A robust Expectation-Maximization algorithm for Multiple Sclerosis lesion segmentation
Daniel García-Lorenzo, Sylvain Prima, Sean Patrick Morrissey, Christian Barillot

To cite this version:

HAL Id: inserm-00421713
https://www.hal.inserm.fr/inserm-00421713
Submitted on 2 Oct 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A robust Expectation-Maximization algorithm for Multiple Sclerosis lesion segmentation

Daniel García-Lorenzo\textsuperscript{1,2,3}, Sylvain Prima\textsuperscript{1,2,3}, Sean P. Morrissey\textsuperscript{1,2,3,4} and Christian Barillot\textsuperscript{1,2,3}

\textsuperscript{1}INRIA, VISAGES U746 Unit/Project, Rennes, France
\textsuperscript{2}University of Rennes I, CNRS IRISA, Rennes, France
\textsuperscript{3}INSERM, VISAGES U746 Unit/Project, Rennes, France
\textsuperscript{4}Neurology Department, University Hospital Pontchaillou, Rennes, France

August 19, 2008

Abstract

A fully automatic workflow for Multiple Sclerosis (MS) lesion segmentation is described. Fully automatic means that no user interaction is performed in any of the steps and that all parameters are fixed for all the images processed in beforehand. Our workflow is composed of three steps: intensity inhomogeneity (IIH) correction, skull-stripping and MS lesions segmentation. A validation comparing our results with two experts is done on MS MRI datasets of 24 MS patients from two different sites.

1 Introduction

Magnetic Resonance Imaging (MRI) has been used as a biomarker for Multiple Sclerosis over the last 25 years. MRI has a high sensitivity to detect white matter lesions (WML) in MS patients. In cross-sectional and longitudinal studies, manual or semi-automatic segmentation have been used to compute the total lesion load (TLL) in T2-w, PD-w or T1-w (either unenhanced or gd-enhanced) MR sequences but with the drawback that these methods are very time consuming and have large intra- and inter-operator variability [9]. Automatic methods show great promise to reduce these variabilities and improve different issues of the analysis of large multi-center study results.

The purpose of this paper is to describe our automatic segmentation workflow, based on a previous segmentation algorithm already published [1]. This paper is structured as follows. In Sections 2 the data employed in the evaluation is described. Then in Section 3, we present each step of the workflow focusing our description in the MS lesion segmentation algorithm. Finally in Section 4 we describe the results and we present our discussion in Section 5.

2 The data

A total of 20 MRI datasets of MS patients were available, in whom we had access to the results of manual MS lesion segmentation (training datasets), and in addition a further 24 subjects in whom we had no access...
to results of manual MS lesion segmentation (evaluation datasets). All datasets include five different MR images for each subject: T1-w, T2-w, FLAIR, Mean Diffusivity (MD), and Fractional Anisotropy (FA). Acquisitions were performed in two different hospitals: Children’s Hospital Boston (CHB) and University of North Carolina (UNC). UNC MR datasets have a slice thickness of 1mm and in-plane resolution of 0.5 mm and CHB datasets have slice thickness of 1.5 mm and in-plane resolution of 0.5 mm. These initial data have been rigidly registered to a common space and up-sampled to an isotropic resolution of 0.5 mm^3 using a B-spline interpolation, we had no access to the original MRI data. A neuroradiologist from each center performed a manual segmentation of the MS lesions in the images. On visual inspection of the manual lesion segmentation from the two sites there is evidence of high inter rater variability.

3 The workflow

In our workflow, we only use T1-w, T2-w and FLAIR sequences. We start with intensity image denoising, inhomogeneity correction and skull stripping before performing the actual automatic MS lesion Segmentation. Each of those image preprocessing steps and the MS lesion segmentation algorithm itself are described in the following.

3.1 Denoising

Image noise corrupts the image intensities and decreases the efficiency of segmentation algorithms. Noise is usually due to several factors thus as the MR hardware or the MR sequence. Several methods have been developed and are widely applied in the literature to denoise images. However, a potential drawback of the denoising algorithms is the smoothing of small lesions and the reduction of contrast with neighboring normal appearing WM.

One of the assumptions that are made in many of the denoising algorithms is the spatial independence of the noise. In the case of the data described in Section 2, the use of an interpolation method is creating some spatial relationship among neighboring voxels, so the assumptions of most of the denoising methods are no longer valid. Therefore, we do not apply any denoising of the datasets which were available for this study. Ideally, denoising is performed on the raw data as the very first step and thus the spatial independence hypothesis remains valid.

3.2 Intensity Inhomogeneity Correction

Intensity non-uniformity in MR images is due to a number of causes during the acquisition of the MR data. In principle, they are due to MR devices, such as B0- or B1-field non-uniformity, and relate to artifacts caused by slow, non-anatomic intensity variations within the same tissue over an image domain.

An entropy-based algorithm for intensity inhomogeneity correction [10] is employed to correct the spatial variations of intensity in the same tissue. Entropy-based methods do not make any assumption of the sequences type or tissue intensity, and therefore they can be applied to all kind of image sequences. In our case, we performed IIH correction only on T1-w and FLAIR images as it was shown experimentally that T2-w images have less inhomogeneity, and even more importantly that IIH methods could potentially degrade the quality of T2-w images [8].
3.3 Skull Stripping

Skull stripping methods remove non-brain voxels from the image to simplify the following lesion segmentation. There are multiple methods described in the literature and several comparison studies were performed [12, 6]. The skull stripping method from the FSL library is employed, called bet [13], used as previously described by Rex et al. [12] to improve its results.

This first algorithm employed is not perfect and usually leaves some non-brain voxels, mainly skull, optic nerve or veins. To improve the brain mask by removing the skull and the veins, we include information coming from T2-w and FLAIR images where veins and skull have low intensity. We use a 3-class model (in T2-w: CSF, brain tissues and skull/veins; and in FLAIR: CSF/skull/veins, brain tissues and lesions) in each sequence with a Expectation Maximization (EM) algorithm [4] to classify the voxels inside the first brain mask. Then only the voxels that have been classified in both sequences as the class with skull and veins are removed from the brain mask. An example can is shown in Figure 1.

3.4 MS Lesions Segmentation

Our original method is called STREM (Spatio Temporal Robust Expectation Maximization)[1], and after introducing improvements to reduce the number of false positives, but also to apply STREM in datasets where only single time point MRI’s are available. MS lesion segmentation is performed with a three-step

Estimation of NABT parameters

NABT image intensities are modeled with a 3-class finite multivariate Gaussian mixture [16], where each class is associated to a different part of the brain: White Matter (WM), Grey Matter (GM) and CSF. All the MR sequences are used to create a multidimensional feature space in order to benefit from the specific inherent information of each sequence. To calculate the NABT parameters we use a modified Expectation Maximization algorithm, called $mEM$, based on the Trimmed Likelihood (TL) Estimator [11]. It was shown to have a monotonous convergence, at least to a local maximum of TL, as the original Expectation Maximization (EM) algorithm. The idea is to use exclusively in our computation of TL the $n-h$ voxels that are closer to the model and reject the $h$ voxels more likely to be outliers.

$$TL = \sum_{i=1}^{n-h} f(x_{v(i)}; \Theta)$$

Where $n$ is the total number of voxels, $h$ the number of rejected voxels, $x_i$ is a vector with the intensities of the $m$ sequences of the voxel $i$, $\Theta$ the parameters of our 3-class model, $f(.)$ the p.d.f. of the model and $\nu(.)$ is a permutation function which orders voxels so that:

$$f(x_{v(1)}; \Theta) \geq f(x_{v(2)}; \Theta) \geq ... \geq f(x_{v(n)}; \Theta)$$

The trimming parameter $h$ is chosen arbitrarily with a high value, to ensure the rejection of all WML voxels from the computation of the NABT parameters. In our workflow the parameter $h$ is set to the 10% of the pixels of the brain.

Refinement of outliers detection

In practice, the $h$ rejected points actually contain some inliers that actually fit the NABT model reasonably well. Thus, to refine the outliers detection, we compute the Mahalanobis distance between each of the $n$ voxels in the image and each NABT given the previously computed parameters. Considering that voxels intensities in each NABT follows a Gaussian law, these Mahalanobis distances follow a $\chi^2$ law with $m$ d.o.f [1, 5]. Each voxel in the image is defined as an outlier if the Mahalanobis distance for every class is greater than the threshold defined by the $\chi^2$ law, for a given p-value. In our workflow p-value is set to 0.4 to ensure all the lesion voxels will be taken into account although this means the existence of many false positives at this stage.

Application of lesion rules

Outliers found with the Mahalanobis distance may be originated from other tissue compartments than WML, basically due to partial volumes, vessels, registration errors, noise, etc. In order to discriminate between the WML and false positives, rules are defined with neurologists and neuroradiologists based on image intensities from the respective MR sequences and voxel connectivity.
Different intensity rules can be implemented for the different types of MS lesions [1]: black holes, Gadolinium-enhanced lesions and T2-w lesions. In this paper we focus in T2-w lesions that are, compared to the normal appearing WM, hyperintense in T2-w and FLAIR, and isointense or hypointense (e.g. black holes) in T1-w. Hyperintense and hypointense voxels are defined by $3.0 \times \sigma_{WM} \pm \mu_{WM}$, where $\sigma_{WM}$ and $\mu_{WM}$ are the standard deviation and the mean of the white matter respectively.

Voxel connectivity allows the use of neighboring rules instead of classifying each voxel independently. In this case, a minimal size of MS lesion is defined [2], so detected lesions that have a size smaller than 3 mm$^3$ are discarded. We also remove detected lesions that are contiguous to brain border or not contiguous to WM tissue.

4 Results

Our workflow does not use any learning steps. Training datasets where not necessary for the final processing of the test images. Yet, we processed these datasets in order to verify that our segmentation workflow could handle these images. In those training datasets no numerical evaluation or optimization of parameters were performed.

4.1 Evaluation measures

Four different measures have been employed in the comparison of the automatic MS lesion segmentation with the expert manual segmentation. A normalization into a 0-100 range was performed [14].

- Volume Diff.: The volume difference captures the absolute percent volume difference to the expert rater segmentation.
- Avg. Dist.: The average distance captures the symmetric average surface distance to the expert rater segmentation.
- True Pos.: Number of lesions in the automatic lesion segmentation that overlaps with a lesion in the expert segmentation divided by the number of overall lesions in the expert segmentation.
- False Pos.: Number of lesions in the automatic lesion segmentation that do not overlap with any lesion in the manual segmentation divided by the number of overall lesions in the automatic segmentation.

In addition a STAPLE algorithm [15] has been performed with the two expert segmentation and other automatic segmentation methods to compare their solutions.

4.2 Test images

The described workflow does not include any manual or semi automatic steps. All the images were processed automatically with the same parameters. Table 4 shows the results for the test images. Average results show that we are far from the value of 90 associated with the inter rater variability in the normalized scale [14], but our results are independent of the center where MR acquisition was performed. An example of good lesion segmentation is given in Figure 2.

There are five datasets with scores under 70. CHB_test1_Case15 was not processed because of a human error and UNC_test1_Case07 T1-w is likely to have been done after Gadolinium injection. In
Figure 2: Dataset CHB test1 Case08: From left to right T1-w, T2-w, FLAIR and automatic segmentation results

Figure 3: Some artifacts (from left to right): CHB test1 06 T2-w and Flair, UNC test1 Case07 T1-w (we think Gadolinium-enhancing T1-w)

UNC_test1_Case10, the low contrast between lesions in the FLAIR image could cause problems in the segmentation. We found that CHB FLAIR images usually have a drop in signal intensity in the superior part of the brain, see Figure 3, that the IIH correction method was not able to correct. In the case of CHB_test1_Case06 and _Case12 this intensity drop was large enough to alter our 3-class model giving low segmentation results. In those two datasets we also find strong movement artifacts but, as our method makes little use of the spatial information, these artifacts should be less critical in the performance as observed in the results of UNC_test1_Case02.

5 Discussion

We have presented our fully automatic workflow for the segmentation of MS lesions. Our objective is to propose a method which is not based on training steps and can be used with different MR protocols or scanners yielding reproducible results. Looking at the available datasets for this validation study we strongly felt that the analysis of the results is not straight forward and there are several aspects that need to be discussed and improved for new future validations studies: definition of MRI lesion, quality of datasets and total lesion load of the patients.
5.1 Definition of MS lesion in MRI

Figure 4: Results for test images. For each measure the value is given in two different metric: real value in the left and a normalized value in the right.

5.1 Definition of MS lesion in MRI

Compared to the segmentation validation of other targets, MS lesion segmentation have an increased complexity. Today there is no consensus on a precise definition of MS lesion in MRI. This situation leads to a very high variability, in first place, in the detection. In second place, and once all the lesions have been detected, there is a still a high variability in contour of each lesion [9, 17, 7], as MS lesions not rarely lack a sharp border, but also because MS lesion size and border vary according to the MR sequence used [7].

5.2 Quality of datasets

Quality of MR acquisition is difficult to evaluate but image artifacts complicate further image analysis and can decrease the detectability of the lesions. For example, UNC datasets were easier to segment as they have a very high variability, in first place, in the detection. In second place, and once all the lesions have been detected, there is a still a high variability in contour of each lesion [9, 17, 7], as MS lesions not rarely lack a sharp border, but also because MS lesion size and border vary according to the MR sequence used [7].

5.3 Total lesion load of the MS patients

Accuracy of segmentation methods vary with the magnitude of lesion load or degree of atrophy, and should be taken into account when interpreting our results. In addition, the measures in the validation are very dependent on the total number of lesions: for example when there are only a few lesions missing one single
lesion will significantly drop True Positives while when there are many lesions True Positives will only decrease very little.

References


