

# Optimizing the Use of Radiologist Seed Points for Improved Multiple Sclerosis Lesion Segmentation

Jon McAusland, Roger C. Tam\*, Erick Wong, Andrew Riddehough, and David K. B. Li

**Abstract**—Many current methods for multiple sclerosis (MS) lesion segmentation require radiologist seed points as input, but do not necessarily allow the expert to work in an intuitive or efficient way. Ironically, most methods also assume that the points are placed optimally. This paper examines how seed points can be processed with intuitive heuristics, which provide improved segmentation accuracy while facilitating quick and natural point placement. Using a large set of MRIs from an MS clinical trial, two radiologists are asked to seed the lesions while unaware that the points would be fed into a classifier, based on Parzen windows, that automatically delineates each marked lesion. To evaluate the impact of the new heuristics, an interactive region-growing method is used to provide ground truth and the Dice coefficient (DC) and Spearman's rank correlation are used as the primary measures of agreement. A stratified analysis is performed to determine the effect on scans with low-, medium-, and high lesion loads. Compared to the unenhanced classifier, the heuristics dramatically improve the DC (+32.91 pt.) and correlation (+0.50) for the scans with low lesion loads, and also improve the DC (+14.55 pt.) and correlation (+0.15) for the scans with medium lesion loads, while having a minimal effect for the scans with high lesion loads, which are already segmented accurately by Parzen windows. With the heuristics, the DC is close to 80% and the correlation is above 0.9 for all three load categories.

**Index Terms**—Magnetic resonance imaging (MRI), multiple sclerosis (MS), seed points, segmentation, white matter lesions (WMLs).

## I. INTRODUCTION

MRI is an extremely sensitive method for visualizing white matter lesions (WMLs) that are associated with a broad spectrum of pathology ranging from normal aging to neurode-

generative diseases, such as multiple sclerosis (MS). The volumetry of WMLs is used as a biomarker in many types of neurological studies [1] and as a result, many methods for the segmentation of WMLs on MRI have been proposed (e.g., [2]–[11]). These methods range from those that require the operator to visually check or interact with every lesion [3], [4], [9] to methods that are largely or fully automatic [5]–[8], [10], [11].

The intensity range of WMLs on MRI typically overlaps with those of healthy tissues, making the inclusion of all lesion voxels without introducing “false positive” regions a difficult challenge. To reduce misclassification, the majority of WML segmentation techniques rely on some input from a human expert who is ideally a radiologist with experience in the particular pathology being studied. This input is most commonly in the form of sample points placed on lesions [3], [9] and in some cases several types of healthy tissue as well [2], [4], [7], [8]. The most interactive of these techniques uses the sample points as seed points from which each lesion is grown under manual control, but using some measure of voxel coherence for guidance, such as local intensity difference [3] or fuzzy connectivity [9]. Other more automated methods use the sample points to compute the candidate regions without further manual interaction [2], [4], but may require the manual identification of false positives as a postprocessing step [4]. There are a number of methods that employ an extrapolative approach in which the radiologist places sample points on a set of training images to tune the algorithm, which can then be applied in a fully automatic fashion to new scans acquired with similar MRI parameters [7], [8]. A few algorithms require no manual input at all to identify the lesions [6], [10], [11], but in some cases, they reduce misclassification by strategies that would bias their suitability against certain pathology. For example, the systematic exclusion of detected WMLs between the lateral ventricles [6] would reduce the effectiveness of the method for analyzing MS lesions, which are common in the corpus callosum [12]. In other cases, the algorithms are either mathematically complex parametric methods [10] or require processing steps that are difficult to optimize, such as nonrigid registration to atlases [11]. As a group, the completely automatic methods have the potential for future applicability, but remain mostly unvalidated for large multicenter studies.

A common approach shared by many of the current techniques that use manually placed points is to apply a non-parametric method of density estimation, such as the Parzen windows [13] or the  $k$ -nearest neighbors ( $k$ -NN) [14], to the samples to compute a probability density function (PDF) of each tissue type [2], [4], [7], [8]. The PDFs express the probability that a voxel with a given intensity value belongs to a certain class. These probabilities are used to compute a statistical

Manuscript received February 24, 2010; revised May 4, 2010; accepted June 18, 2010. Date of publication July 1, 2010; date of current version October 15, 2010. This work was supported by the Natural Sciences and Engineering Research Council of Canada. J. McAusland and R. C. Tam contributed equally to this manuscript. *Asterisk indicates corresponding author.*

J. McAusland, E. Wong, and A. Riddehough are with the Multiple Sclerosis/Magnetic Resonance Imaging Research Group, Division of Neurology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada (e-mail: jon@msmri.medicine.ubc.ca; erick@msmri.medicine.ubc.ca; andrew@msmri.medicine.ubc.ca).

\*R. C. Tam is with the Multiple Sclerosis/Magnetic Resonance Imaging Research Group, Division of Neurology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada, and also with the Department of Radiology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: roger@msmri.medicine.ubc.ca).

D. K. B. Li is with the Multiple Sclerosis/Magnetic Resonance Imaging Research Group, Division of Neurology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada, and also with the Department of Radiology, University of British Columbia, Vancouver, BC V6T 1Z4 Canada (e-mail: david.li@ubc.ca).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TBME.2010.2055865

classification that is sometimes used as the final output [2], but due to the common problem of overlapping distributions, the classification is typically refined further with spatial processing, which can include morphological operations [7] or atlas-based filtering [8].

The majority of methods that use radiologist-placed sample points require the expert to place the points in a way that satisfies the processing constraints of the segmentation algorithm, but is not necessarily intuitive or efficient for the expert. For example, to ensure computational stability, some methods ask the user to place a fixed number of lesion points per region or scan, regardless of the actual number of lesions [2], [4], [8]. These algorithms also typically require that samples from five to seven healthy tissue classes be placed in addition to the lesion class, which can be a distraction from the already difficult task of identifying WMLs, especially for scans with a large number of abnormalities. In some cases, the expert is also asked to perform particularly challenging tasks, such as placing points on the partial volume areas on the boundaries of healthy tissues to simulate WML intensities [4] or applying different types of markers for the same tissue class depending on the perceived intensity [2]. In addition to requiring the expert to perform tasks that may be unintuitive, most algorithms also assume that the positions of the resulting points are optimal for subsequent PDF estimation, even though for small lesions, minor differences in placement can have a significant effect on the estimated distributions. One method attempts to account for placement errors by clustering the samples by intensity then excluding outliers [2], but this can remove potentially valuable information.

This paper examines how expert-placed points can be processed to improve the accuracy of MS lesion segmentation without placing additional burden on the radiologist. Using T2-weighted (T2w) and proton density-weighted (PDw) MRIs, the radiologist is asked to place sample points as visual indicators of the location and extent of the pathological areas, but does not need to be concerned about the actual number of points being placed or their underlying intensities. This freedom is achieved by three heuristics: 1) correct for positional errors in the points; 2) dynamically adjust the number of sample points used for PDF estimation; and 3) refine the shape of each lesion by modeling some basic elements of visual shape perception. In our experiment, the radiologist seeds each lesion, but the rest of the processing pipeline, including the estimation of the normal tissue PDFs, is fully automatic. We choose to use this level of interactivity instead of a more automated approach, such as only asking the expert to label a smaller set of training images that can then be used to compute a lesion atlas, for two reasons. The first is that we would like to observe the effects of the implemented heuristics as directly as possible, without having to be concerned about confounding issues, such as the accuracy of image registration to the atlas [15]. The second reason is that although MRI is very sensitive to white matter (WM) abnormalities, it lacks specificity and, therefore, expert knowledge is required to distinguish MS lesions from other pathology, such as vascular disease [16]. Applying the proposed algorithm to a large set of MRIs from a multicenter MS clinical trial, we are able to demonstrate that the heuristics have a strong positive

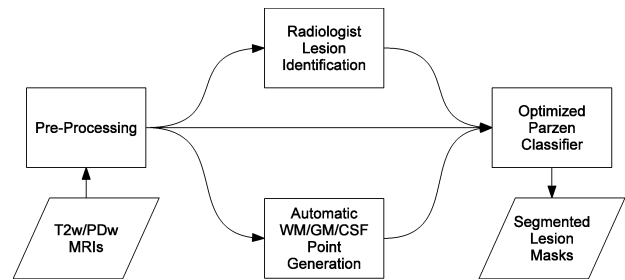


Fig. 1. Lesion segmentation pipeline.

effect on segmentation performance, particularly in the scans with lower lesions loads, for which most current methods show reduced accuracy [5]–[8] when evaluated with relative measures that normalize the amount of overlap or error with the overall lesion volume, such as the commonly used Dice coefficient (DC) (defined in Section II-F).

## II. METHODOLOGY

The lesion segmentation pipeline for testing our optimization heuristics is composed of several main components, as illustrated in Fig. 1. As described in detail in the following sections, a Parzen window classifier takes as input the radiologist-placed lesion points as well as automatically generated sample points of normal-appearing WM, grey matter (GM), and cerebrospinal fluid (CSF). Two of the optimization heuristics are incorporated into the Parzen classifier, whereas the third is applied afterward.

The test dataset was collected for a clinical trial in relapsing–remitting MS and is composed of 234 PDw/T2w MRI pairs (468 individual volumes) from 29 scanning sites. The study protocol for this data was approved by the Institutional Ethics Review Board. The scans were acquired with a dual-echo sequence so that each PDw/T2w pair is inherently registered. The image dimensions are  $256 \times 256 \times 50$  voxels with voxel size  $0.937 \times 0.937 \times 3.0$  mm. The echo time (TE) and repetition time (TR) values vary depending on the site:  $TE_1 = 8.4\text{--}28.1$  ms,  $TE_2 = 61.3\text{--}96.0$  ms, and  $TR = 2500\text{--}3000$  ms. Two radiologists who are highly experienced in MS lesion identification participated in this study. Each radiologist was assigned a set of scans on which they worked independently.

### A. Preprocessing

Each MRI undergoes several preprocessing steps before lesion identification and segmentation are performed. The first is the correction of MR intensity inhomogeneity using a multiscale version [17] of the nonparametric nonuniform intensity normalization (N3) method [18]. The next step is the application of a structure-preserving noise-removal filter named SUSAN [19], applied with conservative parameters that have been empirically determined to improve the visual quality of the images without compromising the appearance of subtle lesions. For the automatic generation of nonlesion sample points and Parzen window classification, the MRI is further processed by removing all nonbrain tissues with the brain extraction tool (BET) [20]. N3,

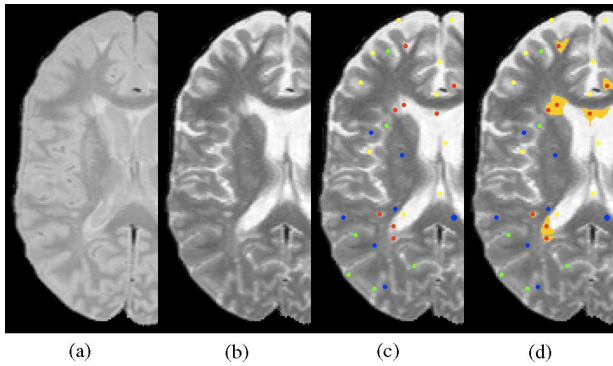


Fig. 2. (a) PDw image. (b) T2w image. (c) Automatically generated sample points (blue = WM, green = GM, and yellow = CSF) and radiologist-placed lesion points (red). The points are enlarged in the image for increased visibility; the radiologists normally use smaller markers for greater precision. (d) T2w image with overlay of lesions segmented by Parzen windows.

SUSAN, and BET are all widely used and publicly available as free software.

### B. Lesion Point Placement

The lesion points are marked on each T2w/PDw scan pair using in-house software that allows the user to view and place points on either image, with the resulting points displayed on both echoes. The interface also provides interslice context by displaying the current slice with the two adjacent slices simultaneously. The placing of the lesion points by the radiologists is guided by a minimal set of instructions that are meant to allow the radiologists to work efficiently and intuitively. Working on one slice at a time, but using the two adjacent slices for reference, the radiologists are asked to place the points as visual indicators of where the lesions are, so that another user, typically a trained technician, would take the points and apply a semi-automatic region-growing method (described in Section II-F) to segment each lesion. For this study, the radiologists were blinded from the fact that their input points would be used to automatically delineate the lesions, nor did they have access to image histograms or other intensity tools to aid them in the placement of the points. The guidelines given to the radiologists consist of three main points:

- 1) mark all focal WMLs consistent with MS pathology, regardless of their size;
- 2) place more than one point on a lesion, if they feel that the additional points would help a technician see the extent of the lesion;
- 3) place at least one point near the center of each lesion.

Fig. 2 illustrates an example of the placement of lesion points.

### C. Automatic Generation of WM, GM, and CSF Sample Points

The sample points for WM, GM, and CSF are generated automatically with an algorithm consisting of the following three main steps, which are performed on the entire scan.

- 1) An approximate CSF mask is computed. We employ a method based on one previously used to perform compu-

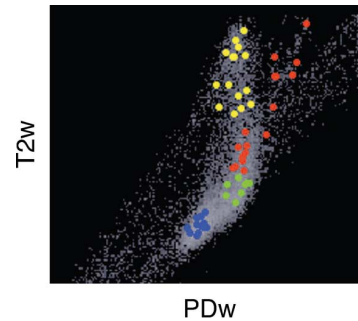


Fig. 3. Two-dimensional histogram of image intensities formed from a dual-echo PDw/T2w pair. The blue, green, and yellow points represent the automatically generated sample points for WM, GM, and CSF, respectively. The red points represent the lesion points placed by a radiologist on the MRIs.

tation of whole-brain volume [21], which has been shown to be stable in longitudinal studies of atrophic brains. First, a T2w/PDw ratio image of the intradural region is computed. In this image, the CSF is brighter than all other tissues. A single Gaussian curve is fitted to the histogram of the ratio image to capture the intensity range spanned by the WM and GM, using the Levenberg–Marquardt nonlinear optimization algorithm to minimize the sum of squared differences. All voxels that are higher than the center of the Gaussian by two standard deviations are classified as CSF.

- 2) The intensity cluster centers of WM and GM are computed from the ratio image excluding CSF by a modified C-means clustering method [22]. An anisotropic diffusion filter [23] is applied to the image before clustering to improve the overall smoothness of the intensity distribution of each tissue without introducing interclass blurring. The clustering algorithm is then initialized with three classes, representing WM, GM, and lesion. The role of the lesion center is to minimize the impact of any bright lesion pixels on the accuracy of the computed WM and GM centers, rather than to estimate the real mean lesion intensity. The WM and GM cluster centers are computed as per the traditional C-means method, but the proportion of lesion is highly variable; therefore, we cannot always assume that there are enough lesion pixels for their cluster center to be computed accurately. Instead, in each iteration of clustering, the lesion center is computed as a fixed offset of the maximum of the other two centers. A reasonable value has been empirically determined to be 30% and is fixed for all of our experiments. After clustering, the 2-D histogram space formed by the PDw and T2w intensities is computed (see Fig. 3), and two separating circles centered on the WM and GM cluster centers are drawn, with their radius being one-half of the distance between the two centers. The voxels within the intensity ranges defined by the circles form the WM and GM masks from which the sample points are taken.
- 3) Sample points for WM, GM, and CSF are randomly selected within the masks computed earlier. To ensure that the intensity distribution of each tissue type is well

represented, a minimum number of sample points of each type (15 each for WM and CSF, 8 for GM) are taken from each slice. In addition, the samples are spatially separated such that a minimum distance (15 mm) is enforced between any two samples of the same type. This distance was determined by searching for the value that maximizes mean spatial distribution with the given number of points, optimized over a large number of development scans.

#### D. Parzen Window Classification

Given a set of  $n$  sample pixels with intensities  $\{\mathbf{x}_i, i = 1, \dots, n\}$  and tissue labels  $\{y_i, i = 1, \dots, n\}$ , the Parzen window method [13] can be used to estimate an intensity PDF of each tissue. Note that each  $\mathbf{x}_i$  can be a vector, with dimensionality equal to the number of image types being used. Because we are using two types of scans (PDw and T2w), our sample points have 2-D intensity vectors. The Parzen window method is well known; therefore, rather than explaining it in detail, we summarize it by providing the equation that expresses the probability that a given intensity vector  $\mathbf{x}$  has tissue label  $y$

$$P(y|\mathbf{x}) = \frac{\sum_{i, y_i=y} K(\mathbf{x}, \mathbf{x}_i)}{\sum_{i=1}^n K(\mathbf{x}, \mathbf{x}_i)} \quad (1)$$

where  $K$  is a fixed-size kernel function, typically a Gaussian for MRI data, centered at each sample point. In all of our experiments, the radius of the kernel is fixed to be 10% of the dynamic range of the scan.

In our study, each voxel inside the skull is labeled as belonging to one of the four types: WM, GM, CSF, and lesion. For lesions that are close in intensity space to GM and/or CSF, the nonlesion sample points help to limit “leakage” of the lesions into normally appearing tissue. The classifier can be seen as attempting to find the optimal boundaries between the lesion cluster and the other clusters in the 2-D histogram space, as illustrated in Fig. 3.

Each slice is processed individually, with the sample points used for Parzen estimation taken from the current slice and the two adjacent slices. We use three slices to estimate a local distribution because most lesions are less than 10 mm in diameter [24], and three slices at 3 mm thickness would span most lesions centered on the middle slice. In addition, using a small slab helps to avoid inaccuracies in PDF estimation due to MR field inhomogeneity and natural variations in tissue intensity in different parts of the brain. However, if the number of lesion seed points in the slab is very low, the number of slices can be increased gradually until there are sufficient points. This procedure is explained in Section II-E2.

For each slice, all of the voxels that have the highest probability of being part of a lesion based on their intensity are grouped using connected component analysis, performed in 2-D due to the anisotropic nature of the scans. Any isolated region that does not contain a radiologist-placed seed point is removed as a false positive. Such regions are common in areas with partial volume.

#### E. Optimizing Use of Radiologist-Placed Seed Points

We propose three simple but effective heuristics that can be used individually or in combination to improve the accuracy of

MS lesion segmentation. Two of these heuristics are designed to improve the accuracy of the estimated PDFs by adjusting the position and the number of seed points used for the Parzen window computation, and the third is a purely spatial method that uses the corrected point positions to approximate visual shape partitioning to identify areas that are likely to be false positives.

1) *Positional Adjustment*: The first heuristic that we apply is a positional correction of each seed point using the intensities of the voxels in the surrounding neighborhood. The motivation for this adjustment is the observation that the typical MRI scans used for visualizing WMLs have voxels that are relatively large compared to many of the smaller MS lesions, causing many of these lesions to be imaged with a large percentage of partial volume. The result is that unless the radiologist can place a seed point precisely and consistently within the bright center spot of each small lesion, which would be impractically slow, if done on a large scale, it is likely that for scans with numerous small lesions, many of the seed points would be placed on voxels with a high degree of partial volume, which would incorrectly skew or shift the estimated lesion intensity distribution. To address this problem, we search the neighborhood of each seed point for the brightest voxel, which is likely to indicate the purest lesion content. The search region is defined as the set of voxels  $\{x_s + k_x + x_n, y_s + k_y + y_n \mid -r \leq x_n \leq r, -r \leq y_n \leq r\}$ , where  $(x_s, y_s)$  are the 2-D coordinates of the seed point,  $k_x$  and  $k_y$  specify a fixed offset to the center of the search region that accounts for a potential systematic bias, and  $\{x_n, y_n\}$  are the voxels that form a square neighborhood with a maximum distance of  $r$  from the offset center. The inclusion of the factors  $k_x$  and  $k_y$  is to account for a positional bias that we hypothesize may exist due to a natural tendency of many computer mouse users to place the pointer immediately next to small targets rather than directly on top. After experimenting with a number of different values for  $k_x$ ,  $k_y$ , and  $r$ , and searching for the intensity maxima in the PDw and T2w values individually as well as their summed values, we have found that using the PDw image alone with  $k_x = -1$ ,  $k_y = 0$ , and  $r = 1$  (window size of  $3 \times 3$  pixels) yields the best overall results from our data used for development, and we have kept these parameters consistent for all of our experiments. However, it should be noted that using a larger window of size  $5 \times 5$  centered on the original radiologist-placed point yields results that are only slightly less than optimal. The value  $k_x = -1$  indicates that the radiologists have a tendency to place lesion points one pixel to the right of the intended target, possibly because of eye and/or hand dominance. A possible explanation for why using the PDw alone gives the best overall results is that in the case of periventricular lesions, relocating to the brightest pixel in a T2w image may land the seed point inside CSF.

2) *Sample Size Adjustment*: The Parzen window method aims to estimate a continuous distribution function from discrete sample points and, therefore, the number of points used can strongly influence the computed result. As explained in Section II-D, our Parzen window algorithm uses a contiguous slab of three slices to estimate the lesion PDF of the center slice. Using a small slab helps to avoid inaccuracies in PDF

estimation due to spatially varying intensity differences, but problems can arise, if the slab contains too few lesion points to form a representative sample, such as when only the partial volume region of a lesion is overlapping the slab, which can cause the classifier to mislabel healthy tissue as lesion. To ensure that there are sufficient lesion points whenever possible, we enforce a lower bound on the number of points by enlarging the slab on both sides until the threshold is reached or we run out of slices. This approach has the same benefit as the  $k$ -NN method that uses a fixed number of nearest neighbors in that a certain level of stability can be expected of the classification process, because they both ensure that a minimum number of samples are used, but our method has the additional advantage of allowing a greater number of samples to be used, if they are available in the current slab. This dynamic adjustment of sample size by adding slices to the slab is designed to strike a balance between providing maximum information to the Parzen classifier and minimizing the effects of inhomogeneity over larger regions. Although increasing the number of slices may artificially widen the estimated PDF, this effect should be slight compared to the impact of insufficient samples, especially because the slab is enlarged only to the extent necessary. After experimenting with a lower bound ranging from 1 to 10 sample points (we saw no improvements beyond 10), we have found that a minimum of five points give the best overall results from our development data, and we have kept this parameter consistent for all of our experiments.

3) *Shape Refinement*: In addition to the two aforementioned heuristics, we explored the possibility of using the corrected sample point positions to refine the shape of the lesions to reduce false positives. The main idea is that some basic concepts of shape perception can be used to model how an observer would visually partition a 2-D region with distinguishable components into subregions, and that this decomposition can be used to remove parts of the segmented area that are unlikely to be part of the true lesion. Our hypothesis is that a radiologist would naturally place at least one seed point in each visually distinct subregion, and that the exclusion of subregions without seed points would result in an overall improvement in accuracy by reducing the false-positive rate. It has been established that sharp bends and narrowings are both important perceptual cues for partitioning shapes and that these properties are represented psychologically by an object's axis of symmetry, or medial axis, and its closest distance to the boundary [25]. We propose that dividing each segmented region at angled narrowings would significantly reduce the amount of misclassification. In particular, we have observed that many false-positive regions are in the form of relatively thin structures, such as cortical GM or partial volume at the WM/CSF interface that snake away from a true lesion, as they follow the curves of the cortex or ventricles. Fig. 4 shows an example of a periventricular lesion with a long false-positive tail that bends around the third ventricle. To approximate the proposed partitioning scheme without actually computing a medial representation, which would be computationally expensive and sensitive to noise, we apply a visibility test from each radiologist point in a segmented region to all interior voxels of the region and remove any voxels that do

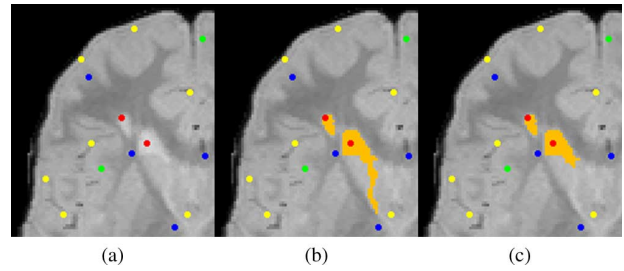


Fig. 4. Example of lesion shape refinement. (a) PDw scan with sample points (blue = WM, green = GM, and yellow = CSF, red = lesion). (b) Lesions segmented with Parzen windows. The larger lesion shows a long false-positive tail. (c) Shape refinement heuristic clips the tail to give a truer result.

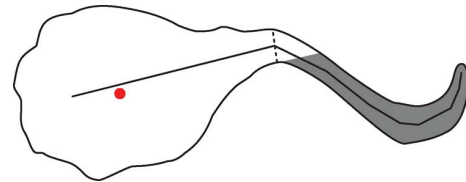


Fig. 5. Approximating visual shape partitioning using line-of-sight from radiologist point. The sample point (red) is typically placed close to the medial axis (solid line), and a visibility test can be used to approximate a partitioning based on the angle of the medial axis and its closest distance to the boundary. The dotted line shows the most natural division, of which the shaded region is an approximation computed from the visibility test. In this case, the tail would be removed as a false positive because it does not contain a lesion point.

not connect to any radiologist points via a straight line without crossing the boundary (see Fig. 5). The rationale is that the radiologists usually place at least one point near the center of each lesion, because they are asked to do so and because of the axis of symmetry's strong relevance to shape perception, and a visibility test from the center is a good first-order approximation of the shape continuity between adjoining components because the bend angle and narrowness of the junction are both taken into account. Fig. 4 shows an example of how the false-positive tail of a segmented region is removed using this method. Although this heuristic is only an approximation to visual shape partitioning, it is quick to implement and provides insight into the feasibility of a shape perception-based approach.

## F. Validation

For validation purposes, an existing semiautomatic 2-D region-growing technique was applied to segment all lesions in the test scans. In contrast to the Parzen window classifier used in combination with the proposed heuristics, the semiautomatic tool requires that each lesion be interactively grown from the seed points. This software has been used successfully in a number of large MS clinical trials (e.g., [26] and [27]), and five experienced technicians who have undergone an extensive and externally audited training procedure were asked to analyze the data in this study. The radiologists were not included in the validation process primarily to avoid influencing the way they place the seed points so that they can maintain their neutrality for possible future experiments.

During interactive segmentation, a technician initiates the growing process for each lesion from the one or more seed

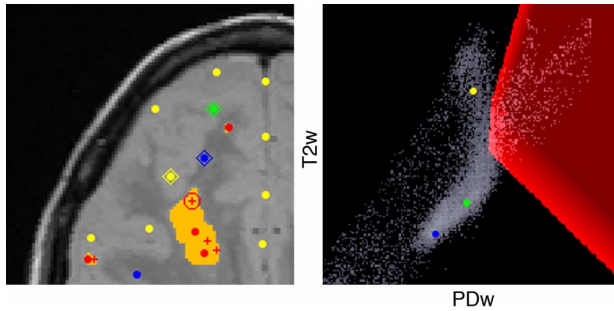


Fig. 6. Semiautomatic method used to produce validation data. On the left is a PDw scan with sample points (blue = WM, green = GM, yellow = CSF, red = lesion). The technician first tries to grow the lesion from the radiologist points (red dots) only, but if the resulting region does not cover the lesion adequately, the technician can add more lesion points, as indicated by the red +s. When the technician selects a red point or + from which to start growing (circled red +), the three sample points representing WM, GM, and CSF that are closest to the lesion point, as indicated by the blue, green, and yellow dots enclosed in a diamond shape, are used to constrain the growth. The orange area is the grown lesion. The right image shows the T2w/PDw histogram with the selected lesion point and the closest WM, GM, and CSF points displayed. The red area shows the intensity space of the grown region allowed from this seed point, as it is constrained by proximity in intensity space to the healthy sample points.

points placed by the radiologist. The operator is aided by an algorithm that uses a local subset of the nonlesion (WM, GM, and CSF) sample points generated by the same automatic method as described in Section II-C to constrain the growing process. For each seed point, the three points representing the closest WM, GM, and CSF sample points are used to restrict the grown region, which is only allowed to include voxels that are closer in 2-D intensity space to the current seed point than the three nonlesion points. The operator has full control of the growing process in that if a lesion is too large or small with the current constraints, he or she can add or delete lesion, WM, GM, or CSF points and regrow the lesion as needed. Therefore, the sample points are only meant as a guide. The 2-D histogram is also displayed to the operator to aid the editing of points. At any one time, only four points are used to constrain the growing process because this allows the operator to predict with some accuracy the effect of adding or deleting a particular point, and also see the change in the segmentation in real time. Fig. 6 illustrates an example in which the operator has added points to fill out a lesion whose expansion is limited by nearby nonlesion points. The operator can also manually draw line segments along the perceived lesion boundary to prevent growing into an adjacent region, but this is done only as a last resort.

To determine the impact of each heuristic on the accuracy of lesion segmentation, we measure the differences between the lesion regions produced by the Parzen window method and those produced by trained technicians using the region-growing method, which serves as the reference standard and the closest available ground truth. The primary measure of accuracy that we employ is the DC [28], which computes a normalized overlap value between the automatic and manual methods as follows:

$$DC = \frac{2 \times (S_{\text{Parzen}} \cap S_{\text{manual}})}{S_{\text{Parzen}} + S_{\text{manual}}}$$

While the DC gives a good idea of overall accuracy, it is often useful to have supplementary measures that provide information on specific aspects of segmentation performance. To determine the extent of the regions falsely classified as lesion by the Parzen window method (i.e., lesions “leaking” into healthy regions), we use the fraction of overestimation (FOE)

$$FOE = \frac{S_{\text{Parzen}} \cap S_{\text{manual}}}{S_{\text{manual}}}$$

To measure the extent of the regions classified as lesion by an operator but missed by the Parzen window method, we use the fraction of underestimation (FUE)

$$FUE = \frac{\overline{S_{\text{Parzen}} \cap S_{\text{manual}}}}{S_{\text{manual}}}$$

Looking at the FOE and FUE together is more informative than either alone and gives an idea of whether an algorithm tends toward overestimation or underestimation, which can be used to explain the DC score. The DC and FUE range between 0 and 1, whereas the FOE can exceed 1, and we typically multiply all three values by 100 to obtain percentage values, mostly for legibility when the numbers are displayed with high precision.

The DC, FUE, and FOE are all relative measures in that the overlap or disagreement is normalized to the lesion load of each scan. The use of relative measures is appropriate for this application because small lesions need to be segmented more accurately in order for the scans with low overall lesion volume, which as explained in Section III tend to have smaller lesions, to have a true volume rank within the population being studied so as to facilitate a meaningful correlation to clinical assessments.

In addition to the DC, it would also be informative to have another primary measure that examines the effect of the heuristics on the total lesion volume of each scan relative to the population. For this purpose, we compute the Spearman’s rank correlation coefficient ( $\rho$ ) between the automatically derived values and the reference values. A correlation coefficient close to unity would suggest that the automatic method can be used in place of the interactive method for the purpose of ranking the scans according to lesion load.

### III. RESULTS

To examine the results, we have stratified the scans into three equally sized groups or tertiles, based on their reference lesion loads. Table I lists the means and standard deviations of the three tertiles. Overall, the dataset represents a wide range of lesion load. The table also shows the summary statistics for the lesion size (as produced by the expert segmentation) and number of distinct lesions, also stratified by lesion load. The reason for breaking the dataset down into lesion load categories is that scans with low lesion loads tend to be more challenging to segment accurately for most current methods (e.g., [5]–[8]) because the lesions tend to be smaller and in lower numbers. Smaller lesions tend to be more subtle in terms of intensity because of pathological state and partial volume, and this combined with a limited sample size makes obtaining a representative and distinct intensity distribution more difficult. For the current dataset, the lesion load and mean lesion size are strongly correlated (0.84), as are

TABLE I  
LESION VOLUME, MEAN LESION SIZE, AND LESION COUNT FOR THREE  
TERTILES OF SCANS AS STRATIFIED BY LESION LOAD

	$q1$	$q2$	$q3$
Volume (mm <sup>3</sup> )			
Mean	498.15	2233.92	9255.31
Standard Deviation	306.64	896.40	5524.59
Mean Lesion Size			
Mean	36.48	50.23	92.80
Standard Deviation	14.25	14.44	50.39
Number of Distinct Lesions			
Mean	32.19	100.19	229.81
Standard Deviation	18.17	46.28	98.50

TABLE II  
SEGMENTATION ACCURACY FOR SCANS IN FIRST TERTILE OF LESION LOAD

Process	DC % (SD)	FOE % (SD)	FUE % (SD)	$\rho$
Parzen	45.77 (23.42)	566.66 (1221.51)	17.89 (16.09)	0.45
Position	73.34 (11.93)	59.31 (75.93)	14.22 (9.32)	0.85
Number	76.58 (8.61)	37.28 (45.61)	17.21 (10.61)	0.90
Shape	78.68 (5.95)	25.09 (17.60)	18.95 (10.88)	0.95

lesion load and lesion count (0.95). We expect our heuristics to have the greatest impact on the cases with low lesion loads. In particular, the positional and sample size adjustments are likely to affect mostly scans with small lesions and a low number of lesions, respectively.

Table II summarizes the validation results for the scans in the first tertile of lesion load. The positional adjustment dramatically improves both the DC (+27.57 pt.) and the correlation coefficient (+0.40). Without the proposed heuristics, the Parzen classifier generally overestimates the lesion areas by a large amount relative to the true lesion size. The positional adjustment greatly reduces overestimation (−507.35 pt.) and slightly reduces underestimation as well (−3.67 pt.). The sample size heuristic also improves the DC (+3.24 pt.) and correlation (+0.05), but to a lesser degree. Similarly, the shape refinement also increases the DC (+2.10 pt.) and correlation (+0.05). Both the sample size and shape refinement heuristics improve the balance between overestimation and underestimation by significantly decreasing the FOE, while only slightly increasing the FUE. Taken together, the three heuristics have a strong impact on the scans in this category of lesion load, improving the overlap similarity by +32.91 pt. to 78.68% and the correlation coefficient by +0.50 to 0.95. Qualitatively, Fig. 7 illustrates an example of how the false-positive regions are greatly reduced on scans with small and subtle lesions.

Table III summarizes the validation results for the scans in the second tertile of lesion load. It is notable that even without any of our heuristics, the Parzen window classifier yielded improved measures over those obtained with the first tertile scans using identical processing, which corroborates the finding from other researchers that small loads generally produce larger errors. Applying the positional adjustment improves both the DC (+12.64 pt.) and the correlation (+0.10) by reducing overestimation (−47.87 pt.). By comparison, the sample size adjust-

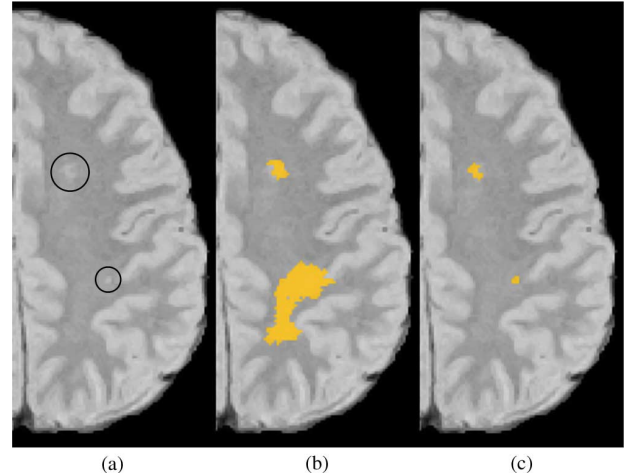


Fig. 7. Types of lesions for which the proposed methods have the greatest impact. (a) PDw scan with two small, subtle focal lesions (circled). (b) Results of Parzen windows without seed-processing optimizations. Both lesions are overestimated, with one having a very large false-positive region. (c) Results of Parzen windows with seed-processing optimizations. The false-positive regions are greatly reduced.

TABLE III  
SEGMENTATION ACCURACY FOR SCANS IN SECOND TERTILE  
OF LESION LOAD

Process	DC % (SD)	FOE % (SD)	FUE % (SD)	$\rho$
Parzen	66.30 (10.80)	74.30 (43.22)	16.52 (7.57)	0.78
Position	78.94 (6.53)	26.33 (19.27)	18.35 (6.62)	0.88
Number	79.45 (5.93)	22.33 (16.08)	19.77 (7.43)	0.89
Shape	80.85 (3.97)	14.03 (7.45)	22.47 (7.36)	0.93

TABLE IV  
SEGMENTATION ACCURACY FOR SCANS IN THIRD TERTILE  
OF LESION LOAD

Process	DC % (SD)	FOE % (SD)	FUE % (SD)	$\rho$
Parzen	81.34 (6.54)	23.56 (14.13)	15.63 (7.31)	0.95
Position	84.24 (4.83)	10.49 (7.70)	19.45 (7.63)	0.95
Number	84.18 (4.91)	10.47 (7.71)	19.57 (7.72)	0.95
Shape	83.11 (4.92)	7.58 (4.58)	23.21 (7.94)	0.95

ment only has a minimal effect, improving the DC by +0.51 pt. and correlation by +0.01. The shape refinement heuristic has a somewhat stronger effect than the sample size adjustment, increasing the DC by +1.40 pt. and correlation by +0.04. For this category of lesion load, the heuristics together account for a +14.55 pt. gain in the DC and +0.15 increase in the correlation coefficient, resulting in final values of 80.85% and 0.93, respectively.

Table IV summarizes the validation results for the scans in the third tertile, which has the highest lesion loads. Even without any of our new heuristics, the Parzen window classifier produces very strong DC (81.34) and correlation (0.95) values. The positional adjustment improves the DC further (+2.90 pt.), but does not change the correlation. This heuristic not only decreases the FOE (−13.07 pt.) but also increases the FUE, albeit by a smaller amount (+3.82 pt.). The sample size adjustment has practically no effect. The shape refinement step actually decreases the DC slightly (−1.07 pt.), even though the FOE continues to drop

TABLE V  
SEGMENTATION ACCURACY OVER ALL SCANS

Process	DC % (SD)	FOE % (SD)	FUE % (SD)	$\rho$
Parzen	64.47 (21.15)	221.51 (744.33)	16.68 (11.09)	0.91
Position	78.84 (9.42)	32.05 (49.63)	17.34 (8.22)	0.98
Number	80.07 (7.34)	23.36 (30.22)	18.85 (8.75)	0.99
Shape	80.88 (5.31)	15.57 (13.42)	21.54 (9.02)	0.99

TABLE VI  
PRE- AND POSTOPTIMIZATION RESULTS FOR EACH RADIOLOGIST WHO  
PERFORMED SEED POINT PLACEMENT

Rater	Scans	DC % (SD)	FOE % (SD)	FUE % (SD)	$\rho$
Without Optimized Seed Processing					
Rad 1	171	65.80(21.07)	203.16(694.58)	15.18(9.29)	0.92
Rad 2	63	60.88(21.12)	271.31(869.36)	20.75(14.24)	0.84
With Optimized Seed Processing					
Rad 1	171	80.45(5.70)	15.91(14.23)	22.00(9.46)	0.99
Rad 2	63	82.05(3.87)	14.63(11.00)	20.29(7.60)	0.99

(−2.89 pt.), likely because underestimation is increased by a greater amount (+3.64 pt.). This is the only case in our experiments in which the application of a heuristic results in a greater increase in underestimation than the decrease in overestimation. However, the correlation coefficient does not suffer as a result, but remains at 0.95.

Table V summarizes the validation results for all of the scans in our experiment. Without any of the heuristics, the correlation is very strong (0.91), but the DC is not (64.47%). This is explained by the large range of lesion loads in the dataset, which allows the correlation to be robust to changes in individual values, and highlights the need to perform a stratified analysis for an understanding of the true impact of the algorithmic components. Nonetheless, these results give some idea of the improvements that can be expected with similar datasets. The positional adjustment improves the DC by +14.37 pt. and the correlation by +0.07. The sample size adjustment has a small effect on the DC (+1.23 pt.) and the correlation (+0.01). The shape refinement process also has a small impact on the DC (+0.81 pt.), but no effect on the correlation that at 0.99 is already close to the theoretical maximum. Overall, the three heuristics are observed to reduce the overestimation significantly, while only increasing the underestimation slightly.

To assess whether the improvements in accuracy are sensitive to the particular radiologist placing the points, we computed the validation statistics for each radiologist individually with and without the optimized seed processing for the entire dataset, and present them in Table VI. Without the optimizations, the Parzen classifier seems to slightly favor one of the radiologists, yielding somewhat more accurate results (DC = 65.80 versus 60.88 and  $\rho = 0.92$  versus 0.84). However, with the proposed heuristics, the differences become very minor, and the accuracy values for each radiologist are comparable to the combined results (DC = 80+ and correlation = 0.99). While the differences between the radiologists are not very large even without the proposed heuristics, the added steps do seem to bring greater

TABLE VII  
SEGMENTATION ACCURACY COMPARED TO SCANS DONE BY EACH TECHNICIAN

Rater	Scans	DC % (SD)	FOE % (SD)	FUE % (SD)	$\rho$
Rater 1	15	78.82 (5.21)	9.06 (5.63)	28.61 (9.61)	0.98
Rater 2	72	82.78 (4.96)	14.71 (15.97)	21.48 (6.38)	0.99
Rater 3	36	79.27 (5.11)	11.65 (7.23)	26.33 (9.00)	0.99
Rater 4	16	80.71 (3.72)	11.01 (9.28)	25.02 (4.79)	0.95
Rater 5	95	80.41 (5.54)	14.37 (13.10)	23.20 (8.11)	0.99

inter-rater agreement in addition to improved accuracy for both radiologists.

Because five technicians were used to produce the validation data, examining how the enhanced classifier compares to each technician individually would help to reveal any problems in consistency across raters. Table VII shows the overall classification performance evaluated against the segmentations produced by each of the five raters. Although different scans were done by each technician, the results are still consistently strong across raters, with a mean DC of 80.40 (SD = 1.54) and mean correlation of 0.98 (SD = 0.02).

#### IV. DISCUSSION

The purpose of this study was to determine whether improvements in the processing of lesion seed points placed by radiologists without concern for the suitability of the points for subsequent automatic lesion segmentation would result in greater accuracy. Our experiments using a large MS clinical trial dataset and technician results as the reference standard, analyzed with measures of overlap similarity, overestimation, underestimation, and rank correlation, show that three relatively simple heuristics can greatly enhance the performance of segmentation methods, particularly those that use the samples to estimate the intensity distribution of the lesions. Two of these heuristics increase the accuracy of the estimated distributions by adjusting the position and number of the input points, and the third refines the shape of each lesion after intensity classification by applying an approximation to visual shape partitioning.

The heuristics work by reducing the overestimation that is prevalent when attempting to separate overlapping distributions using an intensity-based classifier, such as Parzen windows. In some cases, the decrease in false positives is partially offset by an increase in false negatives, but this is acceptable for two reasons. First, overall accuracy as measured by overlap similarity is improved in most cases, even if underestimation is increased. Second, certain guidelines for serial studies of MS call for a conservative approach to lesion volume measurement, and indicate a preference for underestimation rather than overestimation, if both cannot be minimized simultaneously [29].

The magnitude of the heuristics' impact is strongly related to lesion load, with the scans having the lowest lesion loads benefiting the most. It is a commonly accepted notion that reduced segmentation accuracy, as evaluated by relative measures, can be expected from most methods when applied to low lesion loads. It may be tempting to circumvent the problem by only working with scans with relatively large lesion volumes [8], or assuming that small lesions are clinically insignificant and can

be routinely excluded [6], but there is evidence that small lesions can have particular relevance for some clinical considerations of MS [30]. Our optimizations are able to improve the performance of the classifier on the lower two tertiles without negatively affecting the highest tertile. In fact, even though the results for the third tertile yielded by the unmodified Parzen method are already very strong, the positional adjustment is able to produce a slight improvement. However, it should be noted that, unlike the lower two tertiles, the FUE becomes greater than the FOE after the positional adjustment, which suggests that the lesion PDF may be becoming too narrow for the high-load scans, likely with the partial volume areas being the most affected. In spite of this, the large decrease in the FOE and the improvement in the DC make the increase in FUE a worthwhile tradeoff. While the shape refinement process seems to result in increased underestimation that may slightly outweigh the benefit of reduced overestimation for scans with high lesion load, the impact on the overlap similarity and correlation is minimal. For the dataset overall, the improved classifier agrees strongly to the interactive segmentation by trained technicians. However, one of the limitations of the current study is that the validation segmentations were produced with the same seed points as the automatic classifier. While we see the main value of the optimization heuristics as being able to potentially replace the labor-intensive region-growing process, and as a drop-in replacement the heuristics seem to work very well, stronger validation can be obtained with an independent set of seed points, preferably placed by another group of radiologists. Therefore, we see the *improvements* resulting from the heuristics, rather than the absolute agreement between the automatic and interactive methods, as being the most noteworthy aspect of this experiment for many researchers who have existing segmentation pipelines and are looking for relatively simple ways to increase accuracy in scans with low-to-medium lesion loads.

After trying a range of parameter values, the results indicate that even small adjustments in the position (at most moving two pixels) and number of lesion points (requiring only five as a minimum) can have a dramatic effect, particularly for scans with small lesion loads. In general, it is impractical to expect a radiologist to be so mindful of the placement and number of points as to obviate these adjustments. It should be emphasized that while our parameters values are reported for completeness, this study is not intended to propose optimal parameters, but rather to present a strategy by which the use of radiologist-placed points can be improved. In the interest of focus, we have not presented the intermediate results of our parameter optimization procedure, and for researchers aiming to replicate this experiment, our recommendation is to explore a range of parameter values with their own data so that the heuristics can be tuned to the tendencies of their radiologists. We believe that having more radiologists, particularly those external to our group, perform similar experiments is an important step in confirming that the methods are widely applicable.

Relative to the interactive region-growing process, the new pipeline offers much improved efficiency. The average time taken by the technicians to perform the semiautomatic segmentation is 1813.2 s (SD = 1465.3 s), or about 30 min per scan. In

contrast, running the entire automatic pipeline takes an average of 125.0 s (SD = 11.7 s), or about 3 min, on an Intel Core 2 CPU (2.40-GHz, 4096-kB L2 cache). A breakdown of the percentage of time taken by each major component is as follows: multiscale N3 (21.4%), SUSAN (1.0%), BET (1.5%), automatic sampling of healthy tissues (4.1%), and Parzen window classification with the proposed heuristics (72.0%). The publicly available software components (N3, SUSAN, and BET) were used with little modification from the downloaded versions, and all were implemented in C++ by their respective authors. The sampling of healthy tissues was implemented in MATLAB. The Parzen window classifier, including the seed-processing optimizations, was implemented in C++. We expect that greater integration of the above components would further increase the efficiency of the automatic pipeline.

Despite the large gains in performance achieved by the heuristics, the methods presented are only a first step in the investigation of how the use of radiologist sample points can be optimized. For example, while the sample size adjustment procedure seems to improve the stability of PDF estimation with respect to the number of lesions, the method does not account for normal variations in intensity, such as that between the cerebrum and cerebellum, which may widen the estimated distributions. In addition, the shape refinement technique is only a rough approximation to visual shape partitioning. As implemented, the straightline visibility test may be biased against thin curved lesions. In addition, the lack of a boundary model makes defining the precise conditions, for which the heuristic would be beneficial, difficult. Greater discrimination between true and false positives and improved reproducibility may be possible by explicitly using a perceptually relevant shape model that can directly represent the boundary curvature properties. Some direct discussion with the radiologists may also reveal further insight into their visual processing.

In this study, to develop the proposed heuristics, we chose to use Parzen windows with connected component analysis as the classification method. Alternatively, we could have used other seed-based segmentation methods, especially those that inherently incorporate spatial coherence, such as fuzzy connectivity [9] or graph-based methods (e.g., [31] and [32]). The main reason we favor Parzen windows is that this and other non-parametric PDF estimators are much more popular in existing lesion segmentation pipelines, and we wanted to develop heuristics that are immediately and widely applicable. Nonetheless, it would likely be worthwhile to explore the impact of similar seed-processing methods on the more advanced segmentation algorithms.

Another direction for future research is to determine the effect of similar optimizations on segmentation methods that only use expert input on a training set of images to condition the classifier for application to a larger set of completely unlabeled scans. The heuristics would be applied to the training data, which may result in more accurate algorithmic parameters for subsequent unsupervised use. In addition, similar to the use of probability distributions estimated from the training images to perform an intensity classification of the unlabeled scans, the shape refinements performed on the training data can be

analyzed to potentially reveal tendencies that may be applicable to the new scans. In some approaches, the training input is in the form of interactively produced regions instead of points (e.g., [33]). For such cases, the experience of the current experiment can be potentially used to develop an intelligent way of sampling the drawn regions, which may be advantageous to using the regions directly due to possible errors in manual segmentation.

#### ACKNOWLEDGMENT

The authors would like to thank Y. Cheng and G. Zhao for performing the placement of lesion seed points.

#### REFERENCES

- [1] M. Yoshita, E. Fletcher, and C. DeCarli, "Current concepts of analysis of cerebral white matter hyperintensities on magnetic resonance imaging," *Topics Magn. Res. Imag.*, vol. 16, no. 6, pp. 399–407, 2005.
- [2] F. B. Mohamed, S. Vinitzki, S. H. Faro, C. F. Gonzalez, J. Mack, and T. Iwanaga, "Optimization of tissue segmentation of brain MR images based on multispectral 3D feature maps," *Magn. Res. Imag.*, vol. 17, no. 3, pp. 403–409, 1999.
- [3] R. C. Parodi, F. Sardanelli, P. Renzetti, E. Rosso, C. Losacco, A. Ferrari, F. Levvero, A. Pilot, M. Inglese, and G. L. Mancardi, "Growing region segmentation software (GRES) for quantitative magnetic resonance imaging of multiple sclerosis: Intra- and inter-observer agreement variability: A comparison with manual contouring method," *Eur. Radiol.*, vol. 12, no. 4, pp. 866–871, 2002.
- [4] M. E. Payne, D. L. Fetzer, J. R. MacFall, J. M. Provenzale, C. E. Byrum, and K. R. R. Krishnan, "Development of a semi-automated method for quantification of MRI gray and white matter lesions in geriatric subjects," *Psychiatry Res.*, vol. 115, no. 1–2, pp. 63–77, 2002.
- [5] P. Anbeek, K. L. Vincken, M. J. P. van Osch, R. H. C. Bisschops, and J. van der Grond, "Automatic segmentation of different-sized white matter lesions by voxel probability estimation," *Med. Image Anal.*, vol. 8, no. 3, pp. 205–215, 2004.
- [6] F. Admiraal-Behloul, D. M. J. van den Heuvel, H. Olofsen, M. J. P. van Osch, J. van der Grond, M. A. van Buchem, and J. H. C. Reiber, "Fully automatic segmentation of white matter hyperintensities in MR images of the elderly," *NeuroImage*, vol. 28, no. 3, pp. 607–617, 2005.
- [7] B. R. Sajja, S. Datta, R. He, M. Mehta, R. K. Gupta, J. S. Wolinsky, and P. A. Narayana, "Unified approach for multiple sclerosis lesion segmentation on brain MRI," *Ann. Biomed. Eng.*, vol. 34, no. 1, pp. 142–151, 2006.
- [8] Y. Wu, S. Warfield, I. Tan, W. Wells, D. Meier, R. van Schijndel, F. Barkhof, and C. Guttman, "Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI," *NeuroImage*, vol. 32, pp. 1205–1215, 2006.
- [9] M. A. Horsfield, R. Bakshi, M. Rovaris, M. A. Rocca, V. S. R. Dandamudi, P. Valsasina, E. Judica, F. Lucchini, C. R. G. Guttman, M. P. Sormani, and M. Filippi, "Incorporating domain knowledge into the fuzzy connectedness framework: Application to brain lesion volume estimation in multiple sclerosis," *IEEE Trans. Med. Imag.*, vol. 26, no. 12, pp. 1670–1680, Dec. 2007.
- [10] R. Khayati, M. Vafadust, F. Towhidkha, and M. Nabavi, "Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model," *Comput. Biol. Med.*, vol. 38, no. 3, pp. 379–390, 2008.
- [11] R. de Boer, H. A. Vrooman, F. van der Lijn, M. W. Vernooij, M. A. Ikram, A. van der Lugt, M. M. B. Breteler, and W. J. Niessen, "White matter lesion extension to automatic brain tissue segmentation on MRI," *NeuroImage*, vol. 45, no. 4, pp. 1151–1161, 2009.
- [12] A. Traboulsee, G. Zhao, and D. K. B. Li, "Neuroimaging in multiple sclerosis," *Neurologic Clin.*, vol. 23, no. 1, pp. 131–148, 2005.
- [13] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.
- [14] S. Warfield, "Fast k-NN classification for multichannel image data," *Pattern Recognit. Lett.*, vol. 17, no. 7, pp. 713–721, 1996.
- [15] B. T. T. Yeo, M. R. Sabuncu, R. Desikan, B. Fischl, and P. Golland, "Effects of registration regularization and atlas sharpness on segmentation accuracy," *Med. Image Anal.*, vol. 12, no. 5, pp. 603–615, 2008.
- [16] M. Filippi, A. Falini, D. L. Arnold, F. Fazekas, O. Gonen, J. H. Simon, V. Douset, M. Savoiardo, and J. S. Wolinsky, "Magnetic resonance techniques for the in vivo assessment of multiple sclerosis pathology: Consensus report of the white matter study group," *J. Magn. Res. Imag.*, vol. 21, no. 6, pp. 669–675, 2005.
- [17] C. Jones and E. Wong, "Multi-scale application of the N3 method for intensity correction of MR images," in *Proc. SPIE (Med. Imag. 2002: Image Process.)*, pp. 1123–1129.
- [18] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998.
- [19] S. M. Smith and J. M. Brady, "SUSAN—A new approach to low level image processing," *Int. J. Comput. Vis.*, vol. 23, no. 1, pp. 45–78, 1997.
- [20] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapp.*, vol. 17, no. 3, pp. 143–155, 2002.
- [21] C. Jones, D. Li, G. Zhao, D. Paty, and PRISMS Study Group, "Atrophy measurements in multiple sclerosis," in *Proc. Int. Soc. Magn. Reson. Med. Sci. Meeting*, Glasgow, Scotland, 2001, p. 1414.
- [22] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
- [23] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, Jul. 1990.
- [24] M. A. Sahraian and E.-W. Radue, "MS Lesions in T2-Weighted Images," in *MRI Atlas of MS Lesions*. New York: Springer-Verlag, 2008, p. 7.
- [25] K. Siddiqi, B. Kimia, A. Tannenbaum, and S. Zucker, "On the psychophysics of the shape triangle," *Vis. Res.*, vol. 41, no. 9, pp. 1153–1178, 2001.
- [26] L. Kappos, A. Traboulsee, C. Constantinescu, J.-P. Eralinna, F. Forrester, P. Jongen, J. Pollard, M. Sandberg-Wollheim, C. Sindic, B. Stubinski, B. Uitdehaag, and D. Li, "Long-term subcutaneous interferon beta-1a therapy in patients with relapsing-remitting MS," *Neurology*, vol. 67, no. 6, pp. 944–953, 2006.
- [27] A. Traboulsee, A. AL-Sabbagh, R. Bennett, P. Chang, D. Li, the EVIDENCE Study Group, and UBC MS/MRI Research Group, "Reduction in magnetic resonance imaging T2 burden of disease in patients with relapsing-remitting multiple sclerosis: Analysis of 48-week data from the EVIDENCE study," *BMC Neurol.*, vol. 8, no. 1, pp. 11–17, 2008.
- [28] L. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.
- [29] M. Filippi, M. L. Gawne-Cain, C. Gasperini, J. H. van Waesberghe, J. Grimaud, F. Barkhof, M. P. Sormani, and D. H. Miller, "Effect of training and different measurement strategies on the reproducibility of brain MRI lesion load measurements in multiple sclerosis," *Neurology*, vol. 50, no. 1, pp. 238–244, 1998.
- [30] R. Zivadinov, M. Zorzon, R. De Masi, D. Nasuelli, and G. Cazzato, "Effect of intravenous methylprednisolone on the number, size and confluence of plaques in relapsing-remitting multiple sclerosis," *J. Neurol. Sci.*, vol. 267, no. 1–2, pp. 28–35, 2008.
- [31] D. Garcia-Lorenzo, J. Lecoeur, D. L. Arnold, D. L. Collins, and C. Barillot, "Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 2009, pp. 584–591.
- [32] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.
- [33] A. Akselrod-Ballin, M. Galun, J. M. Gomeri, M. Filippi, P. Valsasina, R. Basri, and A. Brandt, "Automatic segmentation and classification of multiple sclerosis in multichannel MRI," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 10, pp. 2461–2469, Oct. 2009.

Authors' photographs and biographies not available at the time of publication.