Automatic Segmentation and Classification of Multiple Sclerosis in Multichannel MRI

Ayelet Akselrod-Ballin*, *Member, IEEE*, Meirav Galun, *Member, IEEE*, John Moshe Gomori, Massimo Filippi, Paola Valsasina, Ronen Basri, *Member, IEEE*, and Achi Brandt

***Abstract*—We introduce a multiscale approach that combines segmentation with classification to detect abnormal brain structures in medical imagery, and demonstrate its utility in automatically detecting multiple sclerosis (MS) lesions in 3-D multichannel magnetic resonance (MR) images. Our method uses segmentation to obtain a hierarchical decomposition of a multichannel, anisotropic MR scans. It then produces a rich set of features describing the segments in terms of intensity, shape, location, neighborhood relations, and anatomical context. These features are then fed into a decision forest classifier, trained with data labeled by experts, enabling the detection of lesions at all scales. Unlike common approaches that use voxel-by-voxel analysis, our system can utilize regional properties that are often important for characterizing abnormal brain structures. We provide experiments on two types of real MR images: a multichannel proton-density-, T2-, and T1-weighted dataset of 25 MS patients and a single-channel fluid attenuated inversion recovery (FLAIR) dataset of 16 MS patients. Comparing our results with lesion delineation by a human expert and with previously extensively validated results shows the promise of the approach.**

***Index Terms*—Brain imaging, MRI, multiple sclerosis, segmentation.**

I. Introduction

IDENTIFYING 3-D brain structures in medical imagery, particularly in MRI scans, is important for early detection of tumors, lesions, and abnormalities, with applications in diagnosis, follow-up, and image-guided surgery [1]–[3]. Computer-aided analysis can assist in identifying brain structures, extract quantitative and qualitative properties of these structures, and evaluate their progress over time. Manual or interactive segmentation by human experts is time-consuming, expensive, and suffers from considerable inter- and intrarater variability. In addition, it is difficult for a human expert to combine information from several slices and multiple channels when multispectral MRI data are examined. While semiautomatic methods [4]–[6] significantly improve the inter- and intrarater variability, they still depend on varying degrees of human intervention, which often are not as robust, reproducible, and reliable as the analysis that would be made by top expert radiologists.

Multiple sclerosis (MS) is one of the most common diseases of the central nervous system (CNS) in young adults, affecting over 2 500 000 patients worldwide. MS is characterized by the destruction of proteins in the myelin surrounding nerve fibers. As a result, multiple areas of scar tissue called sclerosis (also lesions, or plaques) may appear, leading to a progressive decline of motor, vision, sensory, and cognitive function. MRI is a powerful tool for diagnosis of MS and monitoring the disease activity and progression [7]. Consequently, automatic quantitative analysis of MS in MRI has become increasingly important. In this paper, we present a novel method for detecting abnormal brain structures, focusing on 3-D MRI brain data containing scans of MS patients.

A. Background

Automatic segmentation of abnormal brain structures, and particularly MS lesions, is difficult. Abnormal structures exhibit extreme variability. Their shapes are deformable, their location across patients may differ significantly, and their intensity and texture characteristics may vary. Existing systems commonly approach this problem by applying classification algorithms that rely on a voxel-by-voxel analysis, utilizing primarily image intensities and atlas probability values [2], [8]–[10]. Neighborhood relations may be encoded through a Markov random field (MRF) model or other neighborhood statistics [11], [12]. However, we are unaware of an approach that utilizes regional statistical properties at different scales, particularly properties related to the shape, boundaries, and texture statistics.

A number of automatic algorithms have been designed specifically for MS segmentation. Zijdenbos *et al.* [2] developed an automatic pipeline for T1-, T2-, and proton density (PD)-weighted images based on a supervised artificial neural network (ANN) classifier and validated it extensively on multicenter clinical trial. Wells *et al.* [10] use a Gaussian mixture distribution and bias field correction to identify major brain tissues and separate them from the lesions. Van Leemput *et al.* [11] extended this framework by incorporating a probabilistic brain atlas along with neighborhood constraints. Wei *et al.* [9] tested three pipelines, with a pipeline combining template-driven segmentation (TDS), deformable anatomical atlas, and a heuristic connectivity-based partial volume effect (PVE) correction component, demonstrating the highest accuracy. Wu *et al.* [13] expanded this system to a three-channel MRI pipeline for detection

Manuscript received October 1, 2007; revised February 22, 2008. Current version published September 16, 2009. This work was supported in part by the Binational Science foundation under Grant 2002/254, in part by the Israel Institute of Technology, and in part by the European Commission Project IST-2002-506766 Aim Shape. *Asterisk indicates the corresponding author.*

*A. Akselrod-Ballin is with the Computational Radiology Laboratory, Children's Hospital, Harvard Medical School, Boston, MA 02115 USA (e-mail: Ayelet.Akselrod-Ballin@childrens.harvard.edu).

M. Galun, R. Basri, and A. Brandt are with the Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel.

J. M. Gomori is with the Department of Radiology, Hadassah University Hospital, Jerusalem 91120, Israel.

M. Filippi and P. Valsasina are with the Neuroimaging Research Unit, Hospital San Raffaele, Milan 20132, Italy.

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.2008.926671

of subtypes of MS lesions, improving sensitivity, specificity, and accuracy.

B. Our Contribution

This paper introduces a novel approach based on a combination of a powerful multiscale segmentation algorithm [14], [15], a rich feature vocabulary describing the segments, and decision forest classification of the segments. By combining segmentation and classification, we are able to utilize integrative, regional properties that provide regional statistics of segments, characterize their overall shapes, and localize their boundaries.

Our method offers the following advantages. First, it relies on an algebraic multigrid (AMG), *multiscale* graph partitioning approach to provide a hierarchical decomposition of a magnetic resonance (MR) scan in only linear time complexity. Second, we incorporate an novel rich set of *multiscale features* to guide the pyramid construction and to characterize MS lesions. We further use a decision forest along with Fisher linear discriminant (FLD) to utilize this richer set of features. Third, our method is general and flexible, and can be adapted to handle other, similar medical problems. Fourth, similar to other approaches, the method is fully automatic due to the use of a probabilistic brain atlas. We further use atlas data to identify the *cerebellum* due to the difficulty in detection of MS in this area [16]. Finally, our algorithm provides a *soft classification* result with different levels of MS disease probability rather than just a binary result. As the anticipated extent of the lesions may vary significantly between experts [2], [9], [11], this property can be valuable for clinical analysis.

The paper is organized as follows. Section II introduces the segmentation procedure, Section III presents the feature extraction method, and Section IV describes the classification model in our system. In Section V, experiments on two types of brain MRI data are presented. Section VI follows with a discussion and conclusions. An earlier version of this work had appeared in an IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

II. Segmentation Framework

We extended the segmentation by weighted aggregation (SWA) algorithm [14], [15] to handle 3-D multichannel and anisotropic data. In this section, we review the SWA algorithm along with our extensions.

Given a 3-D MRI scan, a six-connected graph $G = (V, W)$ is constructed as follows. Each voxel i is represented by a graph node i ($1 \leq i \leq N$). A weighted edge w_{ij} is associated with each pair of neighboring voxels i and j, reflecting the contrast between them as follows:

$$\omega_{ij} = e^{-\alpha |I_i - I_j|} \tag{1}$$

where I_i and I_j denote the intensities of the two neighboring voxels and α is a positive constant ($\alpha = 15$ in our experiments). We define the saliency of a segment by applying a normalized-cut-like measure as follows. Every segment $S \subseteq V$ is associated with a state vector $u = (u_1, u_2, \ldots, u_N)$ representing the assignments of voxels to a segment S: $u_i = 1$ if $i \in S$ and $u_i = 0$ otherwise.

The *saliency* Γ associated with S is defined by

$$\Gamma(S) \stackrel{def}{=} \frac{u^T L u}{\frac{1}{2} u^T W u} \tag{2}$$

where the similarity matrix W includes the weights w_{ij} and L is the Laplacian matrix of G whose elements are

$$l_{ij} = \begin{cases} \sum_{k\,(k \neq i)} w_{ik}, & i = j \\ -w_{ij}, & i \neq j. \end{cases} \tag{3}$$

Intuitively, the saliency sums the weights along the boundaries of S normalized by its internal weights. Segments that yield small values of $\Gamma(S)$ are considered salient.

If we allow arbitrary real assignments to u, the minimum for Γ is obtained by the minimal generalized eigenvector u of $Lu = \lambda W u$, with the condition that $\lambda > 0$. This equation is in similar to the normalized cuts solution [17].

Our objective is to efficiently find partitions characterized by small values of Γ at all scales. Starting from the initial graph $G^{[0]} \stackrel{def}{=} G$, we create a sequence of graphs $G^{[1]}, \ldots, G^{[k]}$ of decreasing size. This construction is divided into three stages: first, a subset of the fine nodes is chosen to serve as the *seeds*; these will be the nodes of the coarse graph. Then, an *interpolation matrix* is determined, establishing the fraction of each nonseed node to belong to each seed. Finally, the *coupling weights* of the edges between the coarse nodes are calculated.

The construction of the set of *seeds* C, and its complement denoted by F, is guided by the principle that each F-node should be "strongly coupled" to C. To achieve this objective, we start with an empty set C; hence, $F = V$, and sequentially (according to decreasing aggregate size defined in Section III) transfer nodes from F to C until all the remaining $i \in F$ satisfy $\sum_{j \in C} w_{ij} \geq \eta \sum_{j \in V} w_{ij}$, where η is a parameter (we used $\eta = 0.2$).

We define a *sparse interpolation matrix* P (of size $N \times n$, where $n = |C|$) as follows:

$$P_{ij} = \begin{cases} \frac{w_{ij}}{\sum_{k \in C} w_{ik}}, & \text{for} \quad i \in F, j \in C \\ 1, & \text{for} \quad i \in C, j = i \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This matrix satisfies $u \approx PU$, where $U = (U_1, U_2, \ldots, U_n)$ is the coarse level state vector. P_{ij} represents the likelihood of an aggregate i at a fine level to belong to an aggregate j at a coarser level. Finally, an edge connecting two coarse nodes k and l in the coarse graph is assigned with the weight

$$w_{kl}^{\text{coarse}} = \sum_{p \neq q} P_{pk} w_{pq} P_{ql}. \tag{5}$$

w_{kl}^{coarse} is also called the *coupling weight* between aggregates k and l obtained by weighted aggregation. Intuitively, the coupling weight between a pair of coarse aggregates is the weighted sum of the coupling weights between their subaggregates. Using the interpolation matrix P, the saliency measure (2) can be written as

$$\Gamma = \frac{u^T L u}{\frac{1}{2} u^T W u} \approx \frac{U^T P^T L P U}{(1/2) U^T P^T W P U}. \tag{6}$$

TABLE I
OUTLINE OF THE 3-D SEGMENTATION ALGORITHM

- Given a 3D MRI initialize a 6-connected graph $G^{[0]} = (V^{[0]}, W^{[0]})$. $V^{[0]}$ is the set of image voxels, and $W^{[0]}$ is defined according to (5).
- Repeat: for $s = 1, 2, \ldots$ construct $G^{[s]}$ from $G^{[s-1]}$, as follows:
  1) Seed selection: select a representative set of nodes $V^{[s]}$, such that $V^{[s-1]} \setminus V^{[s]}$ is strongly connected to $V^{[s]}$.
  2) Define $P = P^{[s-1]}$ the interscale interpolation matrix (4).
  3) Calculate $W^{[s]} \approx P^T W^{[s-1]} P$ by weighted aggregation (5).
  4) For each node $v \in V^{[s]}$ calculate its aggregative features.
  5) Modify $W^{[s]}$ according to the aggregative features (8).

The right-hand side of (6) determines a coarser graph with n nodes whose weight matrix $W^{\text{coarse}} = P^T W P$ includes the coarse edge weights in the off-diagonal elements. $L^{\text{coarse}} = P^T L P$ is approximated by a relation to W as in (3) [15]. At each level, W^{coarse} is further modified to account for aggregative properties that cannot be expressed at the finest level (Section III).

This coarsening procedure is performed recursively. We denote a coarse scale by s and its predecessor finer scale by $(s-1)$. The scale index is attached to the graph notation, i.e., a graph at scale s is denoted by $G^{[s]} = (V^{[s]}, W^{[s]})$, the appropriate interpolation matrix between scale s and $(s-1)$ is denoted by $P^{[s-1][s]}$ or $P^{[s-1]}$, and $|V^{[s]}|$ is denoted by $N^{[s]}$. Table I summarizes the segmentation algorithm.

A. Handling Anisotropic Data

In many cases, the MRI data are anisotropic (i.e., distances in interslice directions are usually larger than intraslice ones). However, the SWA algorithm assumes that the voxels in the fine level are equally spaced, since the initial graph does not take into account the distances between neighbors [see (1)]. Ignoring this effect may lead to distorted segmentations. To solve this problem, we modify the algorithms as follows. During the first few coarsening steps, we consider each 2-D slice separately while performing seed selection and interscale interpolation (steps 1–2 in Table I), allowing nonzero interpolation weights only between nodes of the same slice. The rest of the steps (steps 3–5 in Table I) are performed on the full 3-D graph. This procedure is repeated until the innerslice and interslice distances are approximately equal. Then, subsequent coarsening steps repeat all steps (steps 1–5 in Table I) recursively considering the full 3-D graph.

B. Multichannel Segmentation

A major aspect of MR imaging is the large variety of pulse sequences that can be applied. The multichannel data are incorporated as follows. First, the different sequences provided for each subject are aligned. In this paper, the alignment between T1-weighted and dual-echo images (used to obtain both T2- and PD-weighted images) was achieved in the acquisition phase, by acquiring the T1-weighted immediately after the dual echo and using the same positioning parameters. Consequently, given the multichannel aligned scans, each voxel now includes a vector of intensities. Therefore, (1) is modified to determine fusion weights exploiting intensity information from all m channels as follows:

$$w_{ij} = \exp\left[-\left(\sum_{c=1}^{m} (\alpha_c)^2 (I_i^c - I_j^c)^2\right)^{1/2}\right] \tag{7}$$

where α_c are predetermined constants and I_i^c is the intensity of voxel i in channel c. Our choice of constants ($\alpha_{T2} = 15, \alpha_{PD} = \alpha_{T1} = 10$) puts more emphasis on T2 intensity contrast effects in the segmentation process.

Following the fusion weight initialization, we maintain different sets of aggregative features for every channel (see Section III) and use these properties to modify the edge weights at coarser levels. Let m and p denote the total number of channels and scales, respectively, then the influence of the multiscale statistics of average intensity and average of variances on the coupling between two aggregates k and l is considered by multiplying the coupling with the following term:

$$\exp\left[-\left(\sum_{c=1}^{m} (\gamma_c)^2 (\bar{I}_k^c - \bar{I}_l^c)^2\right)^{1/2}\right]$$
$$\exp\left[-\left(\sum_{c=1}^{m} (\beta_c)^2 \Delta\nu_{kl}^c\right)^{1/2}\right] \tag{8}$$

where β_c and γ_c are coefficient parameters that control the weight of the different measures in the different c channels ($\beta_c = 0.5$, $\gamma_{\text{FLAIR}} = \gamma_{T2} = 10$, $\gamma_{PD} = \gamma_{T1} = 6$), and $\Delta\nu_{kl}^c$ reflects the dissimilarity of the average of variances for each channel c (see $\nu_k^{[r]}$ in Table II later), which is defined as

$$\Delta\nu_{kl}^c = \frac{1}{p}\sum_{r=1}^{p}\left(\frac{2(\bar{\nu}_k^{c[r]} - \bar{\nu}_l^{c[r]})}{(\bar{\nu}_k^{c[r]} + \bar{\nu}_l^{c[r]})}\right)^2. \tag{9}$$

III. FEATURE EXTRACTION

Lesions can often be characterized by properties of aggregates that emerge at intermediate scales and are difficult to extract by any uniscale procedure. Such properties may include, for instance, intensity homogeneity, principal direction of the lesion, and intensity contrast with respect to neighboring tissues. Voxel-by-voxel analysis is often limited in its ability to utilize such scale-dependent properties.

We refer to such scale-dependent properties as *aggregative features* since the weighted aggregation scheme provides a recursive mechanism for calculating such properties as part of the segmentation process while maintaining the overall linear complexity of the segmentation process. A high-dimensional feature vector containing these aggregative features is constructed for every aggregate in the pyramid. The list of features relevant to the problem domain was selected following interaction with expert radiologists. We use these properties for two purposes. First, we use these aggregative properties to affect the construction of the segmentation pyramid [see (8)]. Second, these properties are available for the classification procedure later (Section IV). However, the actual effect of each of these features is determined

TABLE II
AGGREGATIVE FEATURES IN VECTOR f OF AGGREGATE k

Graph measures:

- **Saliency:** Γ (Eq. 2)

Intensity statistics:

- **Average intensity** of voxels in aggregate k, denoted $\bar{I}_k^{[0]}$. Normalized by the average intensity of the intracranial cavity (IC) (see details in Sec. V-B2).
- **Maximum/Minimum intensity:** denoted $\mu_k^{[2]}$, maximal/minimal average intensity of the sub-aggregates at scale 2. Normalized by $\bar{I}_k^{[0]}$.
- **Variance of average intensities of scale r:** $Var^{[r]} = \bar{I}_k^{2[r]} - (\bar{I}_k^{[0][r]})^2$, where $\bar{I}_k^{2[r]}$ denotes the average of $(\bar{I}_l^{[0][r]})^2$ for all sub-aggregates l of k at scale r. Normalized by $\bar{I}_{k_l}^{2[0]}$.
- **Average of variances:** denoted $\bar{\nu}_k^{[r]}$ where $\nu_k^{[r][r]} = Var^{[0][r]}$.
- **Average intensity Proportions:** Proportion between Average intensity of the different channels.

Shape:

- **Size:** $M^{[0]}$ is the aggregate volume in voxel units. Normalized by the IC volume.
- **Shape moments:** The length, width, depth ($L^{[0]}$, $W^{[0]}$,$D^{[0]}$ respectively), and Orientation are specified by applying principal component analysis to the covariance matrix of the aggregate. Normalized by the corresponding values of shape moments measured for the entire IC.
- **Intensity moments:** averages of products of the intensity and the coordinates of voxels in aggregate k, denoted $\overline{Ix}^{[0]}, \overline{Iy}^{[0]}, \overline{Iz}^{[0]}$. The normalization is performed by the following expression:

$$\frac{\overline{Ix}^{[0]} - \bar{I}^{[0]}\bar{x}^{[0]}}{(Var^{[r]})^{\frac{1}{2}}(\bar{x_k}^{2[0]} - (\bar{x}_k^{[0]})^2)^{\frac{1}{2}}}. \tag{10}$$

Neighborhood statistics:

- **Boundary surface area:** is denoted $B_{kl}^{[0]}$ and refers to the surface area of the common border of aggregates k and l. It is accumulated during the same segmentation process by weighted aggregation of weights that on the finest graph are set to 1. The boundary surface area was normalized by $(M^{[0]})^{\frac{2}{3}}$ where $M^{[0]}$ is the size of aggregate k in voxel units.
- **Neighborhood average intensity, denoted** $\overline{NI}_k^{[0]}$, formulated as:

$$\overline{NI}_k^{[0]} = \frac{\sum_l B_{kl}\bar{I}_l^{[0]}}{\sum_l B_{kl}} \tag{11}$$

- **Selected Neighborhood Statistics:** for the boundary surface (Bs) and neighborhood average intensity (NIs) denoted by Bs_{kl} and $\overline{NIs}_k^{[0]}$ respectively. Computed on a partial neighborhood of the aggregate, by selecting only neighboring aggregates which have an average intensity that is higher than the IC average intensity.

Atlas statistics:

- **Location:** $\bar{x}^{[0]}$, $\bar{y}^{[0]}$, $\bar{z}^{[0]}$ relative to a common coordinate system of a brain atlas.
- **Anatomical probabilities:** the average likelihood of finding gray matter (GM), white matter (WM), cerebro-spinal fluid (CSF) or the cerebellum (CE), in an aggregate k, denoted by $\overline{P}_{WM}^{[0]}$, $\overline{P}_{GM}^{[0]}$, $\overline{P}_{CSF}^{[0]}$, $\overline{P}_{CE}^{[0]}$.

in training by an automatic learning process. Table II provides the list of aggregative features extracted by our system.

In this framework, for each aggregate k emerging at a certain scale s, we calculate a set of aggregative properties. An aggregative property can be expressed as a weighted average over the aggregate k of a property that has first appeared at a scale r ($r \leq s$). The scale s is termed the *aggregate scale* and the scale r is called the *property scale*. At each scale s, the similarity matrix $W^{[s]}$, inherited from finer aggregate scales (5), is modified by the similarities arising from the set of aggregative properties obtained from multiple property scales. For example, the average intensity of aggregate k is an aggregative property, since it is the weighted average over all intensities measured for the voxels (nodes of scale $r = 0$) that belong to k. More complex aggregative properties can be constructed by combining several properties (e.g., variance of average intensities later, which combines the average intensity and average squares of intensities of k) or by taking averages over aggregative properties of finer scales (e.g., average of variances later). A certain property Q, emerging at scale r, of an aggregate k at scale s is denoted by $Q_k^{[r][s]}$. In addition to these properties, we can define binary aggregative properties, reflecting r-scale relations between two aggregates k and l at scale s. Such properties, denoted by $Q_{kl}^{[r][s]}$, are useful for describing boundary relations between neighboring tissues, e.g., surface area of boundary between k and l or the contrast between the average intensity of an aggregate k and the average intensity of its neighbors. The anatomical probabilities features exploit the Statistical Parametric Mapping (SPM) software package [18], to align the subject's data and International Consortium for Brain Mapping (ICBM) atlas probability maps [19], which represent the probability of finding an anatomy type at a specified position. Table II provides the list of features computed with our method.

IV. CLASSIFICATION

Once an MRI scan is segmented, we obtain a full hierarchy of aggregates. Our aim in the classification stage is to identify the aggregates corresponding to lesions. In a training phase, a decision forest classifier [20], [21] is trained based on the aggregative features using data labeled by MS experts. Then, once the system is trained, unlabeled test scans are provided as input, and the classifier is used to discriminate between aggregates corresponding to lesions and nonlesions in these scans. Our classification approach differs from [5], who also used decision trees for MS segmentation, by utilizing a decision forest along with the FLD analysis to deal with multiple features. Next, we describe the training and testing phase and how we use the classification results of segments to determine the classification of individual voxels.

A. Classification With Decision Forest

To construct the decision forest classifier, a training process is applied using MRI scans with MS lesions delineated by experts. The process obtains two kinds of data: 1) a collection of M feature vectors, $C\text{and} = \{f_1, \ldots, f_M\}$, describing M candidate segments (with each feature normalized to have zero mean and unit variance); and 2) a mask indicating the voxels marked as lesions by an expert. We label as a lesion and denote by class c_1, a segment in which $\geq$70% of its voxels were marked by an expert as a lesion. Since the candidate segments may contain a mixed collection of lesion and nonlesion voxels, we selected the 70% threshold in order to include in c_1 only segments that are clearly characterized by lesion properties. We further mark as nonlesions only those segments that do not contain lesion voxels

at all and denote this class by c_2 (other segments are ignored during training).

Given the training set, a subset of the candidate segments are randomly selected and used to construct a tree recursively from the root downwards. To determine a split at each tree node, an FLD [22] is applied to the feature vectors automatically determining the optimal separation direction that achieves maximal impurity decrease.

Throughout the training procedure, multiple decision trees are constructed resulting in a *forest* of K decision trees $T_1, \ldots, T_K$ each trained with a random selection of segments of the training data. During the *testing phase*, an unseen MRI scan is obtained. After segmentation and feature extraction, we classify every high-dimensional feature vector f of a candidate segment by each of the K trees. Each tree T_q then determines a probability measure $P_{T_q}(f \in c_j)$ according to the distribution of training patterns in the terminal leaf node reached. These measures are integrated by taking their mean. Finally, based on this mean probability, a test segment is assigned with the class label c_j.

At this point, candidate segments are classified, but the classifications of overlapping aggregates from different scales may be contradicting. To obtain a result in terms of voxels, we apply the following procedure. For a voxel v and for each scale, we first use the interpolation weights to determine the aggregate to which it belongs with maximal weight. Then, we consider all the aggregates associated with v and take the maximal probability over these aggregates to be the probability of the voxel to be a lesion.

B. Complexity Analysis

The segmentation runtime is linear in the number of voxels with only several dozen of computer operations per voxel [15]. The complexity for generating a tree classifier is

$$O(d^2 N_s \log(N_s) + d^3 N_s + dN_s(\log(N_s))^2) \qquad (12)$$

where d is the number of features (30 for the single channel and 53 in the multichannel case) and N_s ($\leq$15 000) is the number of training patterns for one decision tree. The first term includes the number of operations required to construct the FLD generalized eigenvalue problem; the second term includes the number of operations required for solving it, and the third term refers to the number of operations necessary for optimal splitting of the training points in each tree node. Therefore, the training complexity is dominated by $O(dN_s(\log(N_s))^2)$ and the testing complexity is $O(d\log(N_s))$ per one test sample.

The method is fully automated. Two randomly chosen calibration scans from each type of data were used to determine all parameter values and these scans were not used later in either the training or testing experiments. The segmentation of the single- and multichannel experiments takes approximately 3 and 5 min per subject, respectively, on a standard Xeon 1.7 GHz PC. The training of the classifier takes up to 4 h for one multichannel experiment. The test phase takes about 4 min per subject.

V. Experiments: Application to MS

Next we present validation results on two types of MR datasets along with experiments analyzing the significance of the multiscale features. The first dataset is a multichannel triplet of PD-, T2-, and T1-weighted channels, which is similar to the type of data used in [2], [11], and [13]. The second is a single-channel fluid attenuated inversion recovery (FLAIR) dataset. The data were produced in the Scientific Institute Ospedale San Raffaele and was acquired on a SIEMENS Magnetom Vision 1.5 T MR scanner. The procedure for producing the lesion maps was the following: two neurologists by consensus identified hyperintense lesions on PD and FLAIR films in the multi- and single-channel experiments, respectively. Using the marked films as reference, one trained technician outlined the contours of the lesions using a segmentation technique based on local thresholding. The contours outlined from the technician have been then transformed into binary masks. Before the classification process, several constraints are applied to eliminate candidate segments whose properties differ considerably from those expected from a lesion. The same constraints were used in both the multichannel and the FLAIR experiments and included removing aggregates that include very dark regions (i.e., average intensity < 1 and neighborhood contrast < -0.25) and eliminating aggregates that are not contained in intracranial cavity (IC) based on registering the input MRI data to the ICBM atlas probability maps [19] using the SPM software [18].

A. Significance of Multiscale Features

The role of the various features in the segmentation task was evaluated using the decision forest classifier. The classification process is applied to three sets of segmentation scales: *small, intermediate, and large* segments corresponding to scales $2, 3$, and $\geq$4 in the graph pyramid, respectively. This separation was based on our experience indicating that small, intermediate, and large lesions share different attributes. Therefore, for each of these scales, we construct a separate forest consisting of $K = 50$ trees, trained with a random selection of N_s patterns. In each tree, the size of the random subset selection is determined by 75% of the class size, and twice of this amount for the nonclass, since there were many more nonclass aggregates. At each node in the tree, an optimal separating direction is computed using FLD, which includes coefficients for every one of the features. The features significance was measured by three different summations, over the absolute value of the feature vector coefficients in all the tree nodes: 1) *coefficient summation (Coef)*: uses an equal weight for all the tree nodes (ignoring the node position in the tree); 2) *probability summation (Prob)*: weights the coefficients by the probability of the tree node (proportion of training points at the tree node to the root); and 3) *maximum impurity decrease summation (MI)*: weights the coefficients by the MI decrease obtained by the split in the tree node.

The results for both the multichannel and FLAIR experiment are demonstrated in Fig. 1, presenting the features ordered by their significance in the multiple decision trees from left to right. The features selected correspond to the list in Table II, where subscript and superscript characters were removed or

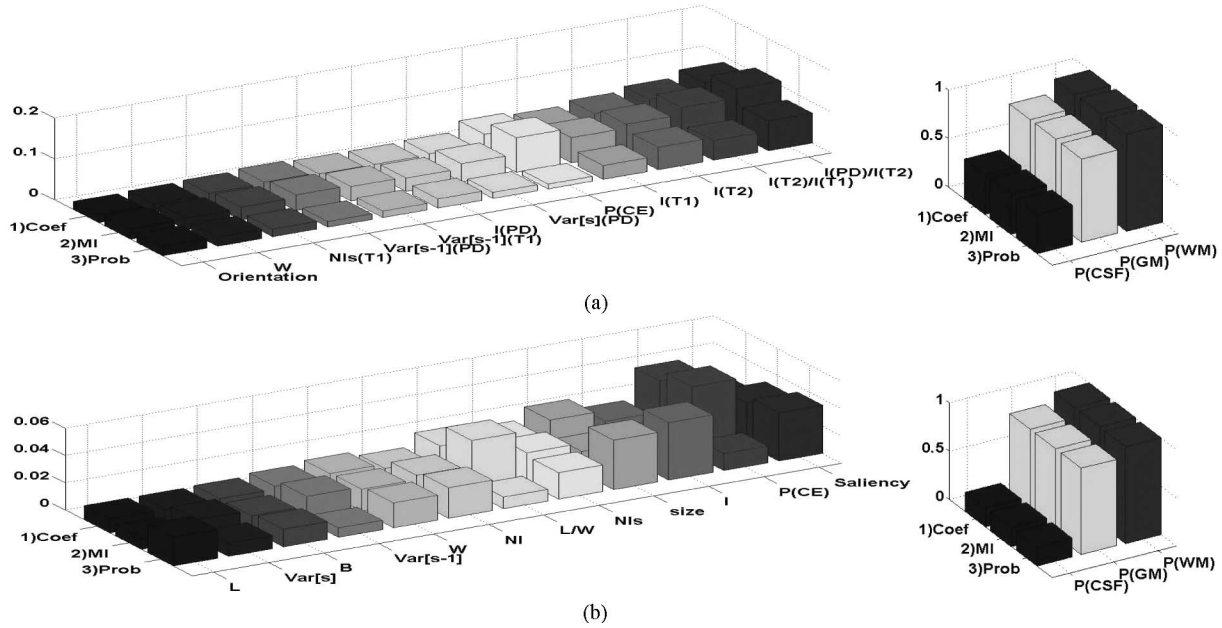


Fig. 1. Significance of features in intermediate scale experiment. The 15 most significant features in increased significance order from left to right for three different summation computations. The three most significant features (right) and the following 12 features (left). (a) Multi-channel: significant features in intermediate scale. (b) FLAIR: significant features in intermediate scale. (see Table I for feature description).

put in braces when the context was clear. The results lead to several conclusions. 1) As expected, prior anatomic knowledge is extremely significant for classification. 2) Commonly used features in MS segmentation, such as the aggregative properties of average intensity and variance of average intensities are also ranked high. 3) Novel multiscale regional features, such as the contrast to the neighborhood, or shape properties (e.g., width, length, and orientation) contribute to classification, as documented in literature on MS and brain segmentation [23]–[25]. 4) The results showed that different recognition tasks (e.g., multi- or single-channel experiment, or detection of lesions in different scales) lead to the selection of different features.

B. Validation

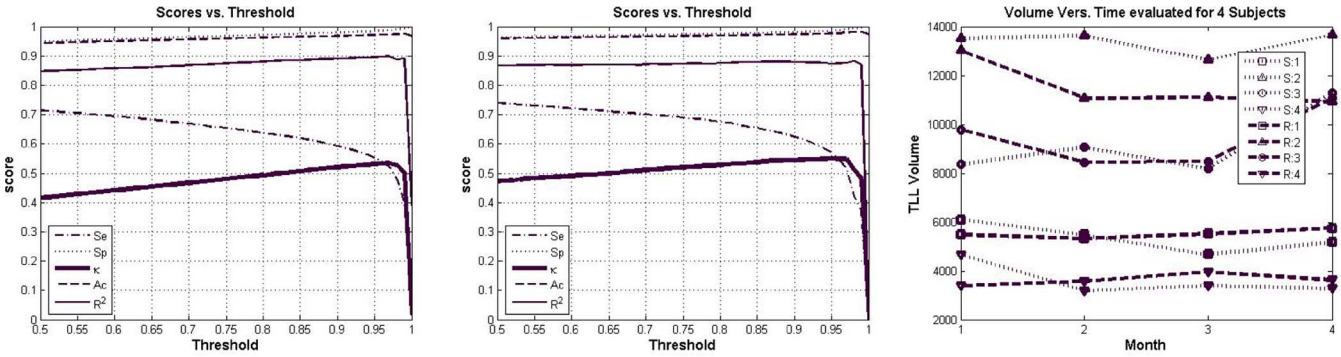
1) Validation Measures: Validation results are presented in terms of voxels. We denote the set of voxels detected as lesions by our automated process by S and in the "ground truth" expert reference by R. Following commonly used definitions [13], true positive voxels are the voxels common to both S and R ($\mathrm{TP} = |S \cap R|$). True negative voxels are all IC voxels not outlined as lesions by experts ($\mathrm{TN} = |IC \cap \bar{R}|$). False positives (FPs) are those detected in S but not by R ($\mathrm{FP} = |S \cap \bar{R}|$) and false negatives are those identified in R but not in S ($\mathrm{FN} = |\bar{S} \cap R|$). The validation measures used include the following.

1) *Sensitivity S_e:* True positive fraction $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$.
2) *Specificity S_p:* True negative fraction $\mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$.
3) *Accuracy Ac:* $(\mathrm{TN} + \mathrm{TP})/(\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP})$.
4) *Dice κ statistics:* $2|S \cap R|/(|S| + |R|)$.
5) *Correlation coefficient R^2:* Analysis between the total lesion load (TLL) detected in R and S.

2) Validation on Multichannel MR Data: The multichannel experiment included 25 patients (13 males and 12 females) aged 47 ± 9 years with secondary progressive (SP) MS. For each subject, the data consist of a dual-echo sequence that is a turbo spin-echo PD-/T2-weighted image pair (TR = 3300 ms; TE = 16/98 ms; echo train length = 5) and a spin-echo T1-weighted image (TR = 768 ms; TE = 15 ms). Each channel contains 24 contiguous axial slices with a pixel size of 0.98 mm × 0.98 mm, slice thickness 5 mm [field of view (FOV) = 250 × 250 mm; matrix = 256 × 256].

Ten experiments were conducted. In each experiment, 75% of the 25 patients were randomly selected for training. The test set consisted of the remaining patients of the multichannel set. Fig. 2(a) presents the average validation measures in terms of voxels over the ten experiments for the multichannel test in the entire brain. The automatic segmentation results is based on the voxel's probability to be a lesion. Therefore, for each experiment, we assessed the scores behavior with varying values of probabilities (ψ) (see Section IV-A). Table III lists several representative results where the rows correspond to $\psi = 0.5, 0.65, 0.8, 0.95$, respectively. In the first row, ψ exceeds the value of 0.5, thus the largest possible set of voxels detected as c_1 is obtained. The last row refers to $\psi \geq 0.95$ where the maximal κ point was obtained in both experiments. Additional motivation for using this graph will be given in Section V-C.

3) Validation on Flair MR Data: To evaluate the generalization ability of the algorithm, it was tested on single-channel FLAIR images. Such images are known for their high sensitivity to lesions, offering a diagnostic capability beyond other sequences. The experiment included 16 patients (5 males and 11 females) aged 43 ± 15 years with relapsing–remitting (14) or SP (2) MS. The parameters of the FLAIR sequence used to acquire the images were: TR = 9500; TE 105; inversion time = 2200; FOV = 250 × 250. The acquisition was interleaved. The voxel size used is 0.97 mm × 0.97 mm or 0.86 mm × 0.86 mm (for



(a) (b) (c)

Fig. 2. Validation and Precision Analysis of results. Scores as function of probability threshold (ψ) on (a) Multi-channel and (b) FLAIR MR Data (c) Comparison of TLL volume obtained by S and R over time on four subjects, exemplified on set B (FLAIR).

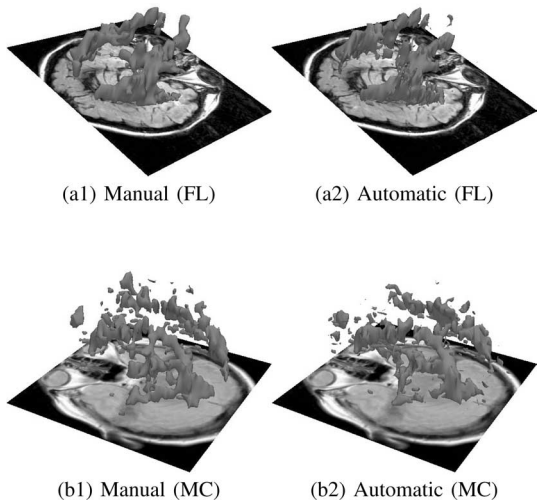
TABLE III
CLASSIFICATION OF MULTICHANNEL (MC) AND FLAIR (FL) SETS, AVERAGED OVER TEN EXPERIMENTS AND COMPARED TO OTHER APPROACHES

Approach	Probability (ψ)	Se	Sp	κ	Ac	R^2
MC:	0.5(**none**)	0.71 ± 0.11	0.95 ± 0.02	0.42 ± 0.09	0.94 ± 0.02	0.85
MC:	0.65	0.68 ± 0.11	0.96 ± 0.01	0.45 ± 0.09	0.95 ± 0.01	0.86
MC:	0.80	0.64 ± 0.12	0.97 ± 0.01	0.49 ± 0.09	0.96 ± 0.01	0.88
MC:	0.95(**optimal κ**)	0.55 ± 0.13	0.98 ± 0.01	0.53 ± 0.1	0.97 ± 0.01	0.90
FL:	0.5(**None**):	0.74 ± 0.1	0.96 ± 0.02	0.47 ± 0.07	0.96 ± 0.02	0.87
FL:	0.65:	0.71 ± 0.11	0.97 ± 0.02	0.50 ± 0.07	0.96 ± 0.02	0.87
FL:	0.80:	0.68 ± 0.11	0.98 ± 0.02	0.53 ± 0.07	0.97 ± 0.02	0.88
FL:	0.95(**Optimal κ**):	0.57 ± 0.14	0.99 ± 0.01	0.55 ± 0.09	0.98 ± 0.01	0.87
[11]:	-	-	-	$0.45, 0.51$	-	$0.96 - 0.98$
[2]:	-	-	-	0.6 ± 0.07	-	-
[13]:	-	$0.70 - 0.752$	$0.987 - 0.999$	-	$0.985 - 0.999$	$0.96 - 0.98$

six and ten subjects, respectively), with slice thickness 5 mm (24 slices). We divide the data as follows: set A includes examination of 12 patients and set B includes four additional patients who had a monthly follow-up, so that four time points were available for each patient.

Throughout the classification stage, ten experiments were conducted. In each experiment, nine patients from set A were randomly selected for training. The test set consists of the remaining patients of set A and all patients of set B. The average validation measures are presented in Fig. 2(b). Table III lists several representative results. Fig. 3 illustrates a 3-D view of MS lesions detected in the experiments on FLAIR and multichannel MRI data using the Slicer software available at "http://www.slicer.org/."

4) Volume Precision Over Time: We analyzed four sets of FLAIR images that were acquired over four months (set B). These datasets obtained validation results which are similar to the ones described in Section V-B3. Generally, tests for robustness of reproducibility analysis should be performed on data rescanned repeatedly from the same brain. Here, since the interval between two scans was not short, the volume may also vary due to actual changes in patient pathology. However, following the measure presented in [26], we performed a serial analysis and computed the ratio of volume difference between our detection and the "ground truth" divided by the mean of the two measurements. The analysis was performed based on the segmentation S obtained at the optimal κ point found in the FLAIR



(a1) Manual (FL) (a2) Automatic (FL)

(b1) Manual (MC) (b2) Automatic (MC)

Fig. 3. 3-D view of MS lesion manual segmentation overlayed on an axial (a1) FLAIR and (b1) PD-weighted slice compared to lesions detected automatically by the algorithm on (a2) FLAIR, and (b2) multi-channel data.

experiment (Section V-B3). The average ratio of volume difference over time for each of the four subjects were 0.1 ± 0.06, 0.15 ± 0.08, 0.07 ± 0.06, and 0.17 ± 0.1, respectively.

Fig. 2(c) presents for each subject the TLL detected by the automatic segmentation and the "ground truth" reference over four points in time. As shown in the graph, the algorithm does not always follow the direction of the change in TLL. However,

computing the average slope of the "ground truth" reference $(R_{t+1} - R_t)/\mathrm{mean}(R)$ over all four subjects shows very little changes in TLL (0.08 ± 0.08, mean $\pm$ S.D), as expected during four months.

C. Validation Analysis

Comparison to results reported in literature demonstrates the difficulty of the MS detection problem and reveal potential obtained by our approach. To our best knowledge, studies reporting extensive validation results for automatic MS segmentation are performed on multichannel data including T2-, PD-, and T1-weighted images only. Thus, both our results on multichannel and on FLAIR data are compared to results reported on multichannel data as presented in Table III. The best correspondence results reported on multichannel data were $\kappa = 0.45, 0.51$ in [11], for 5 mm, 3 mm slice thickness, respectively, with $R^2 = 0.96$–0.98, where a similarity index of $\kappa = 0.58$ was found between two human experts. An average $\kappa = 0.6 \pm 0.07$ was obtained in [2] with $R^2 = 0.93$. Agreement between experts appears to fall in the same range, since the authors found that the κ similarity between pairs of seven experts ranges from 0.51 to 0.67. Recent results published in [13] on several lesions subtypes were reported in terms of sensitivity (70–75.2%), specificity (98.7–99.9%), accuracy (98.5–99.9%), and $R^2 = 0.96$–0.98. The authors also report of correlation and agreement of lesion volume change over time ($R^2 = 0.715$).

Previous papers either provide a κ score [2], [11] or a sensitivity specificity score [13] but not both. We present the entire range for all measures in Fig. 2. Evaluation of the results shows that comparing κ results at its optimal point ($\psi = 0.95$) with papers that report κ values yields $\kappa = 0.53, 0.55$ for the multichannel and FLAIR experiments, respectively, which are higher than [11] but lower than [2]. Comparison of the values reported by [13] at the basic $\psi = 0.5$ level shows $Se = 0.71$, $Sp = 0.95$, $Ac = 0.94$, $R^2 = 0.85$ and $Se = 0.74$, $Sp = 0.96$, $Ac = 0.96$, $R^2 = 0.87$ for the multichannel and FLAIR experiment, respectively. These values are similar in sensitivity values and slightly lower in the specificity and accuracy values. Our correlation coefficients of TLL measured were lower compared to other studies. Yet, the correlations we obtained between manual and automatic measurements are all highly significant ($p < 0.0001$). Additionally, as noted in [11], the correlation coefficient measure does not take into account any spatial correspondence of the segmented lesions. Therefore, it should be considered with respect to spatial similarity metric that were comparable or not far from the state-of-the-art reported measures.

VI. Discussion

We developed a multiscale approach that combines segmentation with classification for detecting abnormal brain structures. Our study focuses on analyzing 3-D MRI brain data of MS patients.

The utility of our method was demonstrated in various experiments using different types of brain MR images. Comparison of our results to other automated MS segmentation methods yields similar κ and sensitivity values with lower specificity accuracy and correlation values. The results obtained with the multichannel data were lower than those obtained in the FLAIR experiment. This can be explained by FLAIR's higher sensitivity to MS lesions and higher specificity, which allows avoiding many of the FPs detected in the multichannel triplet. In particular, FLAIR is better able to suppress the cerebrospinal fluid (CSF) signal, leading to fewer FP in lesions near the ventricles and CSF containing sulci, especially when the CSF is partially volumed with nearby brain parenchyma.

Qualitative inspection of our results shows that our main errors are due to the FP rate. Preliminary assessment indicates that this extra volume is somewhat related to other white matter (WM) classes, e.g., "dirty-appearing" WM (DAWM) [27]. Moreover, in Section V-B4, we found that the algorithm may not be sensitive enough to detect small directional changes over time. An additional limitation of our experiments is due to the use of data labeled by a single rater. Applying our approach on data labeled by several raters and with higher resolution (e.g., 3 mm slice thickness) may lead to improved results, as reported in [11].

Our approach is flexible, with no restrictions on the MRI scan protocol, resolution, or orientation [28]. Unlike common approaches, our method is not limited to finding the lesions in the WM only [5], [8], [11], [29], risking the omission of subcortical lesions. Our learning process requires only a few training examples. The use of a large bank of features along with automatic feature selection can also be useful for other medical imaging applications. Furthermore, based on expert radiologists advice, and as in [11], we consider reporting results on a varying range of probabilities, as an additional benefit of the algorithm.

Future work will explore features that can characterize DAWM and MS lesions subtypes. Finally, we wish to extend our approach and apply it to other tasks and modalities in medical imaging.

Acknowledgment

The authors are grateful for the research conducted at the Moross Laboratory for Vision and Motor Control at the Weizmann Institute of Science.

References

[1] J. S. Duncan and N. Ayache, "Medical image analysis: Progress over two decades and the challenges ahead," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 85–106, Jan. 2000.

[2] A. P. Zijdenbos, R. Forghani, and A. C. Evans, "Automatic pipeline analysis of 3D MRI data for clinical trials: Application to MS," *IEEE Trans. Med. Imag.*, vol. 21, no. 10, pp. 1280–1291, Oct. 2002.

[3] W. E. L. Grimson, G. J. Ettinger, T. Kapur, M. E. Leventon, W. M. Wells, and R. Kikinis, "Utilizing segmented MRI data in image guided surgery," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 11, no. 8, pp. 1367–1397, 1997.

[4] A. Achiron, S. Gicquel, S. Miron, and M. Faibel, "Brain MRI lesion load quantification in multiple sclerosis: A comparison between automated multispectral and semi-automated thresholding computer-assisted techniques," *Magn. Reson. Imag.*, vol. 177, pp. 85–106, 2001.

[5] M. Kamber, R. Shinghal, D. L. Collins, G. S. Francis, and A. C. Evans, "Model-based 3-D segmentation of multiple sclerosis lesions in MR brain images," *IEEE Trans. Med. Imag.*, vol. 14, no. 3, pp. 442–453, Sep. 1995.

[6] J. K. Udupa, L. Wei, S. Samarasekera, Y. Miki, M. van Buchem, and R. I. Grossman, "Multiple sclerosis lesion quantification using

fuzzy-connectedness principles," *IEEE Trans. Med. Imag.*, vol. 16, no. 5, pp. 598–609, Oct. 1997.

[7] D. H. Miller and R. I. Grossman, "The role of magnetic resonance techniques in understanding and managing multiple sclerosis," *Brain*, vol. 121, pp. 3–24, 1998.

[8] S. K. Warfield, K. H. Zou, and W. M. Wells, "Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions," *J. Image Guided Surg.*, vol. 1, no. 6, pp. 326–338, 1995.

[9] X. Wei, S. K. Warfield, K. H. Zou, Y. Wu, X. Li, A. Guimond, J. P. Mugler, R. R. Benson, L. Wolfson, H. L. Weiner, and C. R. Guttmann, "Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy," *J. Magn. Reson. Imag.*, vol. 15, pp. 203–209, 2002.

[10] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 15, no. 4, pp. 429–442, Aug. 1996.

[11] K. Van-Leemput, F. Maes, D. Vandermeulen, A. Colcher, and P. Suetens, "Automated segmentation of multiple sclerosis by model outlier detection," *IEEE Trans. Med. Imag.*, vol. 20, no. 8, pp. 677–688, Aug. 2001.

[12] A. Shahar and H. Greenspan, "A probabilistic framework for the detection and tracking in time of multiple sclerosis lesions," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2004, pp. 440–443.

[13] Y. Wu, S. K. Warfield, I. Tan, W. M. Wells, D. S. Meier, R. van Schijndel, F. Barkhof, and C. R. Guttmann, "Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI," *NeuroImage*, vol. 32, no. 3, pp. 1205–1215, 2006.

[14] E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt, "Hierarchy and adaptivity in segmenting visual scenes," *Nature*, vol. 442, no. 7104, pp. 810–813, 2006.

[15] M. Galun, E. Sharon, R. Basri, and A. Brandt, "Texture segmentation by multiscale aggregation of filter responses and shape elements," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 716–723.

[16] T. A. Yousry, R. I. Grossmanb, and M. Filippi, "Assessment of posterior fossa damage in MS using MRI," *J. Neurol. Sci.*, vol. 172, pp. 50–53, 2000.

[17] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[18] R. S. J. Frackowiak, K. J. Friston, C. Frith, R. Dolan, C. Price, S. Zeki, J. Ashburner, and W. D. Penny, *Human Brain Function*. New York: Academic, 2003.

[19] J. C. Mazziotta, A. W. Toga, A. C. Evans, P. Fox, and J. Lancaster, "A probabilistic atlas of the human brain: Theory and rationale," *NeuroImage*, vol. 2, pp. 89–101, 1995.

[20] L. Breiman, J. H. Olshen, and C. J. Stone, Eds., *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[21] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[22] J. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[23] L. P. Clarke, R. P. Velthuizen, M. A. Camacho, J. J. Heine, M. Vaidyanathan, L. O. Hall, R. W. Thatcher, and M. L. Silbiger, "MRI segmentation: Methods and applications," *Magn. Reson. Imag.*, vol. 13, no. 3, pp. 343–368, 1995.

[24] D. Goldberg-Zimring, A. Achiron, C. R. Guttmann, and H. Azhari, "3D analysis of the geometry of individual ms lesion detection of shape changes over time using spherical harmonics," *J. Magn. Reson. Imag.*, vol. 18, pp. 291–301, 2003.

[25] A. Pitiot, H. Delingette, P. M. Thompson, and N. Ayache, "Expert knowledge guided segmentation system for brain MRI," *NeuroImage*, vol. 23, no. 1, pp. S85–S96, 2004.

[26] C. R. Guttmann, R. Kikinis, M. C. Anderson, M. Jakab, S. K. Warfield, R. J. Kiliany, H. L. Weiner, and F. A. Jolesz, "Quantitive follow-up of patients with multiple-sclerosis using MRI: Reproducibility," *J. Magn. Reson. Imag.*, vol. 9, pp. 509–518, 1999.

[27] Y. Ge, R. I. Grossman, J. S. Babb, J. He, and L. J. Mannon, "Dirty-appearing white matter in multiple sclerosis: Volumetric MRI and magnetization transfer ratio histogram analysis," *AJNR Amer. J. Neuroradiol.*, vol. 24, no. 10, pp. 1935–1940, 2003.

[28] A. Akselrod-Ballin, M. Galun, J. M. Gomori, R. Basri, and A. Brandt, "Atlas guided identification of brain structures by combining 3D segmentation and SVM," in *Proc. Med. Image Comput. Comput.-Assisted Intervention Conf.*, 2006, vol. II, pp. 209–216.

[29] B. Johnston, M. Atkins, B. Mackiewich, and M. Anderson, "Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI," *IEEE Trans. Med. Imag.*, vol. 15, no. 2, pp. 154–169, Apr. 1996.

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