Bayesian Classification of Multiple Sclerosis Lesions in Longitudinal MRI Using Subtraction Images*

Colm Elliott[1], Simon J. Francis[2], Douglas L. Arnold[3], D. Louis Collins[2], and Tal Arbel[1]

[1] Centre for Intelligent Machines, McGill University, Canada
[2] Montreal Neurological Institute, McGill University, Canada
[3] NeuroRx Research, Montreal, Canada

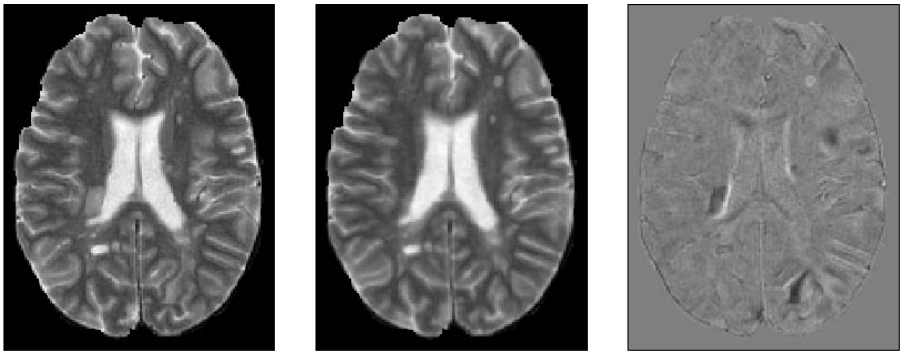
Abstract. Accurate and precise identification of multiple sclerosis (MS) lesions in longitudinal MRI is important for monitoring disease progression and for assessing treatment effects. We present a probabilistic framework to automatically detect new, enlarging and resolving lesions in longitudinal scans of MS patients based on multimodal subtraction magnetic resonance (MR) images. Our Bayesian framework overcomes registration artifact by explicitly modeling the variability in the difference images, the tissue transitions, and the neighbourhood classes in the form of likelihoods, and by embedding a classification of a reference scan as a prior. Our method was evaluated on (a) a scan-rescan data set consisting of 3 MS patients and (b) a multicenter clinical data set consisting of 212 scans from 89 RRMS (relapsing-remitting MS) patients. The proposed method is shown to identify MS lesions in longitudinal MRI with a high degree of precision while remaining sensitive to lesion activity.

1 Introduction

The use of subtraction imaging to identify MS lesion activity on MRI has been shown to increase sensitivity to new and resolving lesions and significantly reduce inter-rater variability [1,2,3,4]. Previous studies using subtraction images for lesion identification were done in a manual or semi-automatic fashion, as the automatic analysis of subtraction images is complicated by the presence of registration errors, flow artifacts and the high variability of signal intensities for lesions [3,5]. For this reason, most automated longitudinal MS lesion segmentation approaches have used more robust statistical approaches that require the inclusion of images from several timepoints [6,7,8], the analysis of lower resolution patches [9], or some form of deformation analysis [5,10]. The automatic lesion classifier presented here attempts to overcome the limitations of subtraction imaging by embedding intensity differences into a Bayesian framework that also incorporates prior classification at a reference timepoint, models that account for registration error and noise, and neighbourhood information. Our probabilistic

* This work was supported by NSERC Strategic Grant (350547-07).

T. Jiang et al. (Eds.): MICCAI 2010, Part II, LNCS 6362, pp. 290–297, 2010.



(a) (b) (c)

Fig. 1. T2 images for 2 timepoints are shown in (a) and (b) while the subtraction image between the two timepoints is shown in (c). New lesions appear as hyperintense on the T2 subtraction image while resolving (disappearing) lesions appear as hypointense.

framework further permits qualification of the degree of confidence to classification results at each voxel in the form of a posterior probability for each tissue class.

Our method was evaluated on (a) a scan-rescan data set consisting of 3 MS patients and (b) a multicenter clinical data set consisting of 212 scans from 89 RRMS patients with 2-4 longitudinal scans each. The overall classification system provides a consistent labelling of lesion voxels while remaining sensitive to lesion activity.

2 Methods

2.1 Problem Formulation

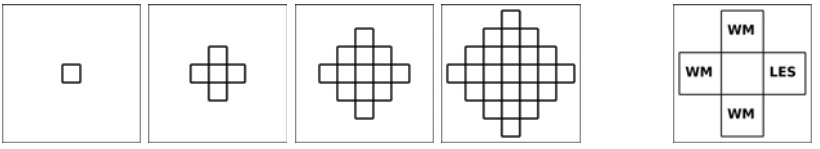
We present the problem of classification as one of inferring a tissue class label, $C_i^{(t)}$, at each voxel i of a multimodal volume at timepoint t, given an image from a reference timepoint, $I^{(r)}$, and a subtraction image, $D^{(t)}$, between timepoint t and the reference timepoint. Tissue class labels are restricted to one of cerebrospinal fluid (csf), gray matter (gm), white matter (wm), MS lesion (les) and a partial volume class (pv).

We first formulate the tissue class inference problem by only considering observations at the voxel in question:

$$p_i^0 = p(C_i^{(t)}|I_i^{(r)}, D_i^{(t)}) = \sum_{C_i^{(r)}} p(C_i^{(r)}, C_i^{(t)}|I_i^{(r)}, D_i^{(t)})$$

$$= \frac{1}{K} \sum_{C_i^{(r)}} p(D_i^{(t)}|C_i^{(r)}, C_i^{(t)}, I_i^{(r)}) p(C_i^{(t)}|C_i^{(r)}) p(C_i^{(r)}|I_i^{(r)}), \quad (1)$$

where we have assumed that $C_i^{(t)}$ is conditionally independent of $I_i^{(r)}$ given $C_i^{(r)}$. The right side of equation (1) can be seen as a product of three terms: a



(a) $\mathcal{N}_i^k$ for k=0-3 (b) $\{C_{\mathcal{N}_i}^{(t)}\}$

Fig. 2. (a) shows $\mathcal{N}_i^k$ for k=0-3 while (b) shows a sample tissue label configuration, where $\mathcal{N}_i$ is defined as the 4-voxel neighbourhood in both (a) and (b)

difference likelihood, a tissue transition likelihood, and a prior classification of our reference image. We incorporate information from neighbouring voxels using a neighbourhood likelihood model and by recursively growing our observation space, $\mathcal{N}_i^k$, defined as

$$\mathcal{N}_i^k = \bigcup_{j \in \mathcal{N}_i^{k-1}} \mathcal{N}_j, \tag{2}$$

where $\mathcal{N}_i^0$ is just the voxel in question (as in (1)), and $\mathcal{N}_i$ is a the first-order neighbourhood. We can express the posterior probability of a class tissue, $C_i^{(t)}$ for a k^{th} inference problem as

$$p_i^k = p(C_i^{(t)}|I_{\mathcal{N}_i^k}^{(r)}, D_{\mathcal{N}_i^k}^{(t)}) = \sum_{\{C_{\mathcal{N}_i}^{(t)}\}} p(C_i^{(t)}, \{C_{\mathcal{N}_i}^{(t)}\}|I_{\mathcal{N}_i^k}^{(r)}, D_{\mathcal{N}_i^k}^{(t)}) \tag{3}$$

$$= \sum_{\{C_{\mathcal{N}_i}^{(t)}\}} p(\{C_{\mathcal{N}_i}^{(t)}\}|I_{\mathcal{N}_i^k}^{(r)}, D_{\mathcal{N}_i^k}^{(t)}) p(C_i^{(t)}|\{C_{\mathcal{N}_i}^{(t)}\}, I_i^{(r)}, D_i^{(t)})$$

$$= \sum_{\{C_{\mathcal{N}_i}^{(t)}\}} \frac{P^k_{\{C_{\mathcal{N}_i}^{(t)}\}}}{K^k} \sum_{C_i^{(r)}} p(D_i^{(t)}|C_i^{(r)}, C_i^{(t)}, I_i^{(r)}) p(\{C_{\mathcal{N}_i}^{(t)}\}|C_i^{(t)}) p(C_i^{(t)}|C_i^{(r)}) p(C_i^{(r)}|I_i^{(r)}), \tag{4}$$

where K^k is a normalization constant, $P^k_{\{C_{\mathcal{N}_i}^{(t)}\}}$ is the probability of a tissue class label configuration, $\{C_{\mathcal{N}_i}^{(t)}\}$ is a configuration of tissue class labels in $\mathcal{N}_i$ (see example in Fig. 2b) and where the summation implies that we consider all possible configurations. We assume Markovianity ($p(C_i^{(t)}|C_{j \notin \mathcal{N}_i}, C_{\mathcal{N}_i}^{(t)}) = p(C_i^{(t)}|C_{\mathcal{N}_i}^{(t)})$), causality ($p(C_i^{(r)}|C_{\mathcal{N}_i}^{(t)}) = p(C_i^{(r)})$), conditional independence of $D_i^{(t)}$ from $C_{\mathcal{N}_i}^{(t)}$ given $C_i^{(t)}$, and conditional independence of $C_{\mathcal{N}_i}^{(t)}$ from $C_i^{(r)}$ given $C_i^{(t)}$. We can expand $P_{\{C_{\mathcal{N}_i}^{(t)}\}}$ as follows:

$$P^k_{\{C_{\mathcal{N}_i}^{(t)}\}} = p(\{C_{\mathcal{N}_i}^{(t)}\}|I_{\mathcal{N}_i^k}^{(r)}, D_{\mathcal{N}_i^k}^{(t)}) = \prod_{j \in \mathcal{N}_i} p(C_j^{(t)}|I_{\mathcal{N}_j^{k-1}}^{(r)}, D_{\mathcal{N}_j^{k-1}}^{(t)}) = \prod_{j \in \mathcal{N}_i} p_j^{k-1}. \tag{5}$$

This iterative process can be seen as modelling tissue label dependencies locally while recursively growing the observation space, which, in the limit, would consider the entire image.

Likelihood Models

Difference likelihood models are learned from training data for all combinations of $C^{(r)}$ and $C^{(t)}$, where tissue classes are restricted to one of 5 classes (csf, gm, wm, les, pv). Transitions from $C^{(r)}$ to $C^{(t)}$ may represent real change (e.g. wm-les), misregistration (e.g. wm-gm), differences in partial volume effects at different timepoints, variability or error in segmentation of the training data, or some combination of these factors. To reduce the number of models that need to be learned, $D^{(t)}$ is assumed to be conditionally independent of $I^{(r)}$ given $C^{(r)}$, except for the case of $C^{(r)}$=les. For this special case, we use PCA to project our multimodal intensities onto a 1-D subspace that best captures intensity variations seen in lesions, and separate this subspace into 3 distinct lesion intensity classes. Difference likelihood models for wm-wm, gm-gm, csf-csf and pv-pv transitions are modeled as 3D Gaussians as they are well approximated as such. All other models are represented by 3D non-parametric distributions using Parzen windows [11].

The neighbourhood likelihood represents the likelihood of observing a neighbourhood configuration $C^{(t)}_{\mathcal{N}_i}$ around a voxel with label $C^{(t)}_i$. A 4-voxel in-plane neighbourhood was used, and a neighbourhood configuration was represented by a count of each tissue class in the 4-voxel neighbourhood. Models were constructed as histograms by observing the frequencies of the different neighbourhood representations that occured in training data for each tissue class.

The transition likelihood represents the prior probability of transitioning from one tissue label to any other (or the same) tissue label. This acts as a bias toward the tissue class at the reference timepoint and towards more plausible tissue label transitions. Transition likelihoods are learned based on frequency of occurence in the training data.

The inclusion of a prior term, $p(C^{(r)}|I^{(r)})$, implies that we have available some form of probabilistic tissue classification for the reference timepoint. This prior on $C^{(r)}$ can come from some other automated tissue segmentation scheme, from manual labeling, or from some combination of both.

3 Experiments

3.1 Preprocessing

All MRI data used in this study consists of sets of T1-weighted (T1), T2-weighted (T2) and PD-weighted (PD) images at a resolution of 1x1x3mm. Each scan was corrected for intensity non-uniformity [12], masked to exclude non-brain and the posterior fossa, linearly (6 DOF) registered to the T2 image at the baseline scan from the same patient, and intensity normalized to a common intensity space [13]. The "ground truth" tissue labels were generated from an automatic tissue segmentation using an in-house classifier based on [14] which then had voxels classified as lesion manually verified and corrected by experts. Tissue class labels were restricted to one of csf, gm, wm, les, and pv. These manually corrected (MC) 5-tissue class segmentations served as a reference for subsequent training and validation.

3.2 Scan-Rescan

A scan-rescan data set consisting of 3 relapsing-remitting MS (RRMS) patients allowed us to validate the precision of the proposed method in the absence of real physical change. Patients were scanned on a Siemens Sonata 1.5T scanner, removed from the scanner, and then rescanned. For the scan-rescan data set, fully manual lesion labels (FM) were also available in addition to the manually corrected lesion labels (MC). Lesions were required to consist of at least two contiguous voxels. The MC labels for the reference timepoint were used as a prior for the Bayesian classifier (BC). Classification was done by first using the scan as the reference timepoint (BC-S) and classifying the rescan, and secondly using the rescan as the reference (BC-R) and classifying the scan. Models used for classification were learned from independent training data. We define new lesion voxels as those that were not labelled as lesion in the reference scan but labelled as lesion in the follow-up scan, and resolved lesions those that were labelled as lesion at reference but not in the follow-up. Means and standard deviations of lesion volume at reference, new lesion voxels, resolved lesion voxels and change in lesion volume over the 3 scan-rescan patients are shown in Table 1. Given that there is no biological change in the scan-rescan period, ideally no lesion activity would be detected. The number of new and resolving lesion voxels are greatly reduced when using the proposed Bayesian classifier as compared to both the FM and MC labels, suggesting greater precision with the proposed method.

Table 1. Scan-Rescan precision for 3 RRMS patients

	MC	FM	BC-S	BC-R
Lesion Volume at Reference (voxels)	7466±4278	7517±4098	7466±4278	7466±4404
New Lesion Voxels	1368±853	1657±887	74±8	26±9
Resolved Lesion Voxels	1313±700	1453±906	49±21	28±15
Net Change in Lesion Voxels	55±155	204±321	25±27	-3±20

3.3 Clinical Data

Increased precision is only meaningful if the classifier is still sensitive to true change. A clinical data set was used to validate the sensitivity of the proposed method to new and enlarging lesions. This data set consists of 212 total scans from 89 RRMS patients with 2-4 longitudinal scans each, taken over a period of 48 weeks with a minimum interval of 12 weeks between scans. Fully manual lesion labels were not available for this data set, so MC labels were used as reference for all 212 scans. Meaningful evaluation based on comparison to reference lesion labels is challenging, due to lack of consensus as to what consitutes a lesion, ambiguity of lesion boundaries, and lack of precision in labelling of the same patient over time. A subset of the new lesion voxels from the MC labels were identified as being new lesions or enlarging portions of existing lesions, based on a minimum of 3 contiguous new lesion voxels and spatial properties

Table 2. Apparent Sensitivity to New and Enlarging Lesions as compared to MC Labels

NE Size	ALL	>=5 voxels	>=10 voxels
Total # NE	63	58	45
Criteria(a)	46 (73%)	45 (78%)	37 (82%)
Criteria(b)	53 (84%)	52 (90%)	44 (98%)
Voxel-wise sensitivity	76.6%	76.9%	77.5%

of connectedness to existing lesions. This set of new or enlarging (NE) lesion labels was manually verified by experts and ensured as much as possible that our ground truth definition of NE lesions corresponds to real change in brain tissue. For each timepoint other than the baseline scan, the scan and MC tissue labels from the previous timepoint were used as the reference image and prior. In this way, all scans except for the baseline scan were classified in a pairwise fashion. Four-fold cross-validation was used, with 66 or 67 patients used for training our models, and 22 or 23 used for testing, on each fold. Performance of our classifier was measured based on the number of NE lesions that were detected. Two different criteria were used to decide whether an NE lesion was considered as detected : (a) identification of a minimum of 50% of voxels in an NE lesion and (b) identification of 3 or more voxels in an NE. Analysis was done separately for all NE lesions and subsets of NE lesions that were greater than 5 and 10 voxels in size. A voxel-wise sensitivity to NE lesions was also measured, which is defined as the percentage of all voxels in new and enlarging lesions that were classified as lesion.

Specificity of newly detected lesions was not quantitatively evaluated as we did not have a filtered subset of MC labels that allowed for a meaningful comparison. Qualitative analysis showed that for a small subset of scans, significant false detection of new lesions occured adjacent to the lateral ventricles. Distortion, atrophy and partial volume effects in the z-direction all contributed to these false detections. Sensitivity and specificity of resolving lesions were also not explicitly measured due to lack of suitable ground truth. Qualitative analysis showed good sensitivity to fully resolving lesions, but resolving portions of partially resolving lesions were generally underestimated, due to the attractive effect of remaining lesion in the neighbourhood likelihood model. Very few falsely resolving lesions were observed.

Table 3. Number of New And Resolving Lesion Voxels for Slices in Figure 3

	Slice 1		Slice 2		Slice 3	
	MC	BC	MC	BC	MC	BC
Lesion Voxels At Reference	156	156	181	181	347	347
New Lesion Voxels	58	57	100	42	137	2
Resolved Lesion Voxels	77	30	77	41	108	3
Net Change in Lesion Voxels	-19	27	33	1	21	-1

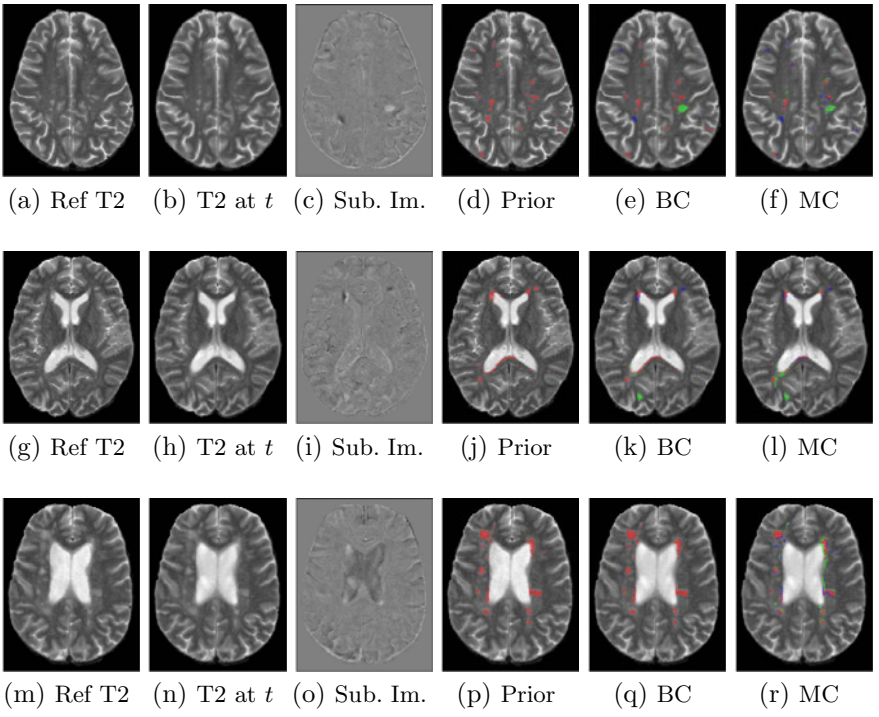


Fig. 3. Sample classification results for three slices from three different patients, where Slice 1 (a-f) and Slice 2 (g-l) both have new and resolving lesions and Slice 3 (m-r) exhibits little or no real change. Reference T2 images (Ref. T2), T2 images at the timepoint t to be classified (T2 at t), subtraction images between the two timepoints (Sub. Im.), prior classifications at the reference timepoints (Prior), output of the proposed Bayesian classifier for time t (BC), and MC labels for time t (MC), are shown for each slice. The BC and MC labels are colour-coded as follows: stable lesion voxels are shown in red, new lesion voxels are shown in green, and voxels that were lesion at the reference timepoint but have resolved are shown in blue.

Sample classification results for 3 slices of 3 different patients in the clinical data set are shown in Figure 3, and the lesion activity in those slices as detected by the proposed classifier and the MC labels is summarized in Table 3. New and resolving lesions are correctly identified while stable lesions are labelled in a much more consistent manner than for the MC labels, where lesion boundaries are shown to fluctuate, and more ambiguous tissue intensities may be labeled differently at the two timepoints despite lack of apparent change.

4 Discussion and Future Work

In this paper, we introduce an automatic Bayesian classifier that detects MS lesion activity in longitudinal scans based on subtraction images. Our approach

attempts to overcome the limitations of subtraction images in terms of registration error and noise by embedding a prior classification at a reference timepoint and by building likelihood models that account for artifact and noise. Our approach was evaluated on both a scan-rescan data set and a large multicenter clinical data set and has demonstrated increased precision as compared to a manual classification, while remaining sensitive to lesion activity.

A quantitative evaluation of sensitivity to resolving lesions and specificity of both new and resolving lesions is needed to fully characterize the performance of the proposed classifier. The incorporation of non-linear registration or explicit segmentation of lateral ventricles may aid in reducing false detection of new lesions in patients where there is significant atrophy or distortion. The preprocessing pipeline used was chosen based on convenience. More optimal pipelines specific to longitudinal data may help reduce noise and artifact in subtraction images [9,15].

References

1. Lee, M.A., Smith, S., et al.: Defining multiple sclerosis disease activity using MRI T2-weighted difference imaging. Brain 121, 2095–2102 (1998)
2. Tan, I.L., van Schijndel, R.A., et al.: Image Registration and subtraction to detect active T_2 lesions in MS: an interobserver study. J. Neurol. 249, 767–773 (2002)
3. Moraal, B., Meier, D.S., et al.: Subtraction MR Images in a Multiple Sclerosis Multicenter Clinical Trial Setting. Radiology 250, 506–514 (2009)
4. Duan, Y., Hildenbrand, P.G., et al.: Segmentation of Subtraction Images for the Measurement of Lesion Change in Multiple Sclerosis. Am. J. Neuroradiol. 29, 340–346 (2008)
5. Rey, D., Subsol, G., et al.: Automatic detection and segmentation of evolving processes in 3D medical images: Application to multiple sclerosis. Med. Image Anal. 6, 163–179 (2002)
6. Welti, D., Gerig, G., et al.: Spatio-temporal Segmentation of Active Multiple Sclerosis Lesions in Serial MRI Data. In: Insana, M.F., Leahy, R.M. (eds.) IPMI 2001. LNCS, vol. 2082, p. 438. Springer, Heidelberg (2001)
7. Prima, S., Arnold, D.L., et al.: Multivariate Statistics for Detection of MS Activity in Serial Multimodal MR Images. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2878, pp. 663–670. Springer, Heidelberg (2003)
8. Aït-Ali, L.S., Prima, S., et al.: STREM: A Robust Multidimensional Parametric Method to Segment MS Lesions in MRI. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 409–416. Springer, Heidelberg (2005)
9. Bosc, M., Heitz, F., et al.: Automatic change detection in mutimodal serial MRI: application to multiple sclerosis lesion evolution. NeuroImage 20, 643–656 (2003)
10. Thirion, J.-P., Calmon, G.: Deformation Analysis to Detect and Quantify Active Lesions in Three-Dimensional Medical Image Sequences. TMI 18, 429–441 (1999)
11. Turlach, B.: Bandwidth selection in kernel density estimation: a review. Discussion paper 9317, Institut de Statistique, UCL, Louvain la Neuve, Belgium (1993)
12. Sled, J.G., Zijdenbos, et. al.: A non-parametric method for automatic correction of intensity nonuniformity in MRI data. TMI 17, 87–97 (1998)
13. Nyùl, L.G., Udupa, J.K., et al.: New variants of a method of MRI scale standardization. TMI 19, 143–150 (2000)
14. Francis, S.: Automatic lesion identification in MRI of MS patients. Master's Thesis, McGill University (2004)
15. Meier, D.S., Guttman, R.G.: Time-series analysis of MRI intensity patterns in multiple sclerosis. NeuroImage 20, 1193–1209 (2003)