An Integrated Segmentation and Classification Approach Applied to Multiple Sclerosis Analysis

Ayelet Akselrod-Ballin[1], Meirav Galun[1], Moshe John Gomori[2]
Massimo Filippi[3], Paula Valsasina[3], Ronen Basri[1] *, Achi Brandt[1] †
Dept. of Computer Science and Applied Math, Weizmann Institute of Science, Rehovot, Israel[1] ‡
Dept. of Radiology, Hadassah University Hospital, Jerusalem, Israel[2]
Neuroimaging Research Unit, Hospital San Raffaele, Milan, Italy[3]

Abstract

We present a novel multiscale approach that combines segmentation with classification to detect abnormal brain structures in medical imagery, and demonstrate its utility in detecting multiple sclerosis lesions in 3D MRI data. Our method uses segmentation to obtain a hierarchical decomposition of a multi-channel, anisotropic MRI scan. It then produces a rich set of features describing the segments in terms of intensity, shape, location, and neighborhood relations. These features are then fed into a decision tree-based classifier, trained with data labeled by experts, enabling the detection of lesions in 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 showing successful detections of lesions in both simulated and real MR images.

1. Introduction

Identifying 3D brain structures in medical imagery, particularly in MRI (Magnetic Resonance Imaging) scans, is important for early detection of tumors, lesions, and abnormalities, with applications in diagnosis, follow-up, and image-guided surgery. Computer aided analysis can assist in identifying brain structures, extract quantitative and qualitative properties of structures, and evaluate their progress over time. In this paper we present a novel method for detecting abnormal brain structures focusing on 3D MRI brain data containing scans of multiple sclerosis (MS) patients.

Automatic detection 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. Detection techniques based on template matching [4] or more recent techniques based on constellations of appearance features (e.g., [5]), which are common in computer vision, are not well suited to handle such amorphous structures. Consequently, with few exceptions (e.g., [11]) medical applications commonly approach this problem by applying classification algorithms that rely on a voxel-by-voxel analysis (e.g., [14, 15, 16, 17]). These approaches, however, are limited in their ability to utilize regional properties, particularly properties related to the shape, boundaries, and texture.

This paper introduces a novel multiscale approach that combines segmentation with classification to detecting abnormal 3D brain structures. Our method is based on a combination of a powerful multiscale segmentation algorithm, Segmentation by Weighted Aggregation (SWA) [12, 7], a rich feature vocabulary describing the segments, and a decision tree-based 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. At the same time, the rich hierarchical decomposition produced by the SWA algorithm allows us to a great extent to circumvent inaccuracies due to the segmentation process. Even when a lesion is not segmented properly we can generally expect to find some aggregate in the hierarchy that sufficiently overlaps it to allow classification.

We adapt the SWA algorithm to handle 3D multi-channel MRI scans and anisotropic voxel resolutions. These allow the algorithm to handle realistic MRI scans. The bank of features we use characterize each aggregate in terms of intensity, texture, shape, and location. These features were selected in consultation with expert radiologists. All the fea-

[*] Research was supported in part by the Binational Science foundation, Grant No. 2002/254.

[†] Research was supported in part by the Israel Institute of Technology.

[‡] Research was supported in part by the European Commission Project IST-2002-506766 Aim Shape. Research was conducted at the Moross Laboratory for Vision and Motor Control at the Weizmann Institute of Science.

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

tures are computed as part of the segmentation process, and they are used in turn to further affect the segmentation process. The classification step examines each aggregate and labels it as either lesion or non-lesion. This classification is integrated across scale to determine the voxel classification of the lesions. We demonstrate the utility of our method through experiments on simulated and real MRI data showing detection of MS lesions.

The paper is organized as follows. Section 2 presents the segmentation procedure, the feature extraction method and the classification model in our system. In section 3 results on simulated and real MRI data are presented. Section 4 follows with a discussion and conclusions.

2. Integrated system

This section describes our system for detecting abnormal brain structures. In a training phase our system obtains as input several MR scans along with a delineation of the lesions in these scans. The system uses segmentation to provide a complete hierarchical decomposition of the 3D data into regions corresponding to both meaningful anatomical structures and lesions. Each aggregate is equipped with a collection of multiscale features. Finally, a classifier is trained to distinguish between aggregates that correspond to lesions from those that correspond to non-lesions.

Once the classifier is trained we proceed to apply our approach to unlabeled test data. At this stage the system obtains as input an MRI scan of a single brain. It then segments the scan and extracts features to describe the aggregates. Finally, each aggregate is classified as either a lesion or a non-lesion, and the voxel occupancy of the lesions is determined.

One of the features we use to describe an aggregate is its location in the brain. To utilize this property we first bring each scan to a common coordinate system. In our implementation this was achieved using the SPM software package [6], which registers a scan to an atlas composed of subject average of 152 T1-weighted scans.

2.1. Segmentation

We use the Segmentation by Weighted Aggregation (SWA) algorithm [12, 7], which we extend to handle 3D multi-channel and anisotropic data. In this section we review the SWA algorithm along with our extensions.

2.1.1 Segmentation framework

Given a 3D MRI scan, a 6-connected graph $G = (V, W)$ is constructed as follows. Each voxel i is represented by a graph node i, so $V = \{1, 2, \ldots, N\}$ where N is the number of voxels. A weight is associated with each pair of neighboring voxels i and j. The weight w_{ij} reflects the contrast between the two neighboring voxels i and j

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

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, i.e $u_i = 1$ if $i \in S$, otherwise $u_i = 0$. The **saliency** Γ associated with S is defined by

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

which sums the weights along the boundaries of S divided by the internal weights. Segments which yield small values of $\Gamma(S)$ are considered salient. The matrix W includes the weights w_{ij}, and L is the Laplacian matrix of G. Our objective is to find those partitions characterized by small values of Γ. To find the minimal cuts in the graph we construct a coarse version of this graph. This coarse version is constructed so that we can use salient segments in the coarse graph to predict salient segments in the fine graph using only local calculations. This coarsening process is repeated recursively, constructing a full pyramid of segments (Fig. 1). Each node at a certain scale represents an **aggregate** of voxels. Each **segment** S, which is a salient aggregate (i.e., $\Gamma(S)$ is low), emerges as a single node at a certain scale. The coarsening procedure proceeds recursively as follows. Starting from the given graph $G^{[0]} \overset{def}{=} G$, we create a sequence of graphs $G^{[1]}, \ldots, G^{[k]}$ of decreasing size (Fig. 1). As in the general AMG setting [1], the construction of a coarse graph from a fine one is divided into three stages: first a subset of the fine nodes is chosen to serve as the **seeds** of the aggregates (the latter being the

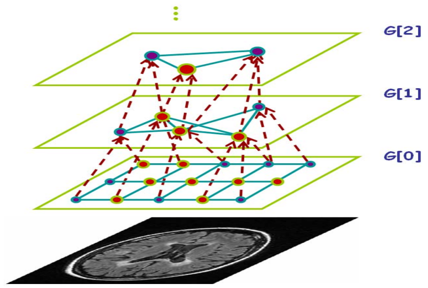


Figure 1. *Illustration of the irregular pyramid notion. The image presents 3 graph levels above one slice from the entire 3D MRI.*

nodes of the coarse graph). Then, the rules for interpolation are determined, establishing the fraction of each non-seed node belonging to each aggregate. Finally, the weights of the edges between the coarse nodes are calculated.

Coarse seeds: 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 Sec. 2.2) transfer nodes from F to C until all the remaining $i \in F$ satisfy $\sum_{j \in C} w_{ij} \geq \beta \sum_{j \in V} w_{ij}$, where β is a parameter (in our experiments $\beta = 0.2$).

The coarse problem: We define for each node $i \in F$ a coarse **neighborhood** $N_i = \{j \in C, w_{ij} > 0\}$. Let $I(j)$ be the index in the coarse graph of the node that represents the aggregate around a seed whose index at the fine scale is j. An **interpolation** matrix P (of size $N \times n$, where $n = |C|$) is defined by

$$P_{iI(j)} = \begin{cases} \frac{w_{ij}}{\sum_{k \in N_i} w_{ik}} & \text{for } i \in F, j \in N_i \\ 1 & \text{for } \ i \in C, j = i \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

This matrix satisfies $u \approx PU$, where $U = (U_1, U_2, ..., U_n)$ is the coarse level state vector. P_{iI} represents the likelihood that an aggregate i at a fine level belongs to an aggregate I at a coarser level. Finally, an edge connecting two coarse aggregates p and q is assigned with the weight:

$$w_{pq}^{coarse} = \sum_{k \neq l} P_{kp} w_{kl} P_{lq}. \qquad (4)$$

Denoting the scale by a superscript $G^{[s]} = (V^{[s]}, W^{[s]})$. Note that since $u^{[s-1]} \approx Pu^{[s]}$, the relation Eq. (2) inductively implies that a similar expression approximates Γ at all levels. However, $W^{[s]}$ is modified to account for aggregative properties (Sec. 2.2). We modify $w_{pq}^{[s]}$ between a pair of aggregates p and q at scale s by multiplying it with an exponentially decreasing function of the their aggregative properties distance. Table 1 summarizes the segmentation algorithm.

2.1.2 Handling anisotropic data

Common MRI data is anisotropic, less vertically resolved. The SWA algorithm, however, assumes that the voxels in the fine level are equally spaced. 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 2D slice separately in performing seed selection and inter-scale interpolation (steps 1-2 in Table 1), allowing non-zero interpolation weights only between nodes of the same slice. The rest of the steps (steps 3-5 in Table 1) are performed on the full 3D graph, i.e., taking into account inter-slice connections. This procedure is repeated until the inner- and inter-slice distances are approximately equal. Subsequent coarsening steps consider the full 3D graph

- Given a 3D MRI construct a 6-connected graph $G^{[0]}$
- 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 (3).
  3. Calculate $W^{[s]}$ by Eq. 4.
  4. For each $v \in V^{[s]}$ calculate aggregative properties (Sec. 2.2).
  5. Modify $W^{[s]}$ according to aggregative properties.

Table 1. *Outline of the 3D segmentation algorithm*

For example, consider data with 5_{mm} slice thickness versus $1_{mm} \times 1_{mm}$ in-slice resolution. Every coarsening step of the SWA algorithm typically reduces the number of nodes by a factor of 2.5-3. Consequently, if we apply the algorithm to a 2D slice, the distance between neighboring nodes in a slice grows at every level by a $\sqrt{2.5}$-$\sqrt{3}$ factor on average, so three coarsening steps are needed to bring the inner- and inter-slice distances to be roughly equal.

2.1.3 Multi-channel segmentation

A major aspect of MR imaging is the large variety of pulse sequences that can be applied. These sequences produce different images for the same tissue, highlighting different properties of the tissue. We incorporate multi-channel data in the algorithm in a fairly straightforward manner. Given a multi-channel scan, each voxel now includes a vector of intensities. The initialization step (Eq. 1) is modified to determine the initial weights utilizing intensity information from all m channels as follows:

$$w_{ij} = exp - \left(\frac{\sum_{c=1}^{m} (\alpha_c)^2 (I_i^c - I_j^c)^2}{\sum_{c=1}^{m} (\alpha_c)^2}\right)^{\frac{1}{2}} \qquad (5)$$

where α_c are pre-determined constants ($\alpha_{T2} = 15, \alpha_{PD} = \alpha_{T1} = 10$) and I_i^c is the intensity of voxel i in channel c. In addition, we maintain different sets of aggregative features for every channel (see Sec. 2.2 below) and use these properties to modify the edge weights at coarser levels.

2.2 Feature extraction

Lesions can often be characterized by properties of aggregates that emerge at intermediate scales, and are difficult to extract by any uni-scale procedure. Such properties may include, for instance, intensity homogeneity, principal direction of the lesion, and intensity contrast with respect to

- **Saliency:** Γ (Eq. 2)

Intensity statistics:

- **Average intensity:** of voxels in aggregate k, denoted $\bar{I}^{[0]}$.
- **Maximum intensity:**$\mu_k^{[2][s]}$ maximal average intensity of the sub-aggregates at scale 2.
- **Variance of average intensities of scale r:** $V^{[r]} = \bar{I}^{2[r]} - (\bar{I}_k^{[0]})^2$, where $\bar{I}^{2[r]}$ denotes the average of $(\bar{I}_l^{[0][r]})^2$ for all sub-aggregates l of k at scale r.
- **Average of variances:** of scale r denoted $\bar{\nu}^{[r]}$ where $\nu_k^{[r][r]} = V^{[0][r]}$.

Shape:

- **Volume:** $m^{[0]}$ is the aggregate volume in voxel units.
- **Location:** $\bar{x}^{[0]}$, $\bar{y}^{[0]}$, $\bar{z}^{[0]}$.
- **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.
- **Intensity moments:** averages of products of the intensity and the coordinates of voxels in aggregate k, denoted $\overline{Ix}^{[0]}, \overline{Iy}^{[0]}, \overline{Iz}^{[0]}$.

Neighborhood statistics:

- **Boundary surface area:** denoted B_{kl}. B_{kl} refers to the surface area of the common border of aggregates k and l. It is accumulated by weighted aggregation such that all the weights on the finest graph are set to 1.
- **Neighborhood Contrast:** defined as the difference between the average intensity of a segment and its **neighborhood average intensity**, formulated as: $< Constrast >_k = \bar{I_k}^{[0]} - \frac{\sum_l B_{kl} \bar{I_l}^{[0]}}{\sum_l B_{kl}}$

Table 2. *Aggregative features for an aggregate k*

neighboring tissues. Voxel-by-voxel analysis is limited in the ability to utilize such scale-dependent properties.

We refer to such properties as *aggregative features*. The weighted-aggregation scheme provides a recursive mechanism for calculating such properties along with the segmentation process. We use these properties for two purposes. First, we use these aggregative properties to affect the construction of the segmentation pyramid. Second, these properties are available for the classification procedure below (Sec. 2.3).

2.2.1 Aggregative features

For an aggregate k at scale s we express an aggregative property as a number reflecting the weighted average of some property q emerged at a finer scale r, ($r \leq s$). For example, the average intensity of k is an aggregative property, since it is the average over all intensities measured at 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 below) or by taking averages over aggregative properties of finer scales (e.g., average of variances below). We denote such a property by $Q_k^{[r][s]}$, and shorten this to $Q^{[r]}$ when the context is clear.

In addition to these properties we can define binary aggregative properties, reflecting relations between two aggregates k and l at scale s. Such properties, denoted by Q_{kl}, 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 aggregative properties of an aggregate k are in fact averages over its sub-aggregates properties. Such properties can be accumulated from one level of scale to the next with the interpolation weights determining the relative weight of every sub-aggregate. For a detailed description on the accumulation of such properties see [7].

Construction of the classifier based on these features requires consideration of the inter-subject and intra-subject variability, therefore all features were normalized for each brain. Table 2 lists the features for aggregate k at scale s. The features were selected based on interaction with expert radiologists. However, the effect of each feature in classification is determined by an automatic learning process.

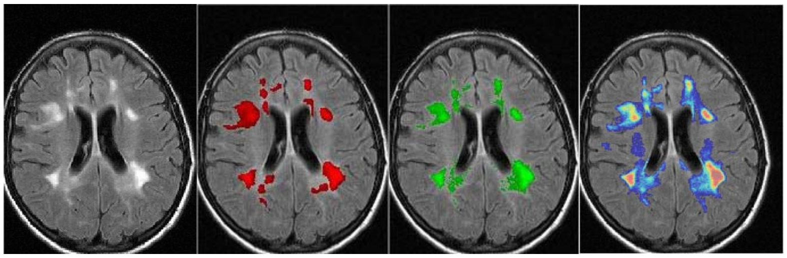
2.3 Classification

Once the MRI scan is segmented and features are computed, so that each aggregate is characterized by a high-dimensional feature vector f (see Table 2), we proceed to the classification stage. A classifier utilizing multiple decision trees [2] is trained using labeled data. Then, given an unlabeled scan the classifier is used to detect the lesions. Below the classification is described.

2.3.1 Multiple decision trees

To construct the decision tree classifier, a learning process is applied using MRI scans with MS lesions delineated by experts. The process obtains two kinds of data. (1) A collection of M candidate segments, $Cand = \{f_1, \ldots, f_M\}$, each is described by a d-dimensional feature vector (each feature is normalized to have zero mean and unit variance), and (2) a mask indicating the voxels marked as lesions by an expert . Since many of the candidate segments may contain a mixed collection of lesion and non-lesion voxels we label as a lesion a segment in which $\geq 70\%$ of its voxels were marked by an expert as lesion. We denote this class by c_1. We further mark as non-lesions only those segments which do not contain lesion voxels at all and denote this class by c_2. The rest of the segments are ignored at the training stage.

We next use the training data to construct multiple decision trees. A subset of the segments are randomly selected and used to construct a tree from the root downwards. At the root node all the labeled segments are considered and



(a)FLAIR (b)Ground-Truth (c)Automatic (d) Soft classifiation

Figure 2. *MS-lesion detection. From left to right: the original data(a), the expert labeling (b) the automatic segmentation (c) and the full range of soft classifiation (d) overlaid on a FLAIR slice. The different colors in (d) refer to different normalized intensity levels (ranging from blue to red).*

are repeatedly split into two subsets. At each tree node we apply a Fisher Linear Discriminant (FLD) [4] to the data determining the optimal separation direction and threshold s that leads to a maximal impurity decrease. This training procedure results in a **forest** of K decision trees $T_1, \ldots, T_K$ each trained with a random selection of segments.

During the testing phase an unseen MRI scan is obtained. After segmentation and feature extraction we classify every segment f 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 $\frac{1}{K}\sum_{q=1}^{K} P_{T_q}(f \in c_j)$. Finally, a test segment is assigned with the label c_j that maximizes this mean.

2.3.2 Classification of voxels

The classification process is applied to three segmentation scales, corresponding to *small, intermediate, and large* segments respectively. For each of these scales we construct a separate forest consisting of $K = 100$ trees, trained with a random selection of $N_s \leq 3000$ patterns. The candidate segments for classification may overlap, so that a voxel may belong to more than one segment. To measure the total lesion load (TLL) it is necessary to generate a result in terms of voxels.

The classifier labels the candidate segments as lesion or non-lesion with some probability (Sec. 2.3.1). All candidates are projected onto the data voxels using the interpolation matrix. Therefore, the interpolation matrix (eq. 3) determines an association weight for each voxel and candidate. A voxel belongs to a candidate if the corresponding association weight ≥ 0.5. The maximum probability over all candidates to which the voxel belongs, determines the probability of the voxel to be a lesion. We further employ both a **hard** and a **soft classification of voxels**. In the hard classification a voxel is classified as a lesion if its probability to be a lesion ≥ 0.5. However, since the 'ground truth' of the lesions may vary among different experts it might be helpful to provide a **soft classification** of the candidates rather than just a binary result. To create the soft classification, each 2D slice is first normalized by the average intensity of the intra-cranial cavity (ICC) in the related 2D slice. Then, by selecting from the hard assignment only voxels with normalized values above a certain threshold (1.75, 1.3 for multi-channel, FLAIR data respectively) one can determine a specific soft assignment, which we denote as automatic classification result.

3 Application to Multiple Sclerosis (MS)

Below we present validation results of employing our integrated system to both simulated and real MR data.

Before applying classification we eliminate candidates whose properties differ considerably from those expected from a lesion. Those include very non salient regions (saliency>7), very large regions (volume>5000 voxels), regions located very close to the midsagittal plain ($|x| < 6$), and very dark regions (intensity < 0.75 and contrast to neighborhood < -0.25, where both are divided by the average ICC intensity). In addition we eliminate aggregates that overlap with anatomical structures where as a rule lesions do not develop. Those include the eyes and the cerebrospinal fluid (CSF). To identify those structures we currently mark the segments corresponding to those structures manually. We further use the automatic skull stripping utility (Brain Extraction Tool [13]) to identify the brain region and eliminate segments that exceed beyond these regions.

The segmentation complexity is linear in the number of voxels. The complexity for generating a tree classifier is $O(d^2 N_s \log(N_s) + d^3 N_s + dN_s(\log(N_s))^2)$ and dominated by $O(dN_s(\log(N_s))^2)$, where N_s is the number of training patterns and d is the number of features. The testing complexity is $O(d\log(N_s))$ per one test sample.

3.1 MR simulator data

We first present results of our integrated system on the Simulated Brain Database (SBD) provided by the McConnell Brain Imaging Center ([3]). Currently, the SBD contains three simulated models of brains (phantoms) with 'mild', 'moderate', and 'severe' levels of MS lesions. We tested our approach on the three MS phantoms each including T1, T2 and PD images (see figure 3) using the default parameters ("normal" [17]): voxel size $1mm^3$, SD of noise 3% and intensity nonuniformity (INU) 20%. The multi-channel experiment was performed on the three channels for 30 slices, which contain 80% of the lesion load. The MS lesions presented in these models are not symmetric between the left and right lobes. Training was performed on the right half of all three brain models and testing on the left half of the brains, where the midpoint was defined by the midsagittal plain. The **detection rate** measures the percentage of correct classifications of candidate segments in the test set (see definitions in sec. 2.3.1). The classification forests of the segments test set on all scales obtained a detection rate of (1,0.99,0.99) for the lesion class (c_1), non-lesion class (c_2) and total candidate set respectively.

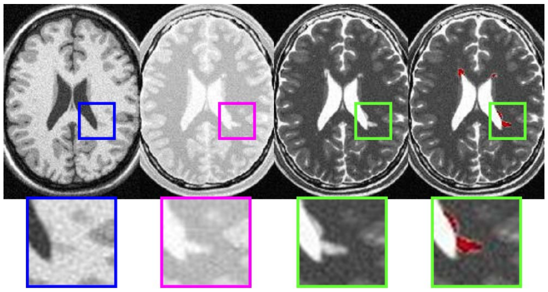


Figure 3. *Multi-channel data. From left to right T1, PD, T2, 'ground-truth' overlaid on the T2 image (red). Below, magnifications of the lesion area.*

Denote (S) as a set of voxels detected as lesions by our system and (R) as the set of voxels labeled as MS lesions in the 'ground truth' reference. n_S,n_R denote the number of connected components (lesions) in S and R correspondingly. Table 3 lists **classification measures** which are commonly used (e.g., [9],[14],[17]). These measures are presented in Table 4 and Table 6. Table 4 shows results obtained after overlaying the candidates from all scales detected as MS by the forest classifiers (see sec. 2.3.2).

To compare our results with other methods we applied the automatic classification of the detected area using one specified threshold for all subjects (Sec. 2.3.2). We obtained an average of $\kappa = 0.80 \pm 0.11$ (mean±S.D) on all three phantoms. In comparison, the authors in [17] tested their pipeline on the simulated data with varying levels of noise and INU. Their best classification accuracy reported for the single condition with the same parameters used in our tests was 0.81.

- **#Hits:** n_S/n_R
- **Overlap:** $|S \cap R|/|R|$. Number of voxels in the intersection divided by the number of voxels in R.
- **FP rate:** $|S \cap \bar{R}|/|R|$.
- **Disconnected FP (DFP) rate:** Number of voxels in extra volume which are disconnected to any ground-truth lesion divided by $|R|$.
- **Similarity measure:** $\{\kappa\} = 2|S \cap R|/(|S| + |R|)$

Table 3. *Classification measures*

Set	#Hit	Overlap	FP	DFP	κ
Mild:	0.74	0.87	1.1	0	0.67
Moderate:	0.85	0.98	0.83	0.04	0.86
Severe:	0.93	0.98	1.02	0.01	0.87
Mean	0.84	0.94	0.99	0.02	0.8
SD	0.1	0.06	0.14	0.02	0.11

Table 4. *Phantom classification measures for each model separately, summarizing with the mean and S.D results on all three models.*

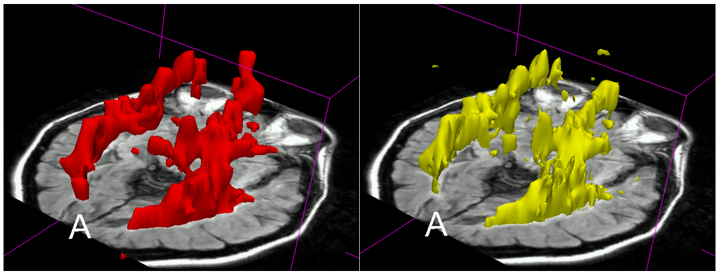
3.2 Real MR Data

To further evaluate our approach on clinical images, which reflect the full range of pathological variability, we tested our algorithm on real MR data [10].

This study consists of 16 subjects for which MS lesions were manually traced by a human expert. In this case we used single channel FLAIR images which are known for their high sensitivity to lesions, offering a diagnostic capability beyond other sequences. The voxel size used is $0.97_{mm} \times 0.97_{mm}$ or $0.86_{mm} \times 0.86_{mm}$ (for 6 and 10 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 examinations of four additional patients which had a monthly follow up, so that four time points were available for each patient.

3.2.1 Validation Results

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. In each one of the ten experiments three multiscale forests were generated. Table 5 presents average detection rates for



(a) 'Ground-Truth' (b)Automatic Segmentation

Figure 4. *3D view of MS lesions detected. Comparison of expert labelling with automatic segmentation overlayed on an axial FLAIR slice.*

Scale	lesion	non-lesion	Total
Small	0.90 ± 0.02	0.97 ± 0.01	0.97 ± 0.01
Interm	0.95 ± 0.02	0.98 ± 0.01	0.98 ± 0.01
Large	0.97 ± 0.02	0.98 ± 0.01	0.978 ± 0.004

Table 5. *Detection rates obtained on real data over ten randomized experiments.*

each scale over ten experiments. Table 6 lists the average classification measures over the ten experiments for test sets A and B. We also assessed the significance of correlation coefficient between the TLL volume detected by expert and automatic segmentation for each set. The two upper rows in Table 6 demonstrate the results obtained for superior slices (above the eyeballs) where on average 0.88 ± 0.05 of lesion volume occurs. The results in two lower rows were obtained on all slices. They are slightly lower due to the many artifacts in FLAIR data found in inferior slices.

Comparing to results reported in literature demonstrates the difficulty of the MS detection problem and reveals the high accuracy obtained by our approach. Correspondence results reported in [14] on multi-channel data were $\kappa = 0.45, 0.51$, for $5_{mm}, 3_{mm}$ slice thickness respectively. In [17] the average $\kappa = 0.6 \pm 0.07$, whereas the κ similarity between pairs of 7 experts ranges from 0.51 to 0.67.

Over superior slices, our average $\kappa \geq 0.64$. Results for all slices is comparable to the state-of-the-art ($\kappa \geq 0.6$). The extra volume exhibited by high FP measure should be further explored. In our experiments, the main extra volume usually surrounds the lesion volume and the DFP is significantly small compared to the FP. Preliminary assessment of our results indicates that this extra volume is somewhat related to other WM classes (e.g. 'dirty-appearing' WM DAWM [8]). Moreover, the delineation of lesion volume varies significantly between different experts, i.e, volume ratios reported in literature may exceed 1.5 and even approach 3 ([14, 16, 17]). Therefore, we may conclude that the FP measure is in the range of the inter-rater variability.

3.2.2 Volume Precision Over Time

We analyzed four sets of images that were acquired over four months (set B). 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 we performed a serial analysis and computed the ratio of volume difference between our detection and the ground-truth divided by the ground truth volume. The average results over time for the four subjects were $(0.1 \pm 0.05, 0.06 \pm 0.06, 0.08 \pm 0.04, 0.39 \pm 0.11)$ respectively. For the last subject significantly worse results were obtained probably due to the considerably smaller TLL relative to the other three subjects.

4 Discussion

We have presented a novel multiscale approach that combines segmentation with classification for detecting abnormal 3D brain structures. Our focus was on analyzing 3D MRI brain data containing brain scans of multiple sclerosis patients. Our method is based on a combination of a powerful multiscale segmentation algorithm, a rich feature vocabulary describing the segments, and a decision tree-based 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.

We adapted the multiscale segmentation algorithm to handle 3D multi-channel MRI scans and anisotropic voxel resolutions. The rich set of features employed were selected in consultation with expert radiologists. All the features are computed as part of the segmentation process, and

Slices	Test set	#Hit	Overlap	FP	DFP	κ	corr. significance
Superior	A:	0.85 ± 0.1	0.91 ± 0.05	1.53 ± 0.72	0.22 ± 0.21	0.64 ± 0.07	$p < 0.005$
	B:	0.83 ± 0.08	0.93 ± 0.02	1.36 ± 0.33	0.12 ± 0.12	0.66 ± 0.05	$p < 0.005$
All	A:	0.82 ± 0.09	0.89 ± 0.05	1.67 ± 0.71	0.36 ± 0.33	0.6 ± 0.07	$p < 0.005$
	B:	0.80 ± 0.08	0.91 ± 0.02	1.37 ± 0.39	0.18 ± 0.16	0.62 ± 0.06	$p < 0.005$

Table 6. *Classification measures for real MR sets, averaged over ten experiments.*

they are used in turn to further affect the segmentation process. The classification step examines each aggregate and labels it as either lesion or non-lesion. This classification is integrated across scale to determine the voxel occupancy of the lesions. We have demonstrated the utility of our method through experiments on simulated and real MRI data, including several modalities (T1, T2, PD and FLAIR). Comparison of the results to other automated segmentation methods applied to Multiple Sclerosis shows the high accuracy rates obtained by our system.

Our approach is flexible with no restrictions on the MRI scan protocol, resolution, or orientation. Unlike common approaches our method does not require a full brain tissue classification into white matter (WM), gray matter (GM), and cerebro-spinal fluid (CSF), and it is not limited to finding the lesions in the WM only, risking the omission of subcortical lesions.Furthermore, our learning process requires only a few training examples as shown specifically in the experiments.

We believe that our method can further be improved by better exploiting the rich information produced by the segmentation procedure. We plan to explore other features that can characterize lesions, as well as features that can characterize dirty appearing white matter (DAWM). Also of importance is to incorporate prior knowledge of anatomic structures into the framework using a brain atlas. Finally, we wish to extend our approach and apply it to other tasks and modalities in medical imaging.

References

[1] A. Brandt, S. McCormick, and J. Ruge, editors. *Algebraic multigrid (AMG) for automatic multigrid solution with application to geodetic computations*. Inst. for Computational Studies, POB 1852, Fort Collins, Colorado, 1982.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

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

[4] J. R. Duda and P. Hart, editors. *Pattern classification and scene analysis*. John Wiley and Sons, New York, 1973.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *CVPR*, 2003.

[6] S. Frackowiak, K. Friston, C. Frith, R. Dolan, C. Price, S. Zeki, J. Ashburner, and W. Penny, editors. *Human Brain Function*. Academic Press, 2003.

[7] M. Galun, E. Sharon, R. Basri, and A. Brandt. Texture segmentation by multiscale aggregation of filter responses and shape elements. *ICCV*, pages 716–723, 2003.

[8] Y. Ge, R. Grossman, J. Babb, J. He, and L. Mannon. Dirty-appearing white matter in Multiple Sclerosis: volumetric MRI and magnetization transfer ratio histogram analysis. *AJNR Am J Neuroradiol*, 24(10):1935–40, 2003.

[9] G. Gerig, M. Jomier, and M. Chakos. Valmet: A new validation tool for assessing and improving 3d object segmentation. *MICCAI*, pages 516–523, 2001.

[10] M. Rovaris, M. Rocca, T. Y. I. Yousry, B. Colombo, G. Comi, and M. Filippi. Lesion load quantification on fast-flair, rapid acquisition relaxation-enhanced, and gradient spin echo brain mri scans from multiple sclerosis patients. *MRI*, 17(8):1105–10, 1999.

[11] A. Shahar and H. Greenspan. A probabilistic framework for the detection and tracking in time of Multiple Sclerosis lesions. *IBSI*, 2004.

[12] E. Sharon, A. Brandt, and R. Basri. Segmentation and boundary detection using multiscale intensity measurements. *CVPR*, pages 469–476, 2001.

[13] S. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

[14] K. Van-Leemput, F. Maes, D. Vandermeulen, A. Colcher, and P. Suetens. Automated segmentation of ms lesions by model outlier detection. *IEEE MI*, 20:677–688, 2001.

[15] S. Warfield, K. Zou, and W. M. Wells. Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions. *J. of image guided surgery*, 1(6):326–338, 1995.

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

[17] A. Zijdenbos, R. Forghani, and A. Evans. Automatic pipeline analysis of 3d MRI data for clinical trials: application to MS. *IEEE MI*, 21:1280–1291, 2002.