left\langle \frac{\partial x_1}{\partial t}, \frac{\partial x_2}{\partial t}, \ldots, \frac{\partial x_d}{\partial t} \right\rangle \\
&= \nabla \Phi \cdot \frac{\partial \mathbf{x}(t)}{\partial t} \\
&= \nabla \Phi \cdot \left( F(\mathbf{x}(t)) \frac{\nabla \Phi}{|\nabla \Phi|} \right) \\
&= \frac{F(\mathbf{x}(t))}{|\nabla \Phi|} (\nabla \Phi) \cdot (\nabla \Phi) \\
&= \frac{F(\mathbf{x}(t))}{|\nabla \Phi|} |\nabla \Phi|^2 \\
&= F(\mathbf{x}(t)) |\nabla \Phi|. \qquad (34)
\end{aligned}$$

Substituting (34) into (33) gives the evolution equation for $\Phi$:

$$\frac{\partial \Phi}{\partial t} + F |\nabla \Phi| = 0. \tag{35}$$

### 4.1.3 The Fast Marching Method

The representation of the implicit front $\gamma$ above can be discretized directly to form an algorithm for segmenting images. However, a faster algorithm can be formulated by embedding $\gamma$ in a conceptually different function called $T$, rather than $\Phi$ [33]. $T(\mathbf{x})$ gives the arrival time of the front at some point $\mathbf{x} \in \mathbb{R}^d$ in the image. Again, we assume the existence of a function $F(\mathbf{x})$ that gives the speed of a point $\mathbf{x}$ in the outward normal direction. Here we will also require that $F(\mathbf{x})$ be strictly positive. In the last section we formulated an evolution equation to construct $\Phi$, here we will take advantage of the restricted speed function to formulate a boundary value problem to construct $T$.

The key to the boundary value representation of surface evolution is that the gradient magnitude of $T$ is inversely proportional to the speed of the surface at all points:

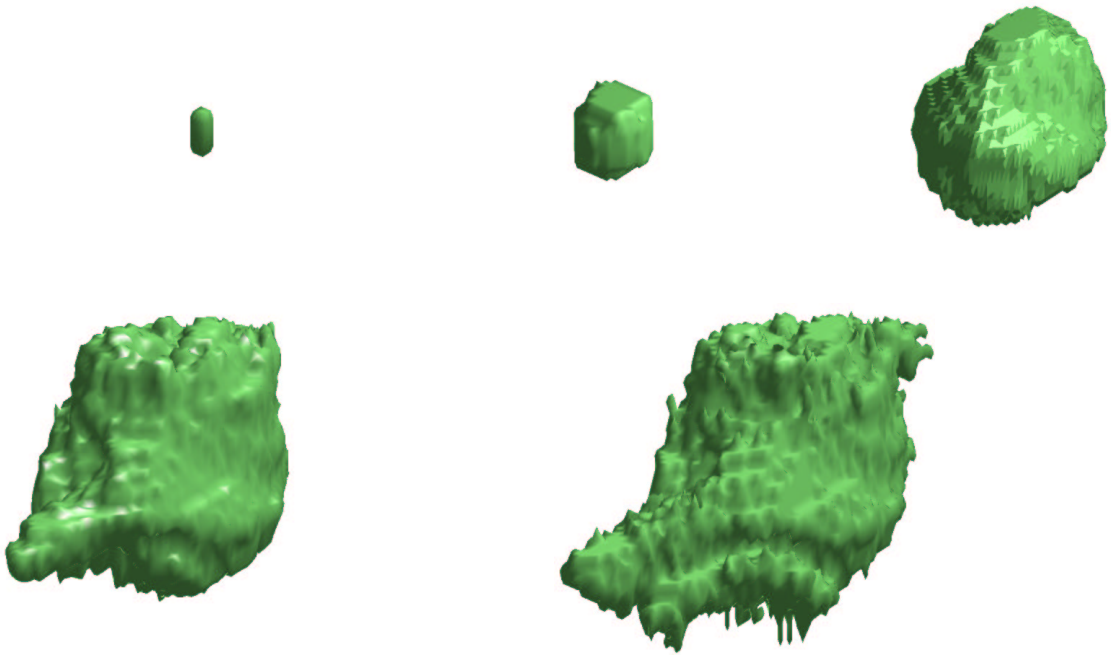


Figure 7: An evolving front inflates to lesion boundaries.

$$|\nabla T| F = 1. \tag{36}$$

FMM is developed in the following section as a numerical scheme for systematically constructing the arrival time function $T$. The boundary condition completing the problem is

$$T(\gamma(0)) = 0, \tag{37}$$

which asserts that the arrival time of the front is zero for any point on the front. It is important to note that a strictly positive speed function is needed to solve the boundary value problem (36). The restricted speed function enforces the idea that the segmenting front should begin at exactly the points specified by the boundary condition (37). Figure 7 shows a front evolving from an initial seed point, and inflating until slowed drastically by the speed function at lesion boundaries. Since we have a strictly positive speed function, the segmenting contour will always have a real arrival time at all points in the image. In practice the actual segmentation will correspond to the set of pixels at which the segmenting surface's arrival time is within a reasonable threshold.

## 4.2 Implementation

So far we have been discussing image segmentation in a somewhat abstract manner. We have assumed that the image hyper-plane is continuous and that the methods of segmentation (snakes, LSM, and FMM) operate on this continuous plane, which is not the case in

reality. In this section we discuss implementation details for discrete images.

Define a discrete image hyper-plane $\Omega \subseteq \mathbb{N}^3$ to be $\Omega := [0, 1, \ldots, a_1] \times [0, 1, \ldots, a_2] \times [0, 1, \ldots, a_3]$, where $a_i$ is the maximum image index value in the dimension $i$. The initial surface $\gamma(0)$ is given by the set of voxels $\Gamma \subseteq \Omega$. A point in the image plane is assumed to be a single voxel indexed by three natural numbers, for instance $\langle i, j, k \rangle \in \Omega$. Often it will be convenient to refer to such points as the vector, $\mathbf{v}$, $\mathbf{u}$ or $\mathbf{w}$. Further notational issues are addressed in context.

### 4.2.1 Implementation of FMM

The basic idea behind FMM is to construct the arrival time function $T$ outward from the initial surface, $\Gamma_0$. The construction process will involve updating points with increasing arrival time until a predefined time threshold is reached. We use a single initial point $\hat{x}$, so that $\Gamma_0 = \{\hat{x}\}$. As the algorithm runs, three disjoint sets of voxels are maintained: accepted ($A$), candidate ($C$), and faraway ($W$) points, where $A \cup C \cup W = \Omega$. When the algorithm terminates, the image segmentation corresponds to the set of accepted voxels $A$.

**Algorithm 1** The Fast Marching Method.

Initialization:

(a) $A \leftarrow \Gamma_0$
$T_{\mathbf{v}} \leftarrow 0 \, \forall \mathbf{v} \in A$

(b) $C \leftarrow \bigcup_{\mathbf{v} \in A} \mathcal{N}(\mathbf{v})$
$T_{\mathbf{v}} \leftarrow \frac{\triangle}{F_{\mathbf{v}}} \forall \mathbf{v} \in C$ where $\triangle$ is the voxel distance from the closest element in $A$.

(c) $W \leftarrow \Omega \setminus (A \cup C)$
$T_{\mathbf{v}} \leftarrow \infty \, \forall \mathbf{v}$

Iteration:

(a) $\mathbf{u} \leftarrow \arg\min_{\mathbf{v} \in C} T_{\mathbf{v}}$

(b) $A \leftarrow A \cup \{\mathbf{u}\}$

(c) $C \leftarrow \mathcal{N}(\mathbf{u}) \cup (C \setminus \{\mathbf{u}\})$
$T_{\mathbf{v}} \leftarrow U_{\mathbf{v}} \, \forall \mathbf{v} \in \mathcal{N}(\mathbf{u})$

(d) The algorithm terminates when the difference between the arrival time at the previous point $\mathbf{u}$ and the current point $\mathbf{u}$ is greater than a predefined threshold.

Algorithm 1 follows the description in [33]. At each iteration, we choose from the candidate set the voxel $\mathbf{u}$ with the smallest arrival time . Let $\mathcal{N} : \Omega \rightarrow \wp(\Omega)$ be the function returning the set of a voxel's neighbors. The value of $U$ (introduced on the last

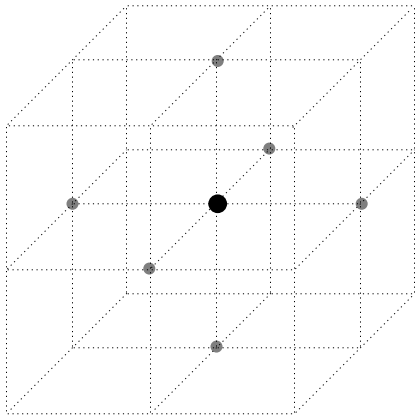


Figure 8: FMM update neighborhood.

line of algorithm 1) is at the heart of the algorithm because it defines the direction of the propagating front. It is given by the numerical approximation of (36)

$$\left[\begin{array}{c} \max\left(D_{\mathbf{v}}^{-\triangle i}T_{\mathbf{v}},0\right)^2+\min\left(D_{\mathbf{v}}^{+\triangle i}T_{\mathbf{v}},0\right)^2 \\ +\max\left(D_{\mathbf{v}}^{-\triangle j}T_{\mathbf{v}},0\right)^2+\min\left(D_{\mathbf{v}}^{+\triangle j}T_{\mathbf{v}},0\right)^2 \\ +\max\left(D_{\mathbf{v}}^{-\triangle k}T_{\mathbf{v}},0\right)^2+\min\left(D_{\mathbf{v}}^{+\triangle k}T_{\mathbf{v}},0\right)^2 \end{array}\right]^{\frac{1}{2}}=\frac{1}{F\left(\mathbf{v}\right)} \tag{38}$$

where

$$D_{\mathbf{v}}^{-\triangle i}T_{\mathbf{v}}\equiv\frac{T_{i,j,k}-T_{i-1,j,k}}{\triangle i}$$

is the backward difference approximation and

$$D_{\mathbf{v}}^{+\triangle i}T_{\mathbf{v}}\equiv\frac{T_{i+1,j,k}-T_{i,j,k}}{\triangle i}$$

is the forward difference approximation to the derivative of $T_{\mathbf{v}}$ with respect to $i$.

During each iteration, (38) is solved for $T_{\mathbf{v}}$ using values of $T$ from the previous iteration at all voxels neighboring $\mathbf{v}$. Figure 8, a visual representation of the update process, shows $\mathbf{v}$ as the large black point in the center of the voxel neighborhood. Each smaller gray point represents one of the voxels that is used to update the black one. The update process works by using (38) to find the largest solution for $T$ at the black point, using the values at the gray points from the previous update.

As a final note, FMM is a fast algorithm in part because all candidate points are stored in a min-heap. This structure keeps the elements with the smallest value (in this case smallest anticipated arrival time) at the top of the heap. Updating the heap with new candidate points is a logarithmic time task.

### 4.2.2 The Speed Function

In section 4.1.3 we discussed how FMM uses a speed function to construct the arrival time of each point on the segmenting surface, and thus the front itself. In this section we define our speed function.

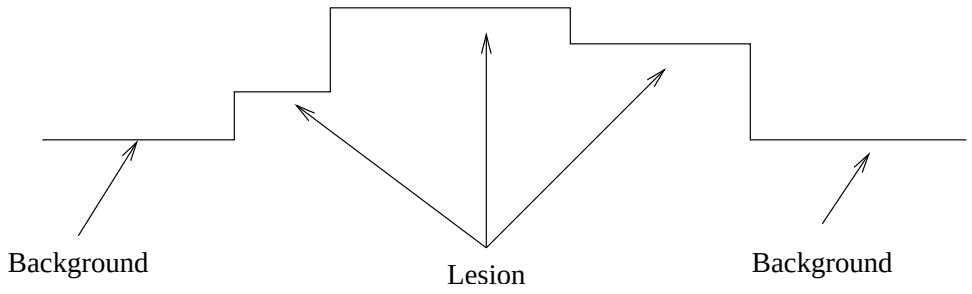



Figure 9: Example intensity profile of a multi-modal lesion.

Section 1.2.4 describes how the intensity of lesions can be multi-modal. Some of these lesion regions might have an intensity that is closer to the background than the rest of the lesion, as the leftmost lesion plateau in Figure 9. In that case, diffusion will never be able to blend all the lesion pixels to uniform intensity because pixels from the fringe mode will blend with the background first. By itself, gradient information for detecting boundaries is not enough to allow the front to envelop the entire lesion.

We also know that lesion intensity is higher than background intensity, but when ambiguous boundaries are present there is never a strong gradient to indicate any type of boundary. Therefore intensity information must be incorporated as a stopping criterion as well.

Because $F$ determines the speed of the segmenting front at every image location, it makes sense to assign $F$ a large value at voxels believed to be part of the lesion and a small value at those believed to be part of the background. Image intensity and gradient information are the the simplest, reliable clues for determining whether a voxel is part of the lesion or not. Segmentation proceeds under the assumption that lesion voxels have higher intensity than non-lesion voxels. Therefore, we will introduce a threshold on voxel intensity that will slow the front down in darker, non-lesion areas. In contrast to simple intensity values, gradient magnitudes are often used to detect borders. Large gradients are indicative of boundaries and are also good clues for where to stop the front. We combine this information into a speed function that is a hybrid of region (intensity) and boundary (gradient) based segmentation strategies.

When we segment an image $u$ that has been diffused to a particular scale, $s$, we assume we know something about the intensity of the lesion and the intensity of the non-lesion background. Therefore, we conclude there is some intensity $\beta$ for which the voxels at least that bright are lesion, and voxels darker than it are background. Thus, we use a speed function that slows the front down when image intensities are below the threshold $\beta$ and encourage propagation otherwise:

$$H\left(\mathbf{v}\right) = \exp\left(-\left(\frac{\beta}{u\left(\mathbf{v};s\right)}\right)^{a}\right). \tag{39}$$

The parameter $a$ controls the slope of this speed function about the inflection point $\beta$. Thus higher values of $a$ cause a steeper dropoffdrop-off in speed at $\beta$, indicating our confidence in that value as a threshold. Figure 10(a) shows this intensity based speed function for

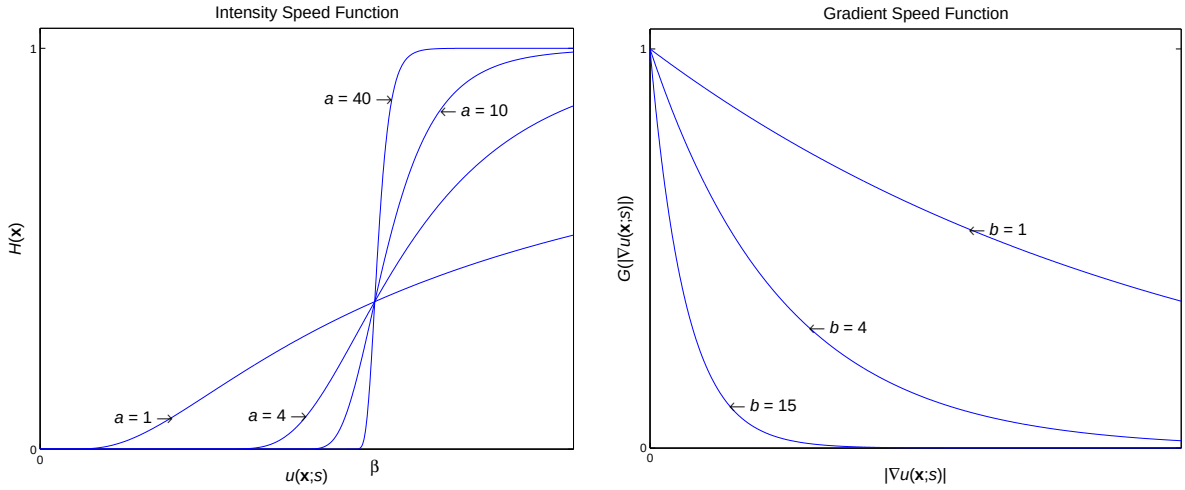



Figure 10: Intensity (left) and gradient magnitude (right) components of the final speed function.

various values of $a$.

The gradient of the filtered image allows us to search for boundaries between regions of different overall intensity. The 3D gradient is estimated by convolving an image with the separable filters

$$\begin{aligned} f_1 &= \begin{bmatrix} 1 & 3 & 1 \end{bmatrix} \\ f_2 &= \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \\ f_3 &= \begin{bmatrix} 1 & 3 & 1 \end{bmatrix} \end{aligned}$$

These are modeled after an optimally rotationally invariant 2D filter in [32]. The filters are combined as follows to produce $\nabla u = \langle u_i, u_j, u_k \rangle$, where

$$\begin{aligned} u_i &= \frac{1}{50 \triangle i} f_2^{\rightarrow} \star f_1^{\uparrow} \star f_3^{\nearrow} \star u \\ u_j &= \frac{1}{50 \triangle j} f_1^{\rightarrow} \star f_2^{\uparrow} \star f_3^{\nearrow} \star u \\ u_k &= \frac{1}{50 \triangle k} f_1^{\rightarrow} \star f_3^{\uparrow} \star f_2^{\nearrow} \star u. \end{aligned}$$

Each $u_w$ is the approximation of the partial derivative of the image with respect to direction $w$. The orientation of each convolution kernel is specified by a superscript arrow, with $f^{\rightarrow}$a filter along $i$, $f^{\uparrow}$a filter along $j$ and $f^{\nearrow}$ a filter along $k$.

To reduce the front's speed at a strong gradient we use

$$G(\mathbf{v}) = \exp\left(-b \left|\nabla u(\mathbf{v}; s)\right|\right), \tag{40}$$

where $b$ is a parameter controlling the sensitivity to gradient magnitude of the image at scale $s$. Figure 10(b) plots $G$ over the domain of gradient magnitudes for various values of $b$.

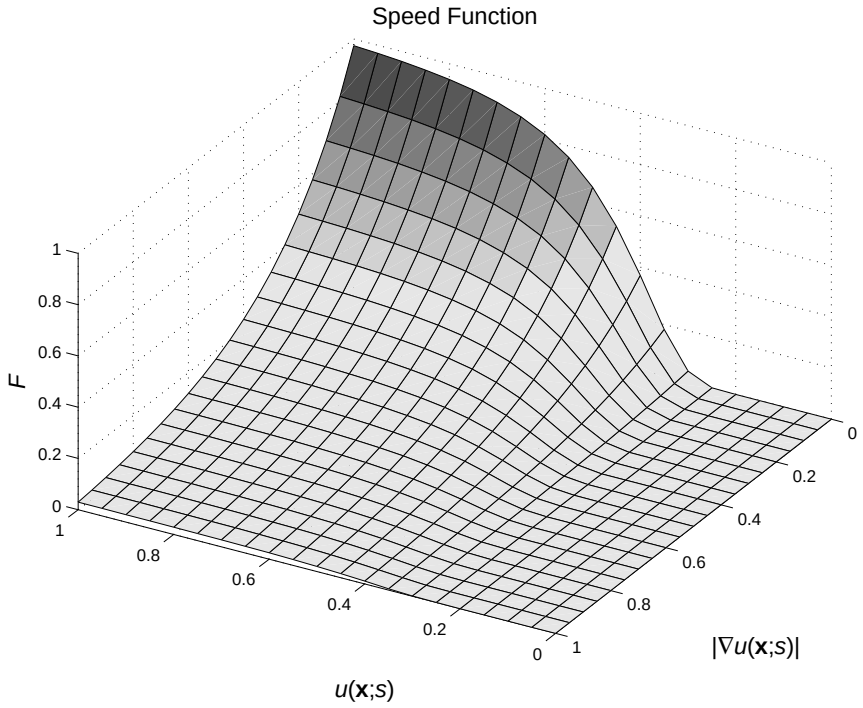



Figure 11: Hybrid gradient and intensity speed function ($\beta = 0.4$, $a = 4$, $b = 3.5$).

Our hybrid speed function is formed as the product of the intensity and gradient function $H$ and $G$ respectively

$$\begin{aligned} F(\mathbf{v}) &= H(\mathbf{v})\, G(\mathbf{v}) \\ &= \exp\left(-\left(\frac{\beta}{u(\mathbf{v};s)}\right)^{a} - b\,|\nabla u(\mathbf{v};s)|\right). \end{aligned} \tag{41}$$

This function allows large speeds where the intensity on the front is above the threshold and there is little boundary evidence but slows when either the intensity drops or a large gradient is encountered. The function approaches zero rapidly when both events occur, as shown in Figure 11.

# 5 Experiments

## 5.1 Experimental Data

Experiments were performed on 10 MR images of various modes and voxel dimensions from 7 patients, as indicated in Table 2. Images with the same number are scans of the same patient from different orientations. The volumes of two hand segmentations from two physicians are reported for 6 of these images (1, 2, 3R, 4, 5, 6) in [23]. For 5 additional images, the segmentations themselves were obtained, thus allowing for a more careful evaluation and comparison of automatic segmentations. The work in [23, 38] used a version of the image of patient 1 that was altered by artifacts from a conversion tool. We have

| Image | Mode | Voxel Size | Orientation |
|---|---|---|---|
| 1 | $T_2$ | $1 \times 1 \times 1\,\text{mm}^3$ | Axial |
| 2 | FLAIR | $1 \times 1 \times 2.5\,\text{mm}^3$ | Axial |
| 3P | $T_1$ | $1 \times 1 \times 2.5\,\text{mm}^3$ | Sagittal |
| 3Q | $T_1$ | $1 \times 1 \times 2.5\,\text{mm}^3$ | Coronal |
| 3R | $T_1$ | $2.5 \times 2.5 \times 2.5\,\text{mm}^3$ | Axial |
| 4 | FLAIR | $1 \times 1 \times 2.5\,\text{mm}^3$ | Coronal |
| 5 | FLAIR | $1 \times 1 \times 2.5\,\text{mm}^3$ | Coronal |
| 6 | FLAIR | $1 \times 1 \times 2.5\,\text{mm}^3$ | Coronal |
| 7Q | FLAIR | $0.5 \times 0.5 \times 2.5\,\text{mm}^3$ | Coronal |
| 7R | FLAIR | $0.5 \times 0.5 \times 2.5\,\text{mm}^3$ | Axial |

Table 2: Test data from ten MRI scans.

chosen to use the original version in our experiments because we cannot be sure of the nature of the alterations to the image. Since it is most likely that the image histogram was stretched, the effect of the alterations should be replicable through our segmentation framework.

## 5.2 Evaluation Criteria

We use several criteria to evaluate different aspects of our method. These include volumetric error to capture the amount of overall agreement with physician segmentations, information retrieval metrics for comparing actual decision strategies, and a new form of variance to measure the sensitivity of the method to initialization.

If, as in §3.2, we view an image as a set of voxels $X = \{x_i\}_{i=1}^N$, then a segmentation is some subset of $X$. For example, $S \subseteq X$ could be a computer segmentation and $H \subseteq X$ a hand segmentation.

### 5.2.1 Volume

Results in [23, 38] were reported in absolute volumes, with physician results alongside those from automatic segmentations. To allow for comparison, we measure the relative disparity between the volumes of automatic and physician segmentations, with volume error given by

$$E := \frac{\text{Volume}_{\text{Estimate}} - \text{Volume}_{\text{Physician}}}{\text{Volume}_{\text{Physician}}} = \frac{|S| - |H|}{|H|}.$$

However, this is not always a meaningful comparison.

In the extreme case, two completely disjoint segmentations of the same image may have the same volume. Of course, results reported earlier did not exhibit behavior this extreme. However, it leaves open the possibility that in certain cases some regions of tissue may be excluded by the computer that were included by physicians, and the *volume* of these

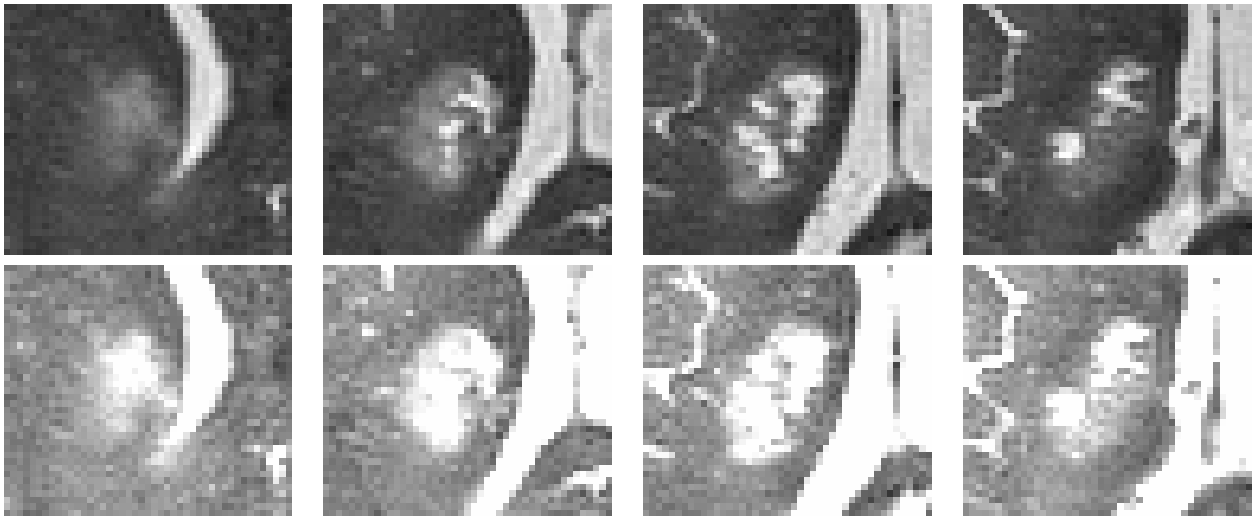


Figure 12: TOP: Slices from the original version of image 1. BOTTOM: The same slices from the altered version of image 1 used in [23, 38].

regions could be compensated for by other regions (incorrectly) included by the computer. This motivates the use of the following measurements.

### 5.2.2 Precision, Recall, and F-measure

It is important to evaluate whether the computer is making the same decisions as physicians in labeling tissue normal and abnormal. In this case, $S$ is the set of "retrieved" voxels and $H$ is the set of "relevant" voxels. Along these lines we measure the precision and recall of the computer segmentations. the precision $P$ and recall $R$ are defined as

$$\begin{aligned} P &:= \frac{|S \cap H|}{|S|} \\ R &:= \frac{|S \cap H|}{|H|}. \end{aligned}$$

The precision is thus the fraction of retrieved voxels that are relevant (i.e. correct), and recall is the fraction of relevant voxels that are retrieved.

A segmentation may have perfect recall by including all the voxels, but it will have poor precision because it includes many voxels that are not lesion. For example, this might happen when the speed function intensity threshold $\beta$ is set too low and the evolving front "leaks" outside the lesion boundary somewhere and eventually includes much of the image. Conversely, a segmentation may have very good precision by including only a few (correct) pixels, but it also has poor recall because it misses many lesion voxels. This might happen when the same threshold is set too high and the propagating front fails to "inflate."

These two values are combined to form the F-measure of a segmentation,

$$F := \frac{2PR}{P+R},$$

which is the harmonic mean of precision and recall. This gives us a single value to maximize and is more sensitive than the average to a difference between precision and recall .

### 5.2.3 Set Variance

Finally, it will be of use to investigate the variation in segmentations. As described before, it is possible for the volume of a segmentation, $|S|$ where $S \subseteq X$, to be misleading. By a similar argument, the standard deviation of segmentation volumes could be misleading. If $n$ segmentations consist of spheres of the same radius, each with a different center, then the volumes of the spheres are precisely the same, and the deviation of the volumes is zero. This counters intuition since the segmentations themselves actually do differ. We are therefore motivated to construct a definition of the "variance" over a collection of segmentations $S_i \subseteq X$, $i = 1 \ldots n$.

Given a random, real-valued sample $y_i$, $i = 1 \ldots n$, the sample mean and sample variance are given by

$$\begin{aligned} \bar{y} &:= \frac{1}{n} \sum_{i=1}^{n} y_i \\ s^2 &:= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 . \end{aligned}$$

The variance is the average squared deviation from the mean. However, the notion of a "mean" set is elusive, so we manipulate the variance expression to reveal

$$s^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - y_j)^2 . \tag{42}$$

We are now prepared to define an analog to sample variance where an observation is not a number, but a set. That is, we view segmentations as observations $S_i$ of a random set $S \in \wp(X)$. In this case, the usual variance (42) is evaluated over set volumes, or cardinalities, with $y_i = |S_i|$.

The symmetric difference between two sets $A, B$ is defined as

$$A \ominus B := (A \cup B) \setminus (A \cap B) . \tag{43}$$

The indicator function of the symmetric difference is the exclusive OR of the operands' indicator functions; an element is included in the resulting set if and only if it is in one or the other, but not both. Thus, if two sets are very similar, their symmetric difference is nearly empty, and if two sets are very dissimilar, the symmetric difference is nearly their union.

We propose to use the following as a measure of variance for sample segmentations $S_i \subseteq X$, $i = 1 \ldots n$:

$$s^2_{\ominus} := \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |S_i \ominus S_j|^2 . \tag{44}$$

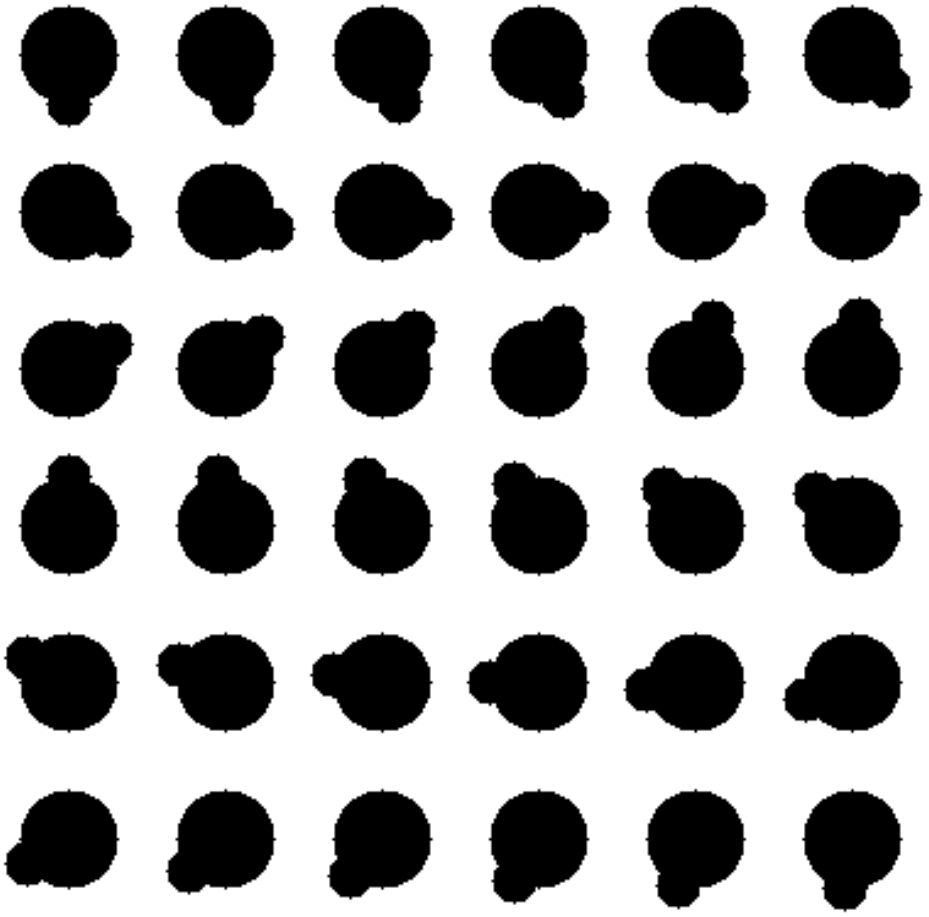


Figure 13: Images for comparing volume variance and set variance. Mean area $\bar{y} = 1395.5$, variance $s^2 = 22.474$, set variance $s^2_{\ominus} = 31677$.

Thus, (44) is analogous to (42) in that it measures the pairwise squared difference between all samples; these pairwise terms are then summed and normalized by the same factor. Whereas (42) squares the difference to yield squared magnitude, (44) measures the cardinality of the symmetric difference in order to obtain a magnitude of the difference between the sets. (The cardinality measurement thereby ignores the particular elements of contention between the pair, since they are not of direct concern here. However, if further examination of the discrepancy between two segmentations were desired, the actual *set* $S_i \ominus S_j$ would likely prove useful.)

As an example, for $n$ disjoint segmentations each of volume $v$, $s^2 = 0$ while $s^2_{\ominus} = \frac{2n}{n-1}v^2$. Figure 13 shows 36 images of a single shape at various rotations having a mean area 1395.5. The variance of the area (caused by discretization) is just 22.474, while the set variance is 31677. This rather dramatic difference indicates how different the black areas are when considered to be segmentations.

These two results show how the measures can be different, but we also show how they can be the same.

**Fact 5.1** *If a sample $S_i$, $i = 1 \ldots n$ has the property that $S_i \subseteq S_j$ or $S_j \subseteq S_i$ for all $i, j = 1 \ldots n$, then $s^2 = s^2_{\ominus}$.*

**Proof:**

If $S_i \subseteq S_j$, then $S_i \cup S_j = S_j$, $S_i \cap S_j = S_i$, and

$$\begin{aligned} |S_i \ominus S_j|^2 &= |(S_i \cup S_j) \setminus (S_i \cap S_j)|^2 \\ &= |S_j \setminus S_i|^2 \\ &= (|S_j| - |S_i|)^2 = (|S_i| - |S_j|)^2 . \end{aligned}$$

Because $S_i$ is a subset of $S_j$, the cardinality of the set difference is merely the numerical difference of their cardinalities, which facilitates the last step above. If $S_j \subseteq S_i$, similar calculations give the same result. Thus,

$$\begin{aligned} s^2_{\ominus} &:= \frac{1}{2n\,(n-1)} \sum_{i,j}^{n} |S_i \ominus S_j|^2 \\ &= \frac{1}{2n\,(n-1)} \sum_{i,j}^{n} (|S_i| - |S_j|)^2 \\ &= s^2. \end{aligned}$$

■

We see that when a sample of sets is like a Russian doll, where each observation fits inside another, the new set variance measure is precisely the same as the usual variance of volumes.

## 5.3 Segmentation Model

### 5.3.1 Scale and Threshold Parameters

Physicians in the process of manually segmenting images exhibit two behaviors worth noting. First, as mentioned in §2, they adjust the brightness and contrast levels of the image so that they may perceive greater discrimination in the voxels. Second, they zoom in to draw careful contours and zoom out when unsure of a particular boundary area. These two behaviors indicate the use of scale in regard to both size and intensity.

These observations led us to conduct segmentations with parameters varying along two axes (Figure 14):

**Scale:** Diffusion time $t$

**Threshold:** Speed function intensity threshold $\beta$

The lowest scale used is a small value that is close to the original image, yet removes a small amount of noise. The largest scale used is a value at which most lesions are reduced to a single blob, presumably near the last stable scale before being mixed with normal brain tissue.

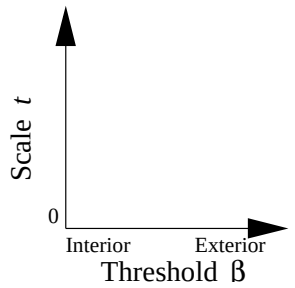



Figure 14: Parameter space induced by segmenting images at several scales and thresholds.

Since different images have different intensity properties, the threshold axis of the parameter space is restricted to bounds particular to the image being segmented. The two intensities of primary interest are those within the lesion, and that of the healthy brain tissue immediately surrounding it. Thus, two points are selected in the original image by a technologist (Figure 15):

1. an interior seed point $\mathbf{l}$, and
2. an exterior baseline point $\mathbf{b}$.

The interior seed point gives both an intensity value of the lesion and a location from which the segmenting surface is inflated. The exterior point gives a baseline on lesion intensity; since it corresponds to healthy tissue, no lesion will have an intensity that low in the filtered images where the intensity distribution within image regions is highly peaked.

The average of intensities in a 3 voxel neighborhood around the interior seed *at a particular scale* is used as an upper bound on the threshold. Similarly, the average of intensities in a 3 voxel neighborhood around the exterior point at the scale is a lower bound on the threshold. The point selected outside the lesion should correspond roughly to the brightest region excluded from the lesion. Under the imaging modalities used in this evaluation, this generally corresponds to normal brain tissue. Figure 15 illustrates an interior seed point and a baseline point in the healthy tissue just outside the lesion. Notice the ambiguous left boundary of the lesion. By choosing a baseline point well into healthy tissue, varying the threshold between the intensities of these two points will place the lesion boundary given by automatic methods at varying places along the ambiguous edge.

It is important to note that while the location of the points is constant through the segmentation process, the intensities at the points vary with scale. What explicitly varies along the horizontal axis of segmentation parameters (Figure 14) is not the threshold used, but the percentage difference *between* the point intensities that are used to calculate the threshold used. Hence, if $\beta(0;t) := u(\mathbf{b};t)$ is the intensity of the exterior baseline point at scale $t$ and $\beta(1;t) := u(\mathbf{l};t)$ is the intensity of the interior seed point at scale $t$, the explicit segmentation parameter along the threshold axis is $p$, where $0 \leq p \leq 1$ and we use

$$\beta(p;t) := \beta(0;t) + p(\beta(1;t) - \beta(0;t))$$

for $0 < p < 1$ as the speed function intensity threshold parameter.

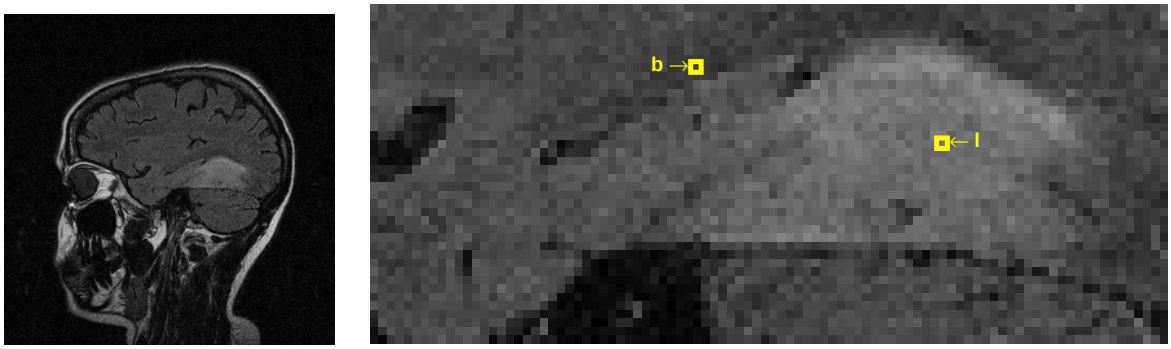



Figure 15: Example of an interior seed point $\mathbf{l}$ and exterior baseline point $\mathbf{b}$.

### 5.3.2 Confidence Parameter

Section 1.4 describes how the frequency of a voxel's inclusion in a segmentation is related to our confidence in it being lesion. We therefore "marginalize" the scale and threshold parameters into a space of segmentations based on confidence. If $\mathcal{T}$ and $\mathcal{P}$ are the (finite) sets of scales and threshold ratio parameters used, let a particular segmentation be $S(t,p) \subseteq \Omega$ for $t \in \mathcal{T}$ and $p \in \mathcal{P}$. The frequency of a voxel's inclusion is given by

$$\phi(\mathbf{v}) = \frac{1}{|\mathcal{T}|\,|\mathcal{P}|} \sum_{t\in\mathcal{T}} \sum_{p\in\mathcal{P}} \chi_{S(t,p)}(\mathbf{v}),$$

which is the total number of times a voxel $\mathbf{v} \in \Omega$ appears in all segmentations divided by the number of segmentations. We define a confidence segmentation to be the set of voxels above some minimum frequency:

$$C(c) = \{\mathbf{v} \in \Omega \mid \phi(\mathbf{v}) \geq c\}$$

for $0 \leq c \leq 1$.

## 5.4 Experimental Details

We use isotropic nonlinear diffusion exclusively in all experiments for the following reasons:

- Edge-enhancing anisotropic diffusion with maximal diffusivity along edges, as in (14) quickly wipes out the fine structures of some lesions.
- The scale-space framework is important to our model of segmentation, and thus a modified, non-differentiable diffusivity such as (16) cannot be used since it does not guarantee the scale-space axioms are upheld.
- Using a constant reduced diffusivity along edges is less effective than isotropic diffusion when $\lambda_1 > \lambda_i$.

Additionally, the diffusivity function (12) is used in an AOS implementation with the following parameters:

$$\begin{aligned} m &= 4 \\ \sigma &= 1 \\ \lambda &= 2.55 \\ \tau &= 2.5. \end{aligned}$$

The Fast Marching Method is used for all image segmentation. Its robust treatment of complex lesion structure coupled with the low time complexity of the method itself make FMM an ideal choice.

The hybrid speed function (41) drives segmentation in the Fast Marching Method. Section 5.3 explains the selection of the intensity threshold parameter $\beta$. The values of the other parameters are:

$$\begin{aligned} a &= 40 \\ b &= 3.5 \end{aligned}$$

The times $t$ at which the image is segmented are discretely sampled on a logarithmic scale, as this is commonly held as the natural scale parameter [11, 20]. Along the threshold axis, $p$ is sampled linearly.

We exclude segmentations in the following cases:

- the propagating front fails to expand beyond the voxels neighboring the interior seed point, i.e. the resulting segmentation volume is less than 28 voxels
- the propagating front reaches a boundary of the image in the X-Y plane

The first case is an example of an invalid under-segmentation. This might happen when the threshold parameter is set so close to the intensity of the initial point that the front has no speed. Regarding the second case, it is fairly easy to specify a box bounding the lesion in two dimensions across all slices. We assume lesions are far enough from such a boundary that if the front reaches a boundary it has over-segmented as a result of poor parameters, i.e. a large scale or a low threshold. (Fronts reaching a boundary in the Z direction are not disqualified because some images exhibit lesion in all slices.) These segmentations are identifiably incorrect and are thus excluded from the measurements to avoid skewing them.

## 5.5 Results

The results of our segmentation model are dependent upon three primary factors: the scale at which the image is examined, the threshold used for stopping the propagating front, and the seed points used to initialize the algorithm. This section evaluates each of these.

First, the space of scale and threshold parameters is scanned for the parameter pair that gives the minimum absolute volume error for each image. Table 3 gives the minimum volume error achieved on each image when compared with the original physician hand segmentation mean volumes reported in [23]. The sensitivity of the optimal parameters is

| Image | Volume Error | Scale $t$ | Threshold $p$ |
|---|---|---|---|
| Doctor 1 | | | |
| 1 | -0.148 | 362217 | 0.45 |
| 2 | -0.001 | 1636 | 0.95 |
| 3R | 0.000 | 330 | 0.95 |
| 4 | 0.007 | 8103 | 0.95 |
| 5 | -0.000 | 270 | 0.75 |
| 6 | 0.000 | 55 | 0.95 |
| Doctor 2 | | | |
| 1 | -0.095 | 362217 | 0.45 |
| 2 | 0.000 | 40135 | 0.95 |
| 3R | 0.000 | 1998 | 0.85 |
| 4 | 0.004 | 8103 | 0.85 |
| 5 | -0.004 | 45 | 0.75 |
| 6 | 0.005 | 4447 | 0.75 |

Table 3: Minimum volume error for each image and the parameters at which each occurs.

illustrated graphically in Figure 16. The means and standard deviations of volume over a three scale, one threshold neighborhood (thus, twenty-one observations) around these minimizing parameters are plotted beside the mean and and standard deviation of each physician's segmentations.

Next, the space of scale and threshold parameters is scanned for the parameter pair that gives the minimum total absolute volume error for the same six patients. For each doctor the parameters (scale and threshold) that gave the least total error magnitude over the six images were identified. Table 4 gives individual volume errors at the parameters for which total absolute volume error occurs for each doctor, thus adapting parameters to a particular physician segmentation style. On the other hand, Table 5 gives individual volume errors at the parameters for which total absolute volume error for both doctors occurs, thus constraining the algorithm to a unique parameter setting and comparing the behavior to that of physicians. Table 6 compares volumes from physician hand segmentations and the results from [23] with new results from diffusion/FMM. As before, means and standard deviations of volume are taken over a 3 scale, 1 threshold neighborhood around parameters minimizing total absolute error for each doctor. The same results are illustrated graphically in Figure 17. The mean and standard deviation over all scales and all thresholds is illustrated in Figure 18.

Once again, the space of scale and threshold parameters is scanned, this time for the pair giving the best F-measure for each image whose physician segmentation we obtained. Table 7 gives the F-measure, precision and recall at these optimal parameters, along with the corresponding volumetric error. Figure 19 shows the hand segmentation of image 2 alongside the closest automatic segmentation.

To evaluate the sensitivity of the segmentations to the interior seed point location (and by extension, intensity), we chose 100 interior seed points uniformly at random from the

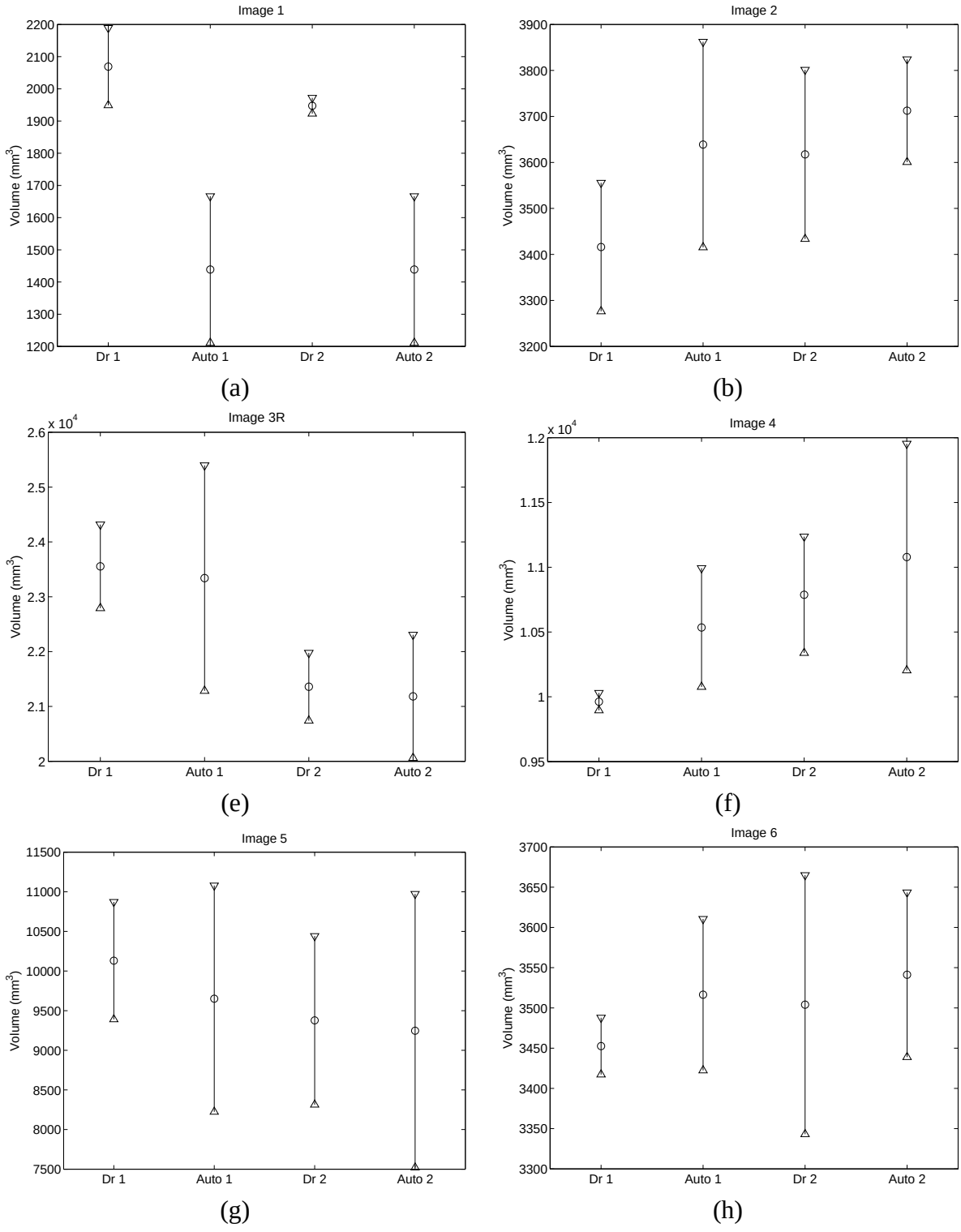



Figure 16: Mean and standard deviations of volumes from physician and automatic segmentations over the neighborhood around optimal parameters for each image and each doctor.

| | Volume Error | |
|---|---|---|
| Image | Doctor 1 | Doctor 2 |
| 1 | -0.766 | -0.705 |
| 2 | 0.020 | 0.102 |
| 3R | -0.152 | 0.004 |
| 4 | 0.007 | 0.051 |
| 5 | 0.009 | -0.006 |
| 6 | -0.046 | -0.030 |

Table 4: Individual errors at the parameters where minimum total absolute volume error occurred for each doctor: $t = 8103$, $p = 0.95$ for doctor 1 and $t = 602$, $p = 0.85$ for doctor 2.

| | Volume Error | |
|---|---|---|
| Image | Doctor 1 | Doctor 2 |
| 1 | -0.723 | -0.705 |
| 2 | 0.110 | 0.048 |
| 3R | -0.095 | -0.001 |
| 4 | 0.150 | 0.062 |
| 5 | -0.008 | 0.071 |
| 6 | -0.012 | -0.026 |

Table 5: Individual errors at the parameters where minimum total absolute volume error occurred for both doctors: $t = 1636$, $p = 0.85$.

| | Doctor 1 | | Snake Hybrid | | Diffusion/FMM | |
|---|---|---|---|---|---|---|
| Image | mean | sd | mean | sd | mean | sd |
| 1 | 2069 | 120 | 1927.6 | 49.3 | 594.3 | 92.5 |
| 2 | 3416 | 139.7 | 3841.5 | 277.8 | 3849.0 | 292.2 |
| 3R | 23555 | 762.3 | 21821.9 | 617.7 | 21096.7 | 1041.4 |
| 4 | 9962.5 | 67.2 | 10031 | 906.8 | 11078.8 | 875.0 |
| 5 | 10131 | 740.7 | 9536.9 | 251.8 | 10623.8 | 673.6 |
| 6 | 3452.5 | 35.3 | 3525 | 117.9 | 3422.1 | 103.8 |
| | Doctor 2 | | Snake Hybrid | | Diffusion/FMM | |
| 1 | 1947.5 | 24.5 | 1927.6 | 49.3 | 579.7 | 81.4 |
| 2 | 3617.5 | 183.8 | 3841.5 | 277.8 | 3933.6 | 362.0 |
| 3R | 21359 | 618.7 | 21821.9 | 617.7 | 22183.0 | 1810.5 |
| 4 | 10787.5 | 449.0 | 10031 | 906.8 | 11358.7 | 1028.9 |
| 5 | 9377.5 | 1064.2 | 9536.9 | 251.8 | 8681.9 | 1205.8 |
| 6 | 3504 | 160.9 | 3525 | 117.9 | 3408.1 | 148.7 |

Table 6: Means and standard deviations of physician segmentations, the previous hybrid method (using five initializations), and the diffusion/FMM method (using one initialization over a range of 7 scales and 3 thresholds).

| Image | Precision | Recall | F-measure | Volume Error | Scale $t$ | Threshold $p$ |
|---|---|---|---|---|---|---|
| 2 | 0.860 | 0.827 | 0.843 | -0.039 | 01998 | 0.45 |
| 3P | 0.935 | 0.833 | 0.881 | -0.109 | 00330 | 0.35 |
| 3Q | 0.934 | 0.834 | 0.881 | -0.108 | 03641 | 0.15 |
| 7Q | 0.861 | 0.690 | 0.766 | -0.198 | 06634 | 0.35 |
| 7R | 0.861 | 0.537 | 0.661 | -0.377 | 22026 | 0.65 |

Table 7: Maximum F-measure for each image and the parameters at which each occurs.

| Image | Segmentations | Mean | Median | Std | Set Std | Volume Error |
|---|---|---|---|---|---|---|
| 1 | 20 | 1536.50 | 1475.50 | 164.93 | 165.48 | -0.2349 |
| 2 | 100 | 3709.80 | 3557.50 | 921.68 | 921.68 | 0.0860 |
| 3P | 63 | 28534.37 | 28952.50 | 993.26 | 993.26 | -0.1223 |
| 3Q | 79 | 27577.91 | 27560.00 | 110.76 | 110.77 | -0.1045 |
| 3R | 84 | 21701.08 | 22398.44 | 5882.13 | 5882.13 | -0.0787 |
| 4 | 100 | 11163.95 | 10716.25 | 1832.79 | 1832.79 | 0.1206 |
| 5 | 67 | 10651.72 | 10440.00 | 494.56 | 494.56 | 0.0514 |
| 6 | 80 | 2879.91 | 3192.50 | 1031.20 | 1031.20 | -0.1658 |
| 7Q | 60 | 28707.75 | 29389.06 | 1952.33 | 1952.33 | -0.2145 |
| 7R | 41 | 22122.21 | 22451.25 | 861.00 | 861.01 | -0.3881 |

Table 8: Segmentation results at one scale and threshold with several seed points.

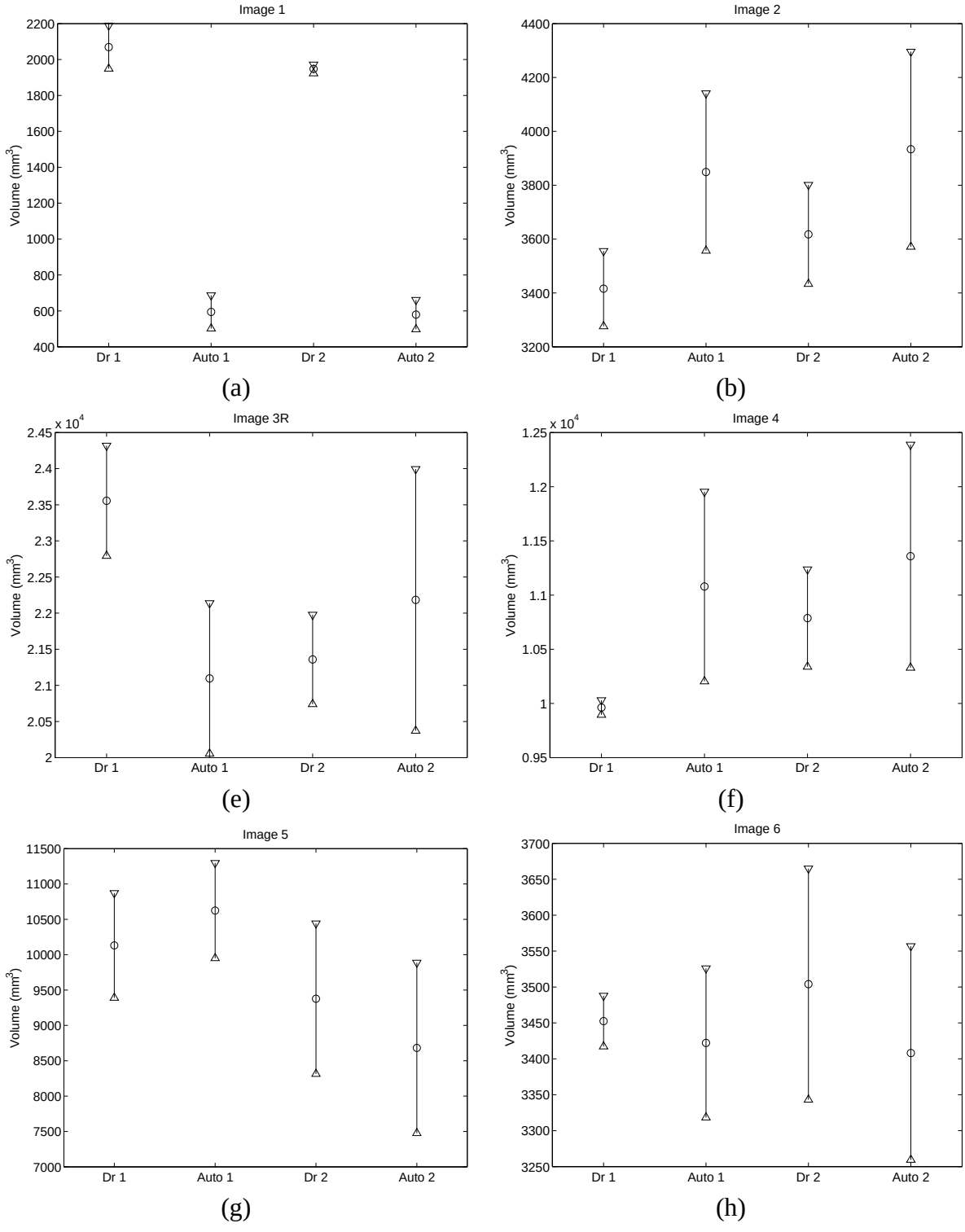



Figure 17: Mean and standard deviations of volumes from physician and automatic segmentations over the neighborhood around the optimal parameters for each doctor.

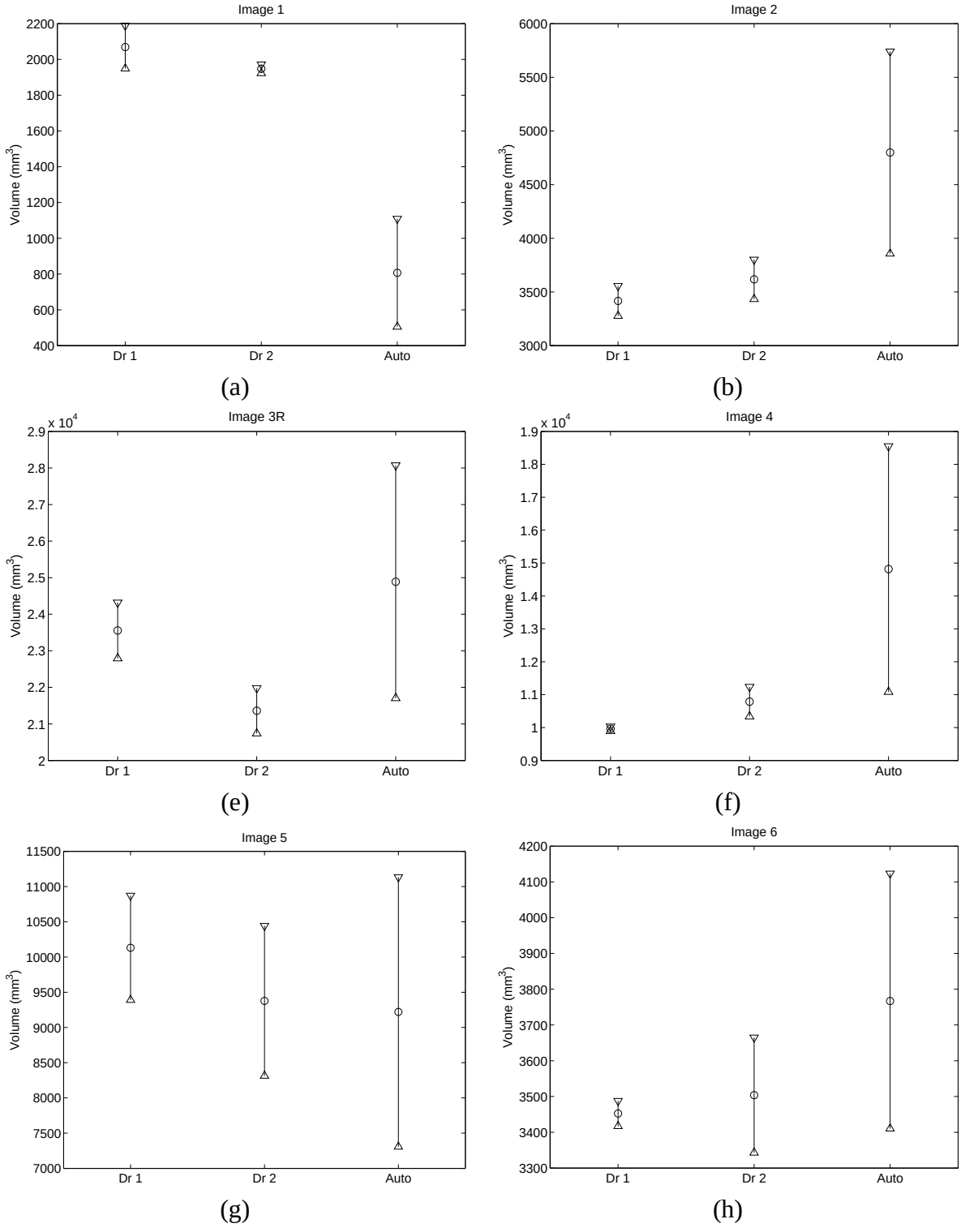



Figure 18: Mean and standard deviations of volumes from physician and automatic segmentations over the entire experiment parameter space.

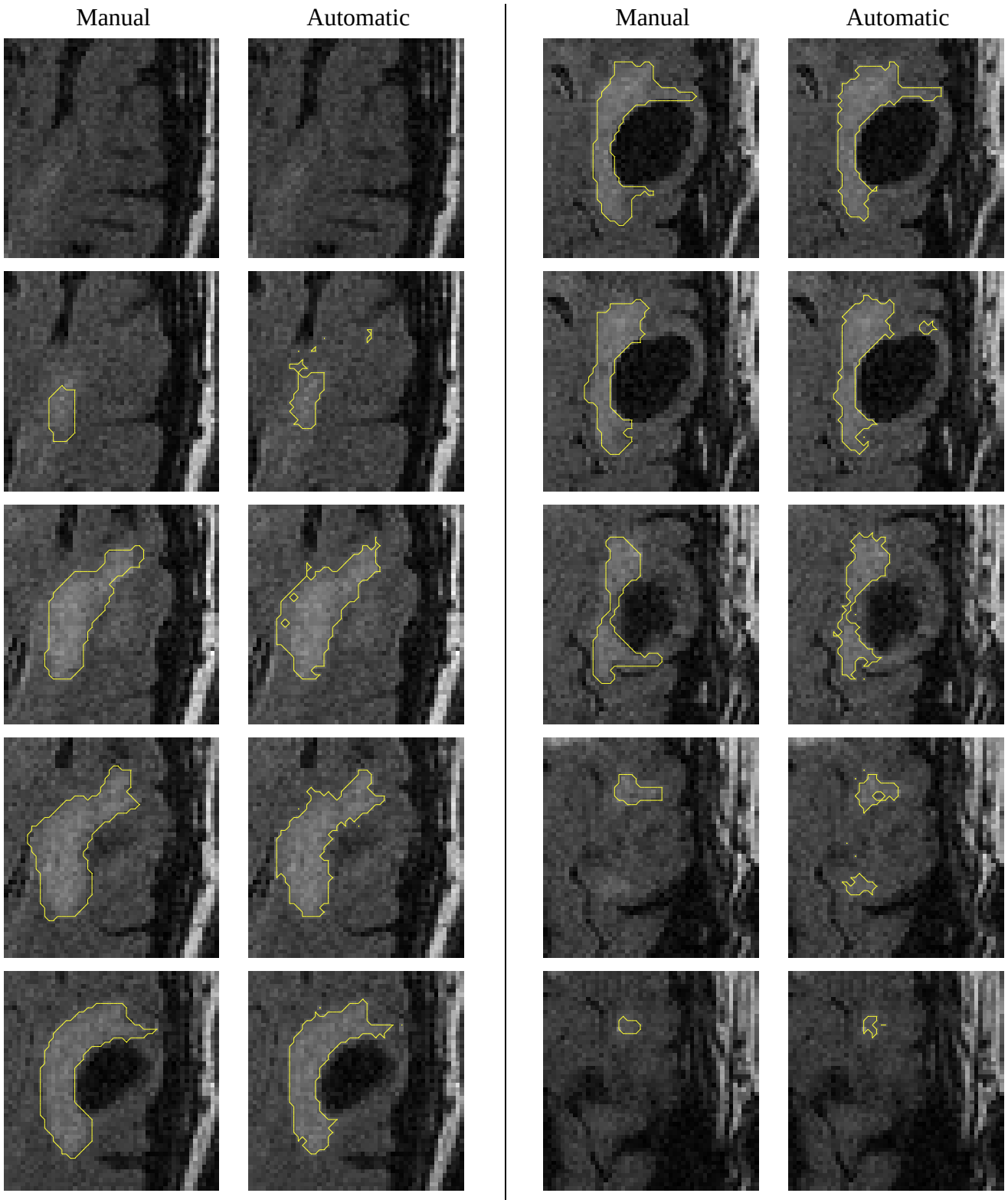
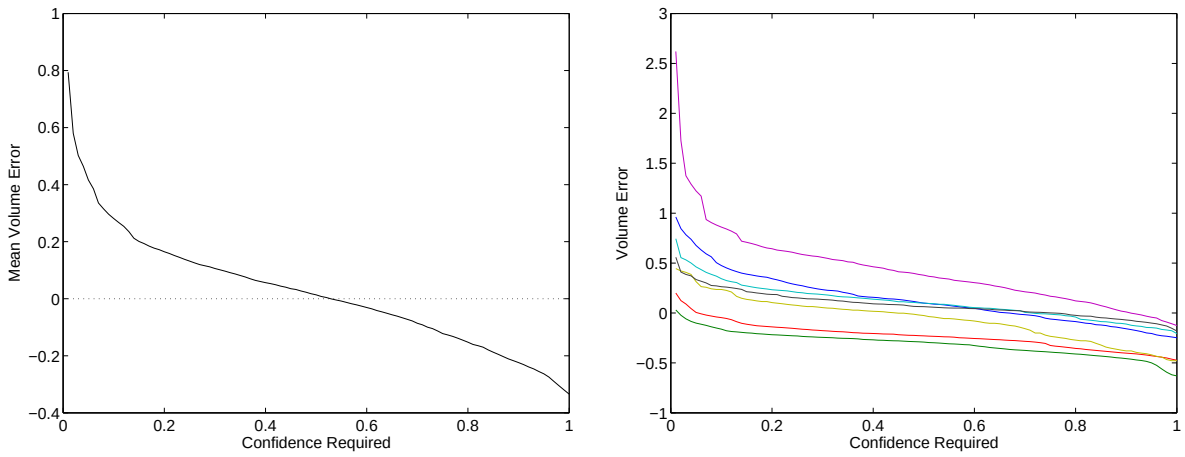



Figure 19: Physician manual segmentation and its closest automatic segmentation for image 2.




(a) (b)

Figure 20: Error vs. Confidence: (a) mean volume error; (b) volume error for each image.

most conservative automatic segmentation of each image (lowest scale, highest threshold) and segmented each image at the most accurate scale and threshold combination as reported in Tables 3 (for doctor 1 only) and 7. As before, invalid segmentations were discarded from the calculations. Table 8 reports the number of valid segmentations out of 100 possible, the mean, median, and standard deviation of the volume, and the set standard deviation, which is the square root of (44). The volume error compares the mean volume to the manually established volume.

For seven lesions in our data set, the error on segmentations of is plotted versus confidence in Figure 20. The swaths of CSF through images 7Q and 7R are large enough that they do not disappear before the maximum scale of the experiments is reached. The segmentation style of the doctor on these two images is too different from the algorithm's and is thus irreconcilable in the current parameter space, so they are omitted from the calculations. Image 1 is omitted as well because its behavior of under segmenting at almost all parameters due to a leaking tendency is atypical. Table 9 shows the individual error on the lesions at the confidence yielding zero mean error.

## 5.6 Analysis

Clearly the segmentation model has the capacity to perform almost identically to physicians. The results of Table 3 show that it is possible to achieve almost zero volume error for each image and each physician at some parameter setting. The only exception, image 1, would certainly achieve a small error comparable to the rest except for the fact it neighbors a brain ventricle that is as bright as the lesion itself. At larger scales, it becomes possible to segment the multimodal lesion of image 1 in a fashion similar to that exhibited by the doctors, but the lesion and ventricle also begin to join, causing the propagating front to leak.

While the model has the capacity to perform perfectly on every lesion, if automatic segmentations are required, then parameters must be constrained. As described in the in-

| | Overall Optimal (Zero Mean Error) | Individual Optima | | |
|---|---|---|---|---|
| Image | Volume Error | Optimal $c$ | Volume Error | F-measure |
| 2* | 0.086 | 0.67 | -0.001 | 0.849 |
| 3P | -0.301 | 0.03 | -0.059 | 0.894 |
| 3Q | -0.236 | 0.13 | 0.087 | 0.886 |
| 3R | 0.087 | 0.71 | 0.002 | - |
| 4 | 0.352 | 0.91 | -0.002 | - |
| 5 | -0.049 | 0.45 | -0.001 | - |
| 6 | 0.055 | 0.76 | -0.001 | - |

Table 9: Individual errors at optimal confidence ($c = 0.53$) with zero mean error, standard deviation 0.220, and individual optimal confidence values with corresponding F-measure and/or volume error. F-measure is the primary optimizing criterion with volume error a secondary fall back. *Image 2 gives individual $c$ and volume results based on original volumetric data, and F-measure based on a segmentation acquired later.

troduction, this means picking the consistent strategy that gives the least error. Earlier results used a single set of parameters over six patient images with five initializations. If we restrict our parameters to a single scale and threshold that seem to fit a doctor's segmentation style, the results are still almost entirely within a 10% error, as shown in Table 4. Image 1 was acquired with a very different MRI mode called $T_2$. Since different tissue types image quite differently under $T_2$, this image requires a much larger scale to achieve results closer to the physician volume. It might be more appropriate to consider this image (and any other $T_2$ images) separately due to the different magnitude of scales required. In any case, the large error on image 1 should not be considered too severely. The lower resolution of image 3R causes a narrow finger of lesion (cf. Figure 24(b)) to be absent at the higher scale parameter ($t = 8103$) used to emulate the segmentations of doctor 1, thus causing the underestimate. These results show that even when one set of rules is required for a doctor, the results are still extremely good. Moreover, the model demonstrates a reasonable insensitivity to the scale and threshold when compared with the physician segmentations and the previous work (Table 6 and Figure 17). If parameters must be constrained even further to one setting for all doctors, the results are similarly good, and the total error for each doctor increases only slightly (by 0.024 for doctor 1 and 0.017 for doctor 2), as shown in Table 5.

Error versus confidence plotted in Figure 20 show the promise of the frequency approach to segmenting. The intuition that over-segmentation should occur when low confidence is required and under-segmentation should occur when high confidence is required is born out in Figure 20(a). Therefore, we might set a default confidence requirement at the zero crossing $c = 0.53$, where the average error is zero. That the confidence giving zero mean error is almost exactly one half is very appealing. Additionally, the standard deviation of volume error is $0.220$, barely above the error deemed clinically useful by physicians. Interestingly, using the confidence measure to agglomerate segmentations increased F-measure slightly over the best possible using only one segmentation (from Table 7).

While we do not have the means to do an evaluation pairing trained technologists with the automatic segmentation, we do not feel the "single parameter for all lesions" setup is the best or most appropriate. A glance at the numbers of Table 9 indicates why. The large negative error on images 3P and 3Q at $c = 0.53$ are because both contain ambiguous regions that doctors ended up including. These areas of uncertainty are only included at lower confidence levels, and similar uncertain structures will always be eliminated at the "average" confidence level. Conversely, image 4 contains healthy tissue that is relatively bright in some regions. When the baseline point does not lie in one of these brighter regions (as it did not in our experiments), the segmentation includes them at lower thresholds. Hence, only with stricter confidence will the correct region be isolated.

Rather than allowing these anatomical nuances to be ignored in pursuit of correct segmentation of the more typical lesions, we propose to allow a user to choose the voxel inclusion frequency isocontour—confidence—he or she feels is the best representative of actual lesion boundaries, as described in §1.4. This is validated by the results in Table 7, which gives the maximum F-measure on the images for which we do have physician segmentations (instead of only volumes). Figures 19, 21, and 22 show select slices of the physicians' manual segmentations alongside the automatic segmentation with the highest precision and recall. The contours are similar in appearance and the differences in volume are within the clinically useful bounds of 20%. Segmentation is a tedious process when done manually because it requires classification consistency over several slices. So-called "ground-truth" data is not a gold standard because it is prone to inconsistency. It might be said that the closest automatic segmentations as given in Table 7 are the closest fit to the manual segmentation, but with a consistent decision process.

The larger errors reported for images 7Q and 7R in Table 7 bear explaining. While the precision is comparable to that of the other images, drastically low recall (and thus large volumetric error) is caused by the refusal of the algorithm to include the large black regions of CSF, as described in §1.2.3. Manual segmentations are strenuous and time consuming, so for these two images doctors drew a contour *surrounding* affected areas. As mentioned before, physicians have informed us that such results are not necessarily incorrect [15].

Figures 23-25 show frequency isocontours for slices from patient images. Cool colors indicate few inclusions, and warm colors indicate many inclusions. The outermost blue contour encloses voxels included in the lesion segmentation at least once, and the innermost red contour encloses voxels included every time.

Image 2 in Figure 23(a) finds "uncertainty" in the finger on the right-hand side the lesion, but it turns out this is similar to the physician segmentation (cf. Figure 19). One outlying contour in 23(b) indicates the ambiguity of the left-hand region. As it happens, this outer contour is also similar to the physician segmentation found in Figure 26(a). An interesting example is found in 24(a). As we witnessed the physician segmenting this image, he spent considerable time evaluating the tissue in the upper-right corner of the lesion, before eventually including the area indicated by the outermost contour, shown in Figure 26(b). Figure 25(a) indicates that the uncertain bottom part of the lesion is included in at least some segmentations, but the parameters at which this occurred made the results too different from the hand segmentation we received to be recorded as the most accurate (cf. Figure 22). Finally, the cool colors in the disconnected upper-right contours of Figure

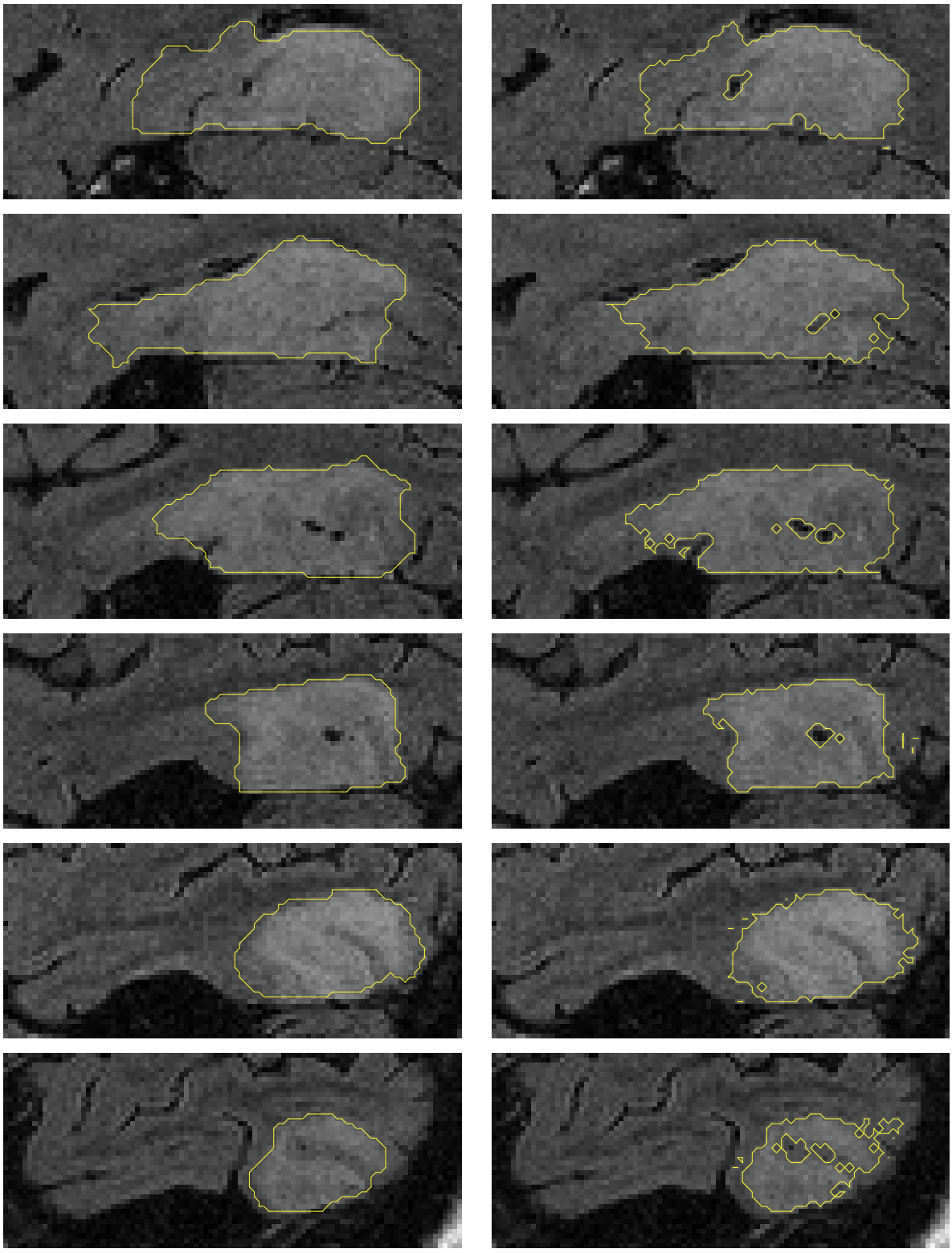


Figure 21: Selected slices from a physician manual segmentation (left) and its closest automatic segmentation (right) for image 3Q.

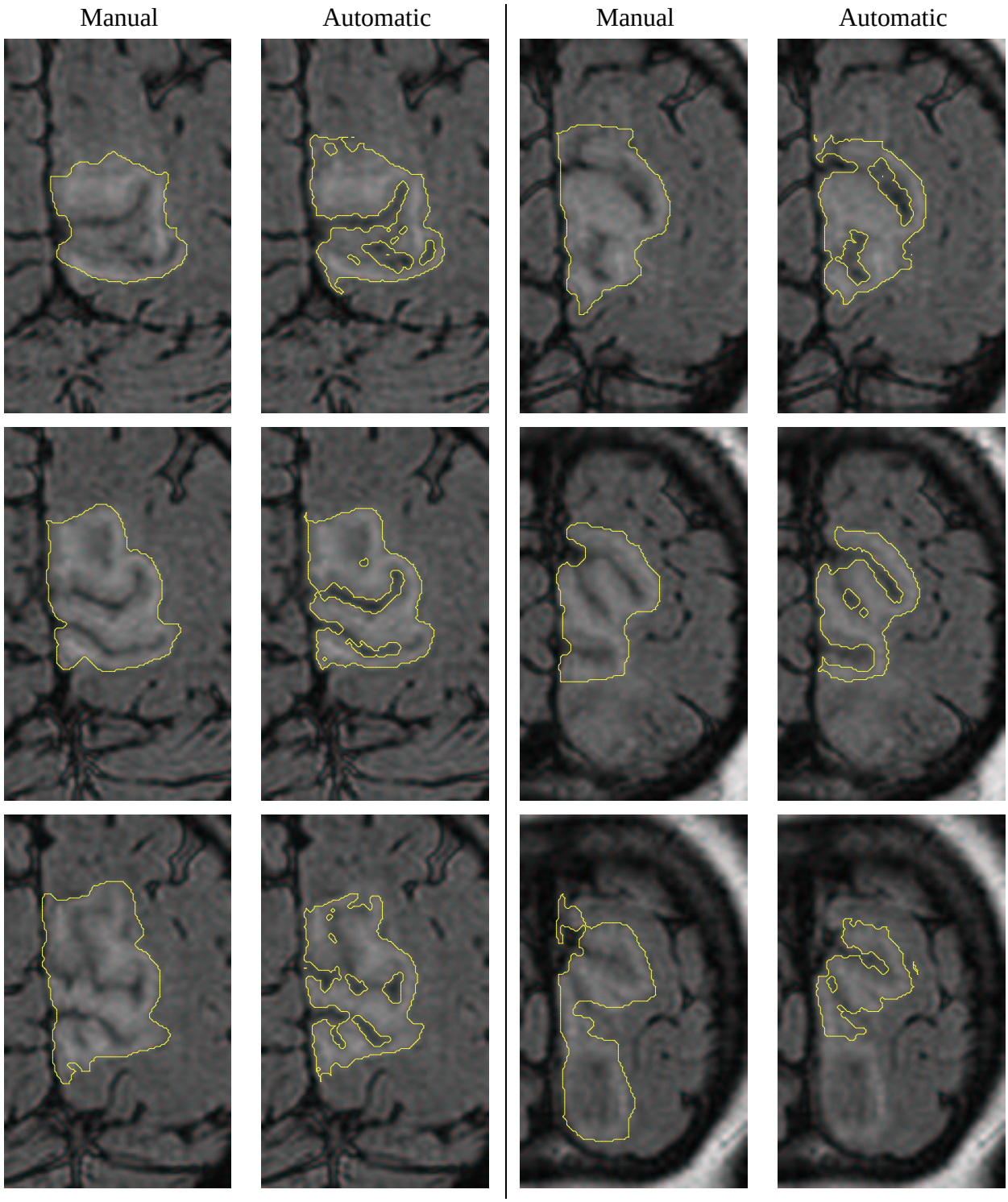
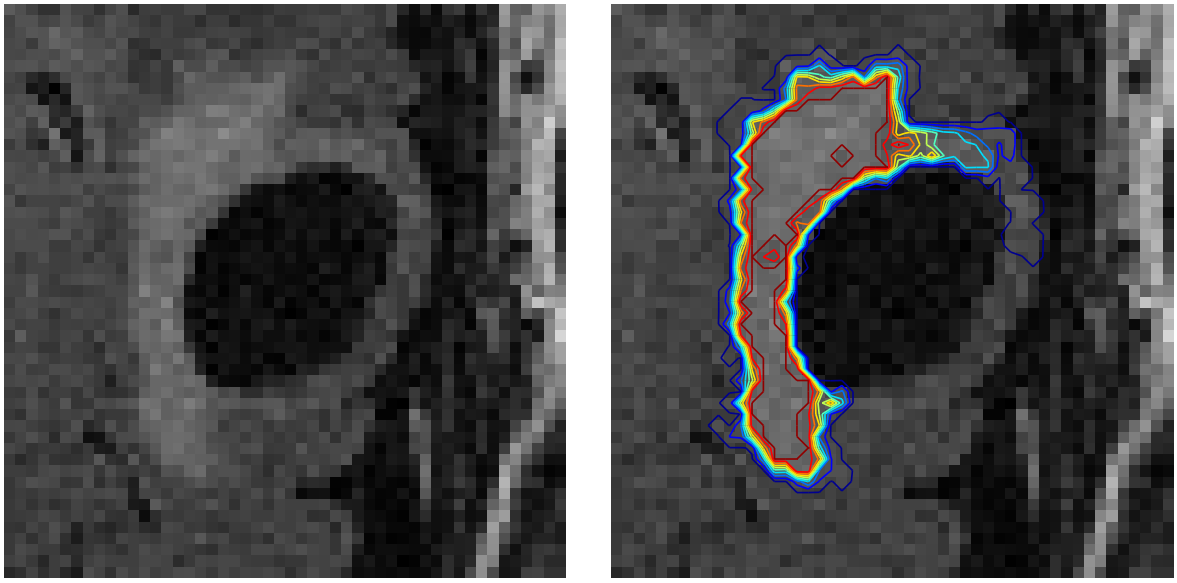
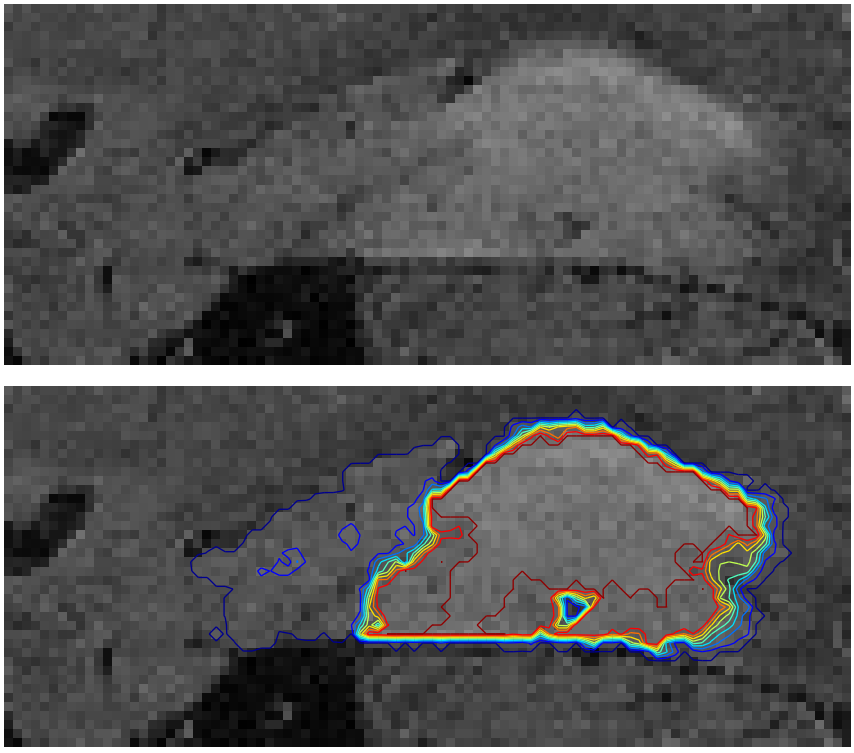



Figure 22: Selected slices from a physician manual segmentation and its closest automatic segmentation for image 7Q.

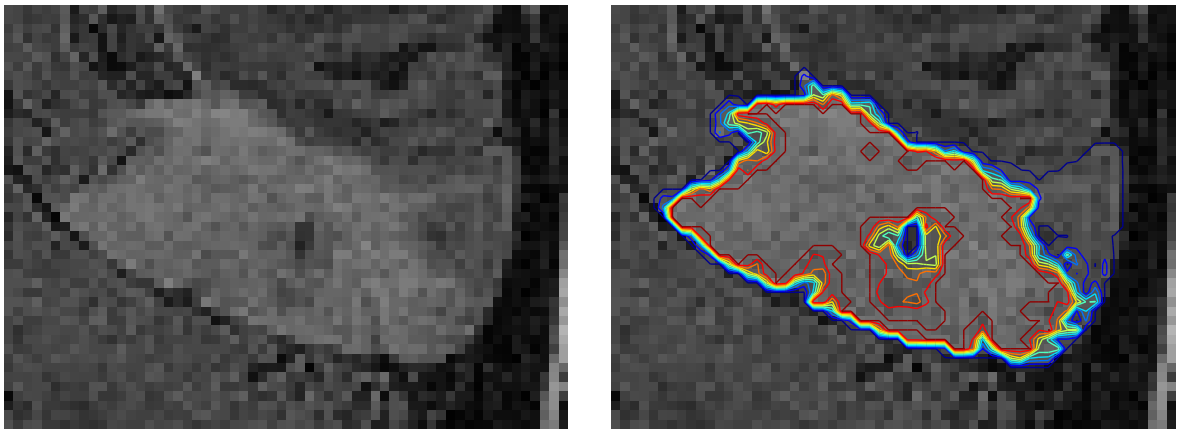


(a)

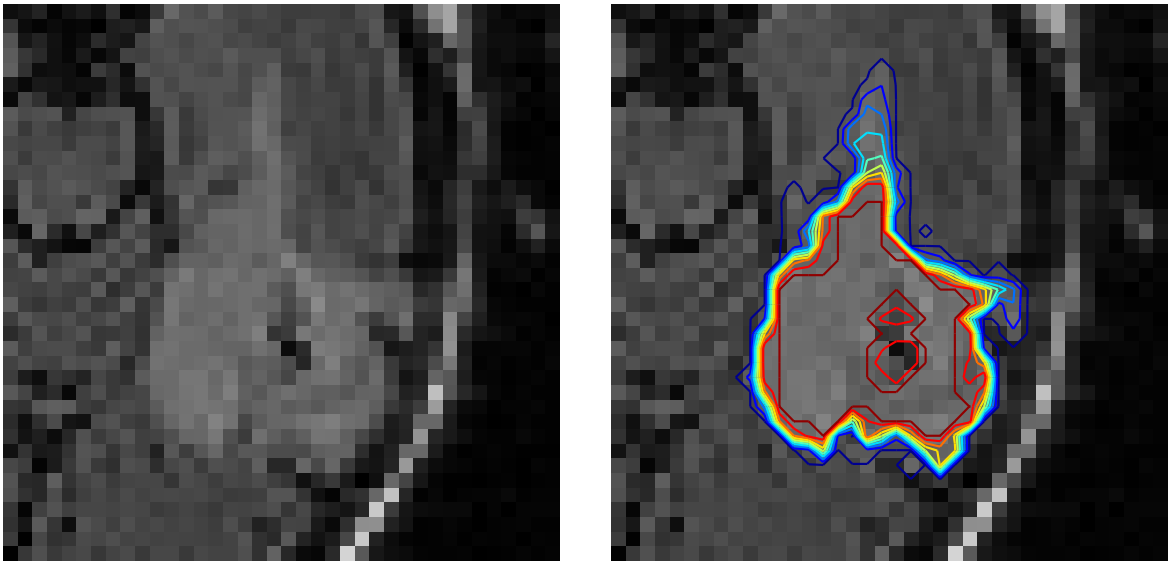


(b)

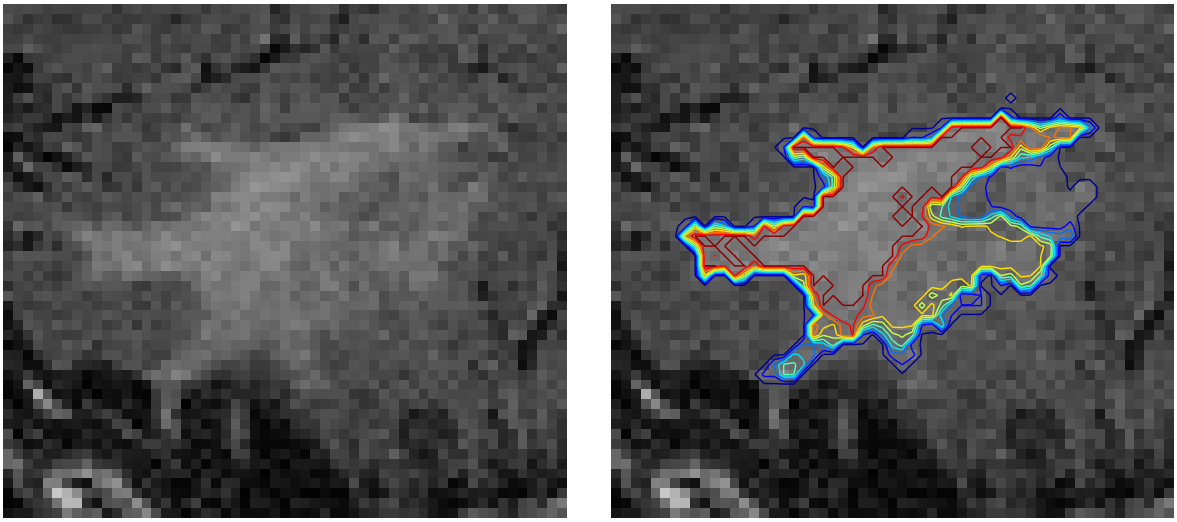
Figure 23: Frequency isocontours plotted at ten confidence levels: (a) image 2; (b) image 3P.



(a)

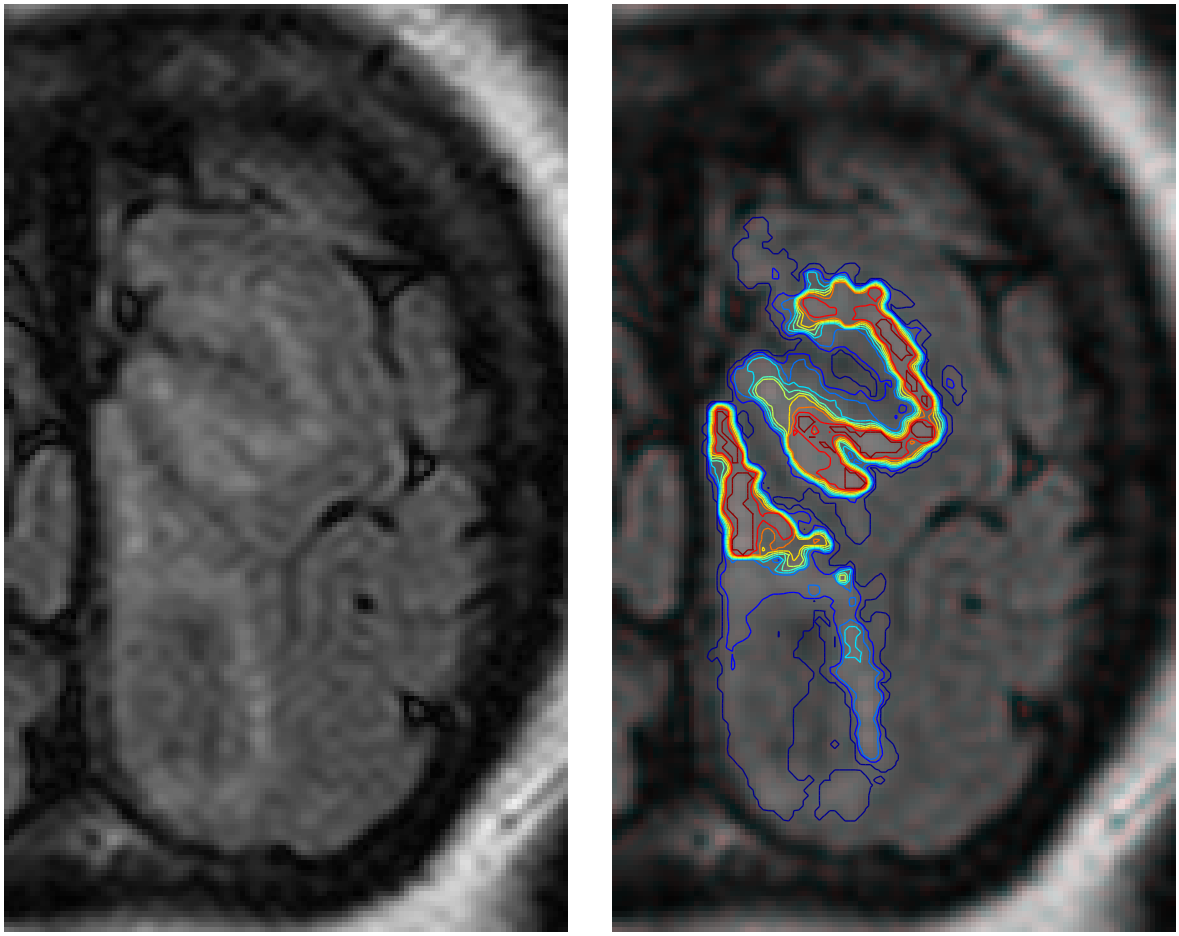


(b)

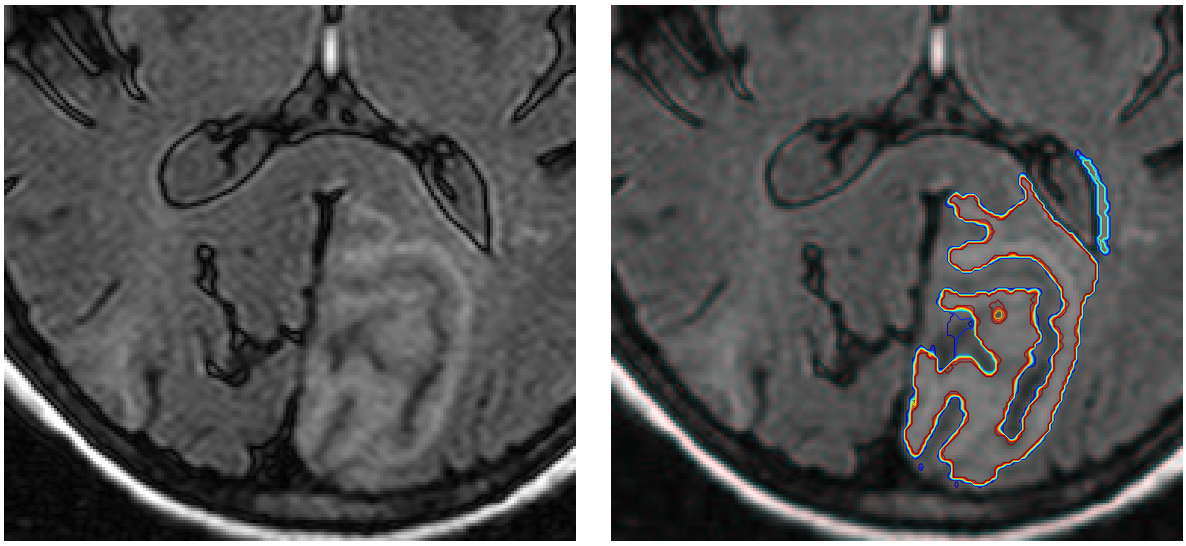


(c)

Figure 24: Frequency isocontours plotted at ten confidence levels: (a) image 3Q; (b) image 3R; (c) image 5.

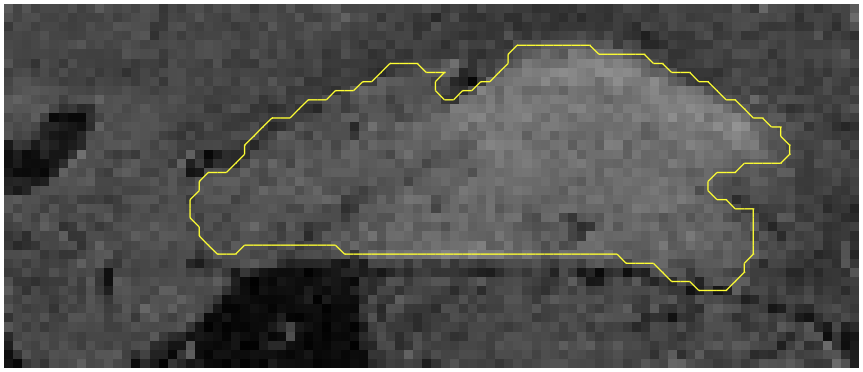


(a)

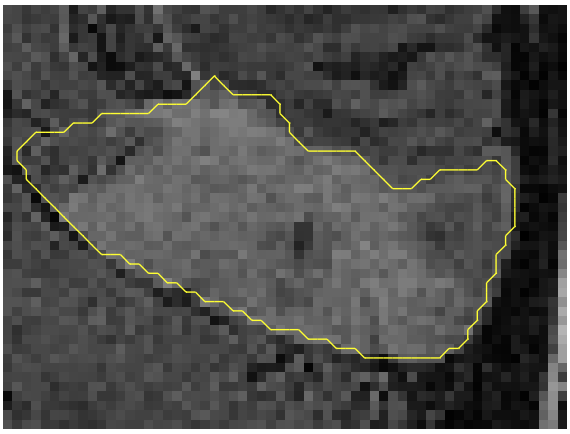


(b)

Figure 25: Frequency isocontours plotted at ten confidence levels: (a) image 7Q; (b) image 7R.



(a)



(b)

Figure 26: Slices of physician segmentations: (a) image 3R; (b) image 3Q.

25(b) indicate the uncertainty of our algorithm concerning whether the region is lesion. Because this region is typically bright since it neighbors the ventricle, it is in fact one that should be excluded (cf. §1.2.1).

Finally, we test the sensitivity of the segmentation process to the location and intensity of the interior point, as demonstrated in Table 8. The best segmentation of image 1 was achieved previously at a very low threshold for the maximum scale we allowed. It is logical that many segmentations of image 1 had to be discarded, because slight deviation from the intensity of the original point at such a low threshold makes the front prone to leaking. If the threshold were to be raised, more consistency would likely be found, but a larger average error would result due to smaller volumes. The average volume error for images 2, 3R and 5, is roughly consistent with, though slightly larger than, the volume errors reported in Table 3; errors for 3P, 3Q, 7Q, and 7R are also consistent with volume errors in Table 7. Image 6 is a multi-modal lesion, so there is still a large variation in the intensity of the seed point because the minimum volume error happened to occur at a very low scale. Finally, note that there is almost no difference whatsoever between the usual measure of standard deviation over volume and our proposed definition over sets. This is due to Fact 5.1. The algorithm is not making inconsistent segmentations by including some areas at one time, and other areas at another time, rather, it is excluding areas to different degrees. That is, all the segmentations are subsets of each other and the interplay between the intensity of the seedpoint relative to the threshold ratio parameter controls just how far the front expands before stopping. This is the only source of inconsistency, which the confidence measure essentially eliminates.

# 6 Conclusions

Without the incorporation of anatomical knowledge or a modest amount of user guidance, it is unlikely that any segmentation method could capture the subtle nuances of stroke lesions in MR imagery. Common segmentation strategies using intensity and intensity changes are a solid starting point, but fall short when medical knowledge must come into play.

Our previous work implicitly incorporated human guidance. The process of drawing an initial contour that surrounds the lesion in a base slice implicitly directs the segmentation toward those edges near the initialization. The work presented in this paper can incorporate a similar amount of user direction, but performs well on average without it.

We have proposed methods for overcoming the variability of stroke lesions. The combination of scale-space and implicit contours can match the segmentations given by physicians. Even when a consistent strategy is chosen, the model hardly deviates from physician volumes.

Our confidence model provides an alternative method for choosing a segmentation strategy. When fully automatic, the intuitive fifty percent confidence strategy gives accurate lesion volumes on average, with a standard deviation only slightly above our clinical target. This method needs only two mouse clicks, which is much less user guidance than previously required. If a small amount of guidance is allowed, zero error can be achieved on almost every lesion. The amount of user guidance in the new method is substantially less than the old because it only requires to choose from a set of contours, rather than carefully

drawing an initial contour.

Because human segmentations are prone to variation, inconsistency, and can be taxing for those involved, it is unlikely we would have enough training data to do a fuller test of the system. However, since the model proposes contours with only a simple initialization step, the system could be tested on a wide scale by having several physicians select the most appropriate contour on a large number of lesions. Such information could eventually indicate whether constraining the model to a particular scale and threshold is the best strategy, or whether a different confidence level might be more appropriate on average. The authors presume it would indicate that no single parameter, whether scale, threshold, or confidence, will make the system robust under all lesion types. Rather, we feel it would verify that the problem can not be fully solved without medical knowledge. It might also illustrate the degree of subjectivity in "ground-truth:" even with automatic segmentations, doctors may individually, if even consistently, prefer different confidence levels in ambiguous lesions.

The statistics of sets is a relatively unexplored area. Along with a new confidence-based segmentation model, we also presented a new direction for thinking about the variability in segmentations. By capturing the fundamental difference between segmentations, rather than volumes, we give a set "statistic" that is similar to standard statistics of volume measurements when the segmentations are similar. Future research may explore the relationship between the usual variance and the set variance, especially as to how they will differ as segmentations contain more discrepancies. Perhaps more interesting is the notion of a "mean" set, which remains to be defined. Like set variance, set mean might relate to a traditional statistical analog, such as the value minimizing the expected squared deviation from the population or sample.

Another possible avenue for investigation is a directional diffusivity function (or the use of an existing one) that meets the continuity and differentiability scale-space requirements yet also reduces the rate at which fine structures are blurred and is not inferior to isotropic diffusion. Such a model might prove more useful in image processing contexts, where how the image looks after diffusion is more important.

We have introduced a model for segmenting lesion images that requires little user input. Clinically useful bounds can be achieved even when all parameters are constrained, but some types of lesions to do not lend themselves to accurate segmentation with a single parameter set. Thus, even more precise volumes may be obtained if a user selects from some candidate segmentations.

# Acknowledgments

This work was supported in part by the Collaborative Biomedical Research Program between the Baystate Medical Center and the Biotechnology Program at the University of Massachusetts.

# References

[1] D. Adalsteinsson and J.A. Sethian. A fast level set method for propagating interfaces. *J. Comp. Phys.*, 118:269–277, 1995.

[2] Luis Alvarez, Frédéric Guichard, Pierre-Louis Lions, and Jean-Michael Morel. Axioms and fundamental equations of image processing. *Arch. Rational Mech. Anal.*, 123:199–257, 1993.

[3] A.H. Anderson and J.E. Krisch. Analysis of noise in phase contrast MR imaging. *Med. Phys.*, 23(6):857–869, 1996.

[4] Michael J. Black, Guillermo Sapiro, David Marimont, and David Heeger. Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, 7(3):421–432, 1998.

[5] Francine Catté, Pierre-Louis Lions, Jean-Michel Morel, and Tomeu Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.*, 29(1):182–193, Feb 1992.

[6] Hwan Soo Choi, David R. Yahnor, and Yongmin Kim. Partial volume tissue classification of multichannel magnetic resonance images—a mixel model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(3):395–407, 1991.

[7] James J. Clark. Singularity theory and phantom edges in scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):720–727, 1988.

[8] Laurent D. Cohen and Isaac Cohen. Finite element methods for active contour models and balloons for 2D and 3D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1131–1147, 1993.

[9] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359, 1996.

[10] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[11] L. Florack, B. ter Haar Romeny, J. Koenderink, and M. Viergever. Linear scale-space. *Journal of Mathematical Imaging and Vision*, 4:325–351, 1994.

[12] Guido Gerig, Olaf Kübler, Ron Kikinis, and Ferenc A. Jolesz. Nonlinear anisotropic filtering of MRI data. *IEEE Transactions on Medical Imaging*, 11(2):221–232, 1992.

[13] Guy Gilboa, Yehoshua Y. Zeevi, and Nir A. Sochen. Complex diffusion processes for image filtering. In Michael Kerckhove, editor, *Scale-Space and Morphology in Computer Vision*, volume 2106 of *Lecture Notes in Computer Science*, pages 299–307. Springer, Jul 2001.

[14] Hákon Gudbjartsson and S. Patz. The Rician distribution of noisy MRI data. *Magn. Reson. Med.*, 34:910–914, 1995.

[15] Richard Hicks, M.D. Personal communication, August 2002.

[16] T. Iijima. Basic theory of pattern normalization (for the case of a typical one-dimensional pattern). *Bulletin of the Electrotechnical Laboratory*, 26:368–388, 1962.

[17] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331, 1987.

[18] Bernd Kawohl and Nikolai Kutev. Maximum and comparison principles for anisotropic diffusion. Preprint, University of Cologne, Mathematical Institute, University of Cologne, Weyertal 86-90, D-50931 Cologne, Germany, 1996.

[19] Satyanad Kichenassamy. The Perona-Malik paradox. *SIAM J. Appl. Math*, 57(5):1328–1342, 1997.

[20] Jan J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[21] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *J. of Applied Statistics*, 21(2):224–270, 1994. (Supplement on Advances in Applied Statistics: Statistics and Images: 2).

[22] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. International Series in Engineering and Computer Science. Kluwer, 1994.

[23] Dima Lisin, Edward Riseman, Gary Whitten, Donald Geman, Joseph Horowitz, and Benjamin Stein. Development of statistical and computer vision techniques to support the clinical study of ischemic stroke treatment. Technical Report UM-CS-2001-019, University of Massachusetts-Amherst, Computer Science Research Center, University of Massachusetts, Amherst, MA 01003-4601, 2001.

[24] Ravikanth Malladi, J.A. Sethian, and B.C. Vemuri. Shape modeling with front propagation: A level set approach. *Institute of Electrical and Electronics Engineers*, 17:158–175, 1995.

[25] Pavel Mrázek. Monotonicity enhancing nonlinear diffusion. *Journal of Visual Communication and Image Representation*, 13:313–323, 2002.

[26] Pavel Mrázek and Mirko Navara. Consistent positive directional splitting of anisotropic diffusion. In Bostjan Likar, editor, *Proceedings of Computer Vision Winter Workshop*, pages 37–48, 2001.

[27] B.W. Murphy, P.L. Carson, J.H. Ellis, Y.T. Zhang, R.J. Hyde, and T.L. Chenevert. Signal-to-noise measures of magnetic resonance imagers. *Magnetic Resonance Imaging*, 11:425–428, 1993.

[28] Mark Nitzberg and Takahiro Shiota. Nonlinear image filtering with edge and corner enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):826–833, 1992.

[29] S. Osher and J.A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *J. Comp. Phys.*, 79:12–49, 1988.

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

[31] Pietro Perona, Takahiro Shiota, and Jitendra Malik. Anisotropic diffusion. In Bart M. Ter Haar Romeney, editor, *Geometry-driven diffusion in computer vision*, pages 73–92. Kluwer Academic Publishers, 1994.

[32] Hanno Scharr and Joachim Weickert. An anisotropic diffusion algorithm with optimized rotation invariance. In *Mustererkennung 2000*, pages 460–467. Springer, 2000.

[33] J.A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93:1591–1595, 1996.

[34] Jan Sijbers, Arnold J. den Dekker, Paul Scheunders, and Dirk Van Dyck. Maximum-likelihood estimation of Rician distribution parameters. *IEEE Transactions on Medical Imaging*, 17(3):357–361, 1998.

[35] Jan Sijbers, Arnold J. den Dekker, Marleen Verhoye, and Dirk Van Dyck. Adaptive anisotropic noise filtering for magnitude MR data. *Magnetic Resonance Imaging*, 17(10):1533–1539, 1999.

[36] Jon Sporring and Joachim Weickert. On generalized entropies and scale-space. Technical Report DIKU-96/37, University of Copenhagen, Dept. of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copehnhagen, Denmark, 1996.

[37] Jon Sporring and Joachim Weickert. Information measures in scale-spaces. *IEEE Trans. Information Theory*, 45(3):1051–1058, 1999.

[38] Ben Stein, Dima Lisin, Joseph Horowitz, Edward Riseman, and Gary Whitten. Statistical and deformable model approaches to the segmentation of MR imagery and volume estimation of stroke lesions. In Wiro J. Niessen and Max A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention*, volume 2208 of *Lecture Notes in Computer Science*, pages 829–836. Springer, Oct 2001.

[39] Benjamin Reece Stein. *Signal Formulation, Segmentation, and Lesion Volume Estimation in Magnetic Resonance Images*. PhD thesis, University of Massachusetts-Amherst, University of Massachusetts, Amherst, MA 01003-4601, 2001.

[40] B. ter Haar Romeny, L. Florack, J. Koenderink, and M. Viergever, editors. *Scale-Space Theory in Computer Vision*, volume 1252 of *Lecture Notes in Computer Science*. Springer, 1997.

[41] Rein van den Boomgard. *Algorithms for Non-Linear Diffusion*, 2001. http://carol.wins.uva.nl/ rein/nldiffusionweb/material.html.

[42] Joachim Weickert. Scale-space properties of nonlinear diffusion filtering with a diffusion tensor. Technical Report 110, University of Kaiserslautern, Laboratory of Technomathematics, University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern, Germany, 1994.

[43] Joachim Weickert. Nonlinear diffusion scale-spaces: From the continuous to the discrete setting. In Marie Odile Berger, Rachid Deriche, Isabelle Herlin, Jerome Jaffré, and Jean-Michel Morel, editors, *12th International Conference on Analysis and Optimization of Systems: Images, Wavelets and PDEs*, volume 219 of *Lecture Notes in Control and Information Sciences*, pages 111–118, New York, Jun 1996. Springer-Verlag.

[44] Joachim Weickert. Theoretical foundations of anisotropic diffusion in image processing. In *Computing Supplement*, volume 11, pages 221–236. Springer, Wien, 1996.

[45] Joachim Weickert. Recursive separable schemes for nonlinear diffusion filters. In ter Haar Romeny et al. [40], pages 260–271.

[46] Joachim Weickert. A review of nonlinear diffusion filtering. In ter Haar Romeny et al. [40], pages 3–28.

[47] Joachim Weickert. *Anisotropic diffusion in image processing*. ECMI. Teubner-Verlag, Stuttgart, Germany, 1998. Out of print.

[48] Joachim Weickert. Coherence-enhancing diffusion filtering. *International Journal of Computer Vision*, 31(2/3):111–127, 1999.

[49] Joachim Weickert and Brahim Benhamouda. Why the Perona-Malik filter works. Technical Report DIKU-97/22, University of Copenhagen, Dept. of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copehnhagen, Denmark, 1997.

[50] Joachim Weickert, Bart M. ter Haar Romeny, and Max A. Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Transactions on Image Processing*, 7(3):398–410, 1998.

[51] Ross T. Whitaker and Stephen M. Pizer. A multi-scale approach to nonuniform diffusion. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 57(1):99–110, 1993.

[52] Andrew P. Witkin. Scale-space filtering. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, volume 2, pages 1019–1022, 1983.

# Appendix A

The following is code for MATLAB 6 that implements the additive operator splitting scheme for isotropic diffusion in three dimensions. It is modeled on the pseudo-code detailing the algorithm in [50] and Matlab code for an explicit scheme in [41].

```
function u = snld3D( f, lambda, m, sigma, dim, tau, k )
% SNLD3D Scalar (Isotropic) Nonlinear Diffu-
sion in 3D
%
%       U = SNLD3D( f, lambda, m, sigma, dim, tau, k )
%
% f := input image
% lambda := constrast parameter
% m := diffusitivity decay
% sigma := regularization scale
% dim := relative coordinate dimension size
% tau := stepsize
% k := numsteps
%
% Uses Weickert's isotropic edge enhancing diffu-
sion filter.

  Cm = CMconstant(m);

Cm = -lambertw(-1,-1/(2*m) * exp(- 1/(2*m))) -
1/(2*m);

  u = f;

  for i = 1:k
    % Gradient
    ux = gD3D( u, sigma, dim, 1, 0, 0 );
    uy = gD3D( u, sigma, dim, 0, 1, 0 );
    uz = gD3D( u, sigma, dim, 0, 0, 1 );

    % Gradient Magnitude Squared
    uw2 = ux.^2 + uy.^2 +uz.^2;

    % Prevent Division by Zero
    uwPOS = (uw2>0);
    uw2NZ  = uw2.*uwPOS + ones(size(uw2)).*~uwPOS;

    g = ones(size(uw2)) - uwPOS.*exp(-
```

```
Cm./((uw2NZ/lambda).^m));

    u = aosstep3D( u, tau, dim, g );

  end


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function g = gD3D( f, sigma, dim, ox, oy, oz )
% GD3D Gaussian (Derivative) Convolution in 3D
%
% G = GD3D(f, sigma, dim, ox, oy, oz)
%
% sigma := scale
% dim := relative coordinate dimension size
% ox := order of derivative wrt x (cols)
% oy := order of derivative wrt y (rows)
% oy := order of derivative wrt z (slices)

% Modeled on 2D code by R. van den Boomgaard

        K = ceil( 3 * sigma );
        x = -K:K;

        Gsx = gDeriva-
tive( ox, x.*dim(1), gauss(x.*dim(1),sigma), sigma );
        Gsy = gDeriva-
tive( oy, x.*dim(2), gauss(x.*dim(2),sigma), sigma );
        Gsz = gDeriva-
tive( oz, x.*dim(3), gauss(x.*dim(3),sigma), sigma );


        g = conv3Brd( f, Gsx, Gsy, Gsz );

function r = gDerivative( order, x, Gs, sigma )
    switch order
    case 0
        r = Gs;
    case 1
        r = -x/(sigma^2) .* Gs;
    case 2
        r = (x.^2-sigma^2)/(sigma^4) .* Gs;
    otherwise
        error('only derivatives up to second or-
```

```
der are supported’);
    end
    r = r/(sum(abs(r));
function gs = gauss( x, sigma)
  gs = exp( - x.^2 / (2*sigma^2) );


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function g = conv3Brd( f, w1, w2, w3 )
% CONV3BRD Separable convolution with reflected bor-
ders in 3D
%
%    G = CONV3BRD( F, W1, W2, W3)
%
% f := input
% w1 := x direction (cols) kernel
% w2 := y direction (rows) kernel
% w3 := z direction (slices) kernel

% Modeled on 2D code by R. van den Boomgaard

    N = size(f,1);
    M = size(f,2);
    P = size(f,3);

    I = (size(w1(:),1)-1)/2;
    J = (size(w2(:),1)-1)/2;
    K = (size(w3(:),1)-1)/2;

    iind = min(max((1:(N+2*I))-I,1),N);
    jind = min(max((1:(M+2*J))-J,1),M);
    kind = min(max((1:(P+2*K))-K,1),P);

    fwb = f(:,jind,:);

    g1 = convn(fwb,w1,’valid’);

    fwb = g1(iind,:,:);
    clear g1;

    g2 = convn(fwb, w2’, ’valid’);

    fwb = g2(:,:,kind);
    clear g2;
```

```
    g=convn(fwb,reshape(w3,1,1,length(w3)),'valid');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function u = aosstep3D(f, tau, dim, g )
% AOSSTEP3D Additive Operator Splitting Step in 3D
%
%    U = AOSSTEP3D( F, TAU, DIM, G )
%
% f := input image
% tau := step size
% dim := relative coordinate dimension (voxel) size
% g := diffusivity

  tol = 0; % set to eps to avoid divide by zero er-
rors/warnings
  d = 3; % number of dimensions
  cycle = [2 3 1];

  u = zeros(size(f));

  for k=1:d % loop over each dimension

    [N1, N2, N3] = size(f);

    % Current dimension size
    h1 = dim(k);

    alpha = zeros(N1,N2,N3);
    m = zeros(N1,N2,N3);
    l = zeros(N1-1,N2,N3);
    y = zeros(N1,N2,N3);
    w = zeros(N1,N2,N3);
     %-----------------------------------------------
-----------------------------
    % Diffusion along row dimension

    % Step Zero -- Create the tridiagonal matrix

    % single neighbor at 'left' border
    al-
pha(1,:,:) = 1 + d*tau*( g(1,:,:) + g(2,:,:) )/( 2 * h1^2 );

    % two neighbors
```

```
    alpha(2:N1-1,:,:) = 1 + d*tau* ...
        ( 2*g(2:N1-1,:,:) + g(1:N1-
2,:,:) + g(3:N1,:,:) )/( 2 * h1^2 );

    % single neighbor at 'right' border
    alpha(N1,:,:) = 1 + d*tau*( g(N1,:,:) + g(N1-
1,:,:) )/( 2 * h1^2 );


    gamma =  - d*tau*( g(1:N1-
1,:,:) + g(2:N1,:,:) )/( 2 * h1^2 );


    % Step One -- LR Decomposition

    m(1,:,:) = alpha(1,:,:);

    for i=1:N1-1

      l(i,:,:) = gamma(i,:,:) ./ ( m(i,:,:) + tol);

      m(i+1,:,:) = alpha(i+1,:,:) -
l(i,:,:) .* gamma(i,:,:);

    end;

    % Step Two -- Forward Substitution

    y(1,:,:) = f(1,:,:);

    for i=2:N1
      y(i,:,:) = f(i,:,:) - l(i-1,:,:) .* y(i-
1,:,:);
    end;

    % Step Three -- Backward Substitution

    w(N1,:,:) = y(N1,:,:) ./ ( m(N1,:,:) + tol );

    for i=N1-1:-1:1
      w(i,:,:) = ( y(i,:,:) -
gamma(i,:,:).*w(i+1,:,:) ) ./ ( m(i,:,:) + tol );
    end;
```

```
    % cycle input image and diffusivity dimensions
    f = permute(f,cycle);
    g = permute(g,cycle);

    % inversely cycle w back to original dimen-
sion order
    for j=1:k-1
      w = ipermute(w,cycle);
    end;

    u = u + w;

  end;

  u = (1/d) * u;
```