The impact of processing workflow in performance of automatic white matter lesion segmentation in Multiple Sclerosis

Daniel García-Lorenzo, Sylvain Prima, Laura Parkes, Jean-Christophe Ferré, Sean Patrick Morrissey, Christian Barillot

The impact of processing workflow in performance of automatic white matter lesion segmentation in Multiple Sclerosis

Daniel García-Lorenzo[1,2,3] ⋆ , Sylvain Prima[1,2,3], Laura Parkes[4], Jean-Christophe Ferré[5], Sean Patrick Morrissey[1,2,3,6] and Christian Barillot[1,2,3]

[1] INRIA, VisAGeS Unit/Project, IRISA, Rennes, France
[2] University of Rennes I, CNRS IRISA, Rennes, France
[3] INSERM, U746 Unit/Project, IRISA, Rennes, France
[4] Magnetic Resonance and Image Analysis Research Centre, University of Liverpool, UK
[5] Department of Neuroradiology, University hospital Pontchaillou, Rennes, France
[6] Department of Neurology, University hospital Pontchaillou, Rennes, France

Abstract. The design of a robust automatic segmentation workflow is crucial to deal with the shortcomings of images that impacts on their analysis. In this paper, different workflows, using state-of-the-art tools, are compared in order to evaluate the role of the different preprocessing tasks. We propose some methods in order to improve the computing time, robustness and accuracy of the segmentation method. We compare with manual segmentation as ground truth the workflows and improvements in the segmentation method. Finally, we present a new automatic workflow for white matter lesions segmentation in Multiple Sclerosis.

1 Introduction

Multiple Sclerosis (MS) is a chronic inflammatory-demyelinating disease of the central nervous system. Magnetic Resonance Imaging (MRI) serves as a biomarker that detects with high sensitivity white matter lesions (WML) in patients with MS. Over the last 25 years, it has been increasingly used for diagnosis, prognosis and as a surrogate marker in MS trials. Conventional Magnetic Resonance (MR) sequences for MS include pre- and post-gadolinium (gd) T1-weighted (T1-w), T2-weighted (T2-w), proton density (PD) or FLuid Attenuating Inversion Recovery (FLAIR). These sequences have been developed to optimize the detection of the lesions in the white matter (WM) [1].

In cross-sectional and longitudinal studies, manual segmentation has been used to compute the total lesion load in T2-w, PD-w, unenhanced and gd-enhanced T1-w MR sequences but this method is very time consuming and has large intra- and inter-operator variability [2]. Semi-automatic methods tend to reduce this variability, but there is a great promise that automatic methods

⋆ We are thankful to ARSEP for funding

will improve considerably lesion segmentation reproducibility, which is of critical importance when processing huge amounts of MR images, as in large multicenter clinical trials.

Automatic segmentation of WML is a complex task that requires considerable preprocessing of MRI data [3], but so far no standards have been defined for preprocessing. Our assumption is that WML segmentation methods significantly depend on the preprocessing tasks, most notably: registration, skull stripping, image denoising and intensity inhomogeneity correction. In this paper, the segmentation of WML is performed with STREM [4]. It is a parametric multidimensional and multitemporal approach for the segmentation of WML and Normal Appearing Brain Tissues (NABT).

The purpose of this paper is to investigate how the different preprocessing procedures impact on the segmentation results, comparing manual segmentation of WML performed by an expert as the ground truth. Another outcome of this work is to propose improvements of the WML segmentation method in terms of computation time, robustness and accuracy.

This paper is structured as follows. In Sections 2.1 and 2.2, we briefly describe the STREM algorithm, its original version and the modifications we implement to shorten its computing time and to increase its robustness and accuracy. Then in Section 2.3, we present each of the typical preprocessing tasks. In Section 3 we describe the different data, the different workflows we used in our experiments and the results. Then we present conclusions of this work in Section 4.

2 Methods

2.1 The original segmentation algorithm: STREMv1 (version 1)

A 3-class Finite Multivariate Gaussian Mixture Model [5] is used to model NABT intensities. All the MR sequences are used to create a multidimensional feature space in order to benefit from the specific inherent information of each sequence. The main idea of this algorithm is to find the atypical intensities of the WML voxels not as a specific class but as outliers to the model. This is performed through a three-step process: 1. Robust estimation of NABT parameters, 2. Refinement of outliers detection and 3. Application of lesion rules

Estimation of NABT parameters To calculate the NABT parameters we use a modified Expectation Maximization algorithm, called mEM, based on the Trimmed Likelihood (TL) Estimator [**?**]. 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.

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

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, Θ 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:

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

Once the algorithm arrives to a maximum of TL, each one of the three classes is assignated to its corresponding tissue using the mean value of the Gaussian in the T1-w sequece, thus the class with lower mean in is CSF, the next one GM and the last one WM. In order to initialize the EM algorithm, an atlas is registered to the subject and the probability maps for the three tissues associated to the atlas are used to obtain the initial mean and variance for each class [6].

Refinement of outliers detection 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 practice, these 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 χ^2 law with m d.o.f [4, 7]. 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 χ^2 law, for a given p-value.

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 [4].

2.2 Some improvements: STREMv2 (version 2)

In StremV2 we perform two major changes of the algorithm. In order to speed up the process and to reduce the probability to reach a local maximum of TL, the initialization of mEM is modified. Then, new rules are introduced to take into account simple spatial constraints to reduce the number of false positives.

Improvement of the initialization of the mEM algorithm As mentioned earlier, the mEM algorithm, as the original EM, may converge to a local maximum of TL. That is why a proper initialization is critical to find the global maximum. In MRI tissue segmentation, a typical approach is to use an atlas to initialize the EM algorithm [6], but the atlas registration may be time consuming and not very appropriate in MS patients displaying important tissue atrophy or a huge lesion load. Another option is to perform multiple random initializations, let the algorithm converge and keep the solution which maximizes the TL of the image [8]. The drawback of this approach is that it is computationally expensive as multiple mEM must be performed to guarantee the global convergence. There exists no general agreement on how many mEM must be performed but as the number of parameters is a $O\left(m^2\right)$ the number of possible local maxima also increases, and so does the computational complexity.

To guarantee reaching to the global maximum, we choose to implement multiple initializations. In order to reduce the computation time, two shortcuts are implemented. T1-w is known to have the best contrast between NABT, so we decide to perform a two-step procedure. First, the classification of T1-w images is performed with a mEM with multiple random initializations, then the a posteriori probabilities obtained by this initial mEM (at the end of the E-step) are used to initialize a multidimensional mEM with all the MR sequences (applying the multidimensional M-step with the output of the T1-w E-step).

The second shortcut is to use a selection strategy when initializing multiple mEM. Instead of waiting for the convergence of each mEM, selections can be done after a fixed number of iterations. This strategy works as follows:

i n_0 random initializations
ii Use mEM with $iter_0$ iterations using these initializations
iii Keep n_1 best solutions
iv Use mEM with n_1 partial solutions until convergence
v Keep the best solution

In practice, we use $n_0 = 300$, $iter_0 = 10$ and $n_1 = 10$, as results where similar to higher values for the three parameters. After modifying the initialization of the mEM algorithm, we perform the same steps 2 and 3 of STREMv1 and we call the new algorithm STREMv1.5 (version 1.5).

Spatial constraints One of the assumptions of STREMv1 is that WML can be segmented only based on MR intensities, without any spatial constraints. This hypothesis is only partially satisfactory because, in practice, this leads to the detection of a number of false positives:

1. In brain WM, isolated hyperintensity voxels are misclassified due to image noise.
2. In the cortex and CSF, voxels can have the same MR intensity as WML.

To further improve WML segmentation, we apply spatial constraints to the output of STREMv1.5 and we will call this algorithm STREMv2. To avoid the first type of false positives, a minimal size of WML is defined [9]. With morphological operations we eliminate all WML that have a size smaller than 3 mm^3. For the second type of errors, a rule is applied to eliminate all WML that are not contiguous to the WM.

2.3 Preprocessing stage

Intensity inhomogeneity (IIH) correction Several factors cause IIH in MRI, for example the inhomogeneity of the magnetic B0 field. Usually this small spatial variation of intensity does not significantly affect conventional radiological assessments of the images but reduces the performance of many image processing algorithms. To correct for this IIH, the image intensity profile is modified using B-spline functions which minimize the entropy of the image histogram [10][2].

[2] available at: http://brainvisa.info/index.html

Denoising Image additive noise due to hardware acquisition and thermal noise may also impact the processing performances, and thus have to be removed. Several edge preserving denoising techniques have been recently developed and we choose the Non-Local Means (NLM) algorithm which has shown better performance compared to other state-of-the-art methods [11][1].

Registration In order to process multiple sequences, images have to be aligned in the same reference frame. One of the images has to be chosen as reference and the other images are registered to this reference. The optimal rigid body transformation which maximizes the mutual information is calculated using the NEWUOA optimizer that has been shown to be faster and more robust than traditional minimization techniques [12][1]. In this study FLAIR sequence is used as reference and registration is applied after IIH correction and denoising if they are employed.

Skull Stripping As in most brain segmentation techniques, a preprocessing task has to be performed in order to extract the brain from non-brain tissues. A deformable model is applied in the T1-w images for this task [13][3].

It has been largely studied that skull-stripping can cause very unstable effects [14]. To minimize this effect we choose always to correct the automatic skull-stripping results in order to focus the validation only in the IIH and denoising steps.

3 Evaluation & Results

In order to evaluate the new tissue segmentation method in the context of the workflow, we have retained three types of data acquisition protocols and five combinations of workflows.

3.1 Data

Three different MR protocols are used in the experiments:

subject1 Images acquired on a 3T Siemens TRIO: 3D 1mm isotropic T1-w, 2D 3-mm axial slice thickness Dual Echo (T2-w and PD) and 2D 3-mm axial slice thickness FLAIR.

subject2 Images acquired on a 3T Philips ACHIEVA: 3D 1mm isotropic T1-w, 2D 3-mm axial slice thickness Dual Echo (T2-w and PD) and 2D 3-mm axial slice thickness FLAIR.

subject3 Images acquired on a 3T Siemens TRIO: 3D 1mm isotropic T1-w, 3D 1mm isotropic T2-w and 3D 1mm isotropic FLAIR.

For datasets including Dual Echo acquisitions, the PD image is discarded because of the poor contrast in the sequences. For each subject, an expert reader manually segmented the hyperintensity WML in the FLAIR sequence that is used as the ground truth.

[1] available at: http://www.irisa.fr/visages/benchmarks/

[3] available at: http://www.fmrib.ox.ac.uk/fsl/index.html

3.2 Evaluation

Preprocessing Workflows Five different workflows are tested in this experiment. For all the subjects, the same parameters are used in every step. STREMv1 is the last step in the different workflows:

1. Basic : No preprocessing before registration.
2. NLM : Denoising before registration.
3. IIH : Intensity correction before registration.
4. IIH+NLM: Intensity correction and then denoising before registration.
5. NLM+IIH: Denoising and then intensity correction before registration.

STREMv1, STEMv1.5, STREMv2 In order to compare the different versions of STREM with each other, we choose a fixed workflow for the three versions of STREM and we evaluate the segmentation among the three approaches. NLM+IIH is chosen because the first experiment demonstrated that it is the best preprocessing workflow for STREMv1.

3.3 Evaluation metrics

The comparison of the different workflows and segmentation methods is done by assessing the segmentation quality and the processing time. In order to evaluate the differences between the automatic segmentation and the ground truth, we use the Dice Similarity Coefficient (DSC) that is commonly used in all kinds of segmentation evaluations [15]. A $DSC > 0.70$ is usually regarded as excellent agreement [15].

Computation time is also measured in order to compare the two different initializations methods and how the initialization affects the computation time. The PC used in the experiments is an Intel(R) Core(TM)2 CPU 2.66GHz with 2GB RAM.

3.4 Results

STREMv1, STREMv1.5, STREMv2 Table 1 (left) clearly demonstrates that STREMv2 in combination with NLM+IIH yields the best results in terms of DSC. Comparison of v1 and v1.5 algorithms shows that the results are exactly the same. This means that both algorithms give the same model parameters, but, as shown in Table 1 (right), the initialization of STREMv1 outperforms as the computation time is shorter. Multiple mEM give a better initialization than the atlas registration for the multidimensional mEM so this last one needs less iterations to converge. If we compare STREMv1.5 and STREMv2.0, we can see how the new rules defined to remove false positives increase DSC values without significantly increasing the computing time.

Preprocessing workflows Table 1 (left) and Figure 2 show the results of all different workflows. The NLM+IIH workflow displays better results for each subject. The preprocessing task which most impacts the segmentation results is the IIH correction. This result is comprehensible as STREMv1 is based exclusively on intensities, thus spatial IIH causes a poor detectability of WML.

On the other hand the effects of denoising are more difficult to evaluate. Denoising without IIH correction may not be of any help if images have a large IIH as in Subject 3. In addition, the order of preprocessing steps is not obvious. Denoising methods are supposed to work better with piecewise constant regions, that is why IIH correction is usually used before denoising [16]. The drawback is that IIH correction changes locally the nature of the noise when correcting a multiplicative inhomogeneity field affecting the denoising algorithm. In our experiments NLM+IIH works better than IIH+NLM.

	Subj1	Subj2	Subj3
Basic(v1)	0.31	0.20	0.42
NLM(v1)	0.33	0.29	0.42
IIH(v1)	0.31	0.49	0.53
IIH+NLM(v1)	0.30	0.46	0.48
NLM+IIH(v1)	0.31	0.49	0.56
NLM+IIH(v1.5)	0.31	0.49	0.56
NLM+IIH(v2)	0.38	0.66	0.64

	IIH	NLM.	Reg.	Atl.	STREM	Total
Basic(v1)	0	0	339	99	2741	3179
NLM(v1)	0	429	347	136	2486	3398
IIH(v1)	68	0	387	186	1082	1723
IIH+NLM(v1)	68	424	355	48	911	1806
NLM+IIH(v1)	80	429	362	140	923	1934
NLM+IIH(v1.5)	80	429	362	0	584	1455
NLM+IIH(v2)	80	429	362	0	588	1459

Fig. 1. Left Table: DSC values for all subjects and workflows. Right Table: Computation time in seconds for subject1 of each step and the total time, where Atl. means atlas registration for the initialization of STREMv1.

4 Discussion & Conclusion

Segmentation methods are usually evaluated without considering the preprocessing workflow, but, as we have shown, the choice of the preprocessing tasks and their order in the overall workflow have a great impact on the segmentation performance. From our point of view, this topic should be studied for every method which requires a complex workflow.

We presented a fully automated workflow for WML segmentation, but further studies are necessary in order to address open questions. In this study, we have not compared different algorithms for the same task and we have evaluated the influence of using generic algorithms, but specific algorithms for MS patients may be more efficient for our workflow. This study will be extended recruiting more subjects and different algorithms in order to create a more reliable and more specific tool for its application in clinical MS trials.

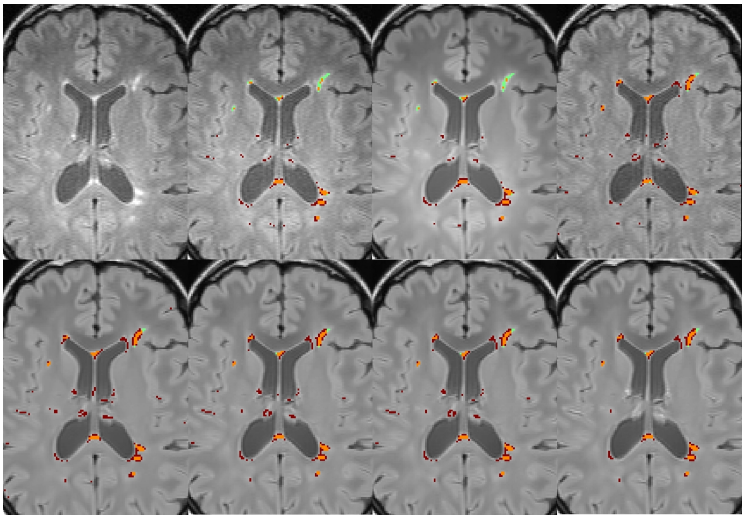

Fig. 2. Comparison of results for subject 3: From left to right, and from top to bottom: FLAIR sequence, Basic(v1), NLM(v1), IIH(v1), IIH+NLM(v1), NLM+IIH(v1), NLM+IIH(v1.5), NLM+IIH(v2). Color coding is: Orange: True Positives, Green: False Negatives, Red: False Positives

References

1. Miller, D.H., Grossman, R.I., Reingold, S.C., McFarland, H.F.: The role of magnetic resonance techniques in understanding and managing multiple sclerosis. Brain **121 (Pt 1)** (Jan 1998) 3–24
2. Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G.J., Plummer, D.L., Tofts, P.S., McDonald, W.I., Miller, D.H.: Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn Reson Imaging **14**(5) (1996) 495–505
3. Zijdenbos, A.P., Forghani, R., Evans, A.C.: Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Trans Med Imaging **21**(10) (Oct 2002) 1280–1291
4. Aït-Ali, L.S., Prima, S., Hellier, P., Carsin, B., Edan, G., Barillot, C.: STREM: a robust multidimensional parametric method to segment MS lesions in MRI. Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv **8**(Pt 1) (2005) 409–416
5. Wells, W.M., I., Grimson, W., Kikinis, R., Jolesz, F.: Adaptive segmentation of MRI data. Medical Imaging, IEEE Transactions on **15**(4) (Aug. 1996) 429–442
6. Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P.: Automated segmentation of multiple sclerosis lesions by model outlier detection. Medical Imaging, IEEE Transactions on **20**(8) (Aug. 2001) 677–688
7. Dugas-Phocion, G., Gonzalez, M., Lebrun, C., Chanalet, S., Bensa, C., Malandain, G., Ayache, N.: Hierarchical segmentation of multiple sclerosis lesions in multisequence MRI. In: Biomedical Imaging: Macro to Nano, 2004. IEEE International Symposium on. Volume 1. (April 2004) 157–160

8. Biernacki, C., Celeux, G., Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics and Data Analysis **41**(3-4) (2003) 561–575
9. Barkhof, F., Filippi, M., Miller, D.H., Scheltens, P., Campi, A., Polman, C.H., Comi, G., Adèr, H.J., Losseff, N., Valk, J.: Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis. Brain **120 (Pt 11)** (Nov 1997) 2059–2069
10. Mangin, J.F.: Entropy minimization for automatic correction of intensity nonuniformity. In: Mathematical Methods in Biomedical Image Analysis. IEEE Workshop on. (June 2000) 162–169
11. Coupé, P., Yger, P., Barillot, C.: Fast Non Local Means Denoising for 3D MR Images. In Larsen, R., Nielsen, M., Sporring, J., eds.: 9th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'2006. Volume 4191 of Lecture Notes in Computer Science., Copenhagen, Denmark, Springer (October 2006) 33–40
12. Wiest-Daesslé, N., Prima, S., Morrissey, S., Barillot, C.: Validation of a new optimisation algorithm for registration tasks in medical imaging. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI'2007, Washington, USA (April 2007) 41–44
13. Smith, S.M.: Fast robust automated brain extraction. Hum Brain Mapp **17**(3) (November 2002) 143–155
14. Fennema-Notestine, C., Ozyurt, I.B., Clark, C.P., Morris, S., Bischoff-Grethe, A., Bondi, M.W., Jernigan, T.L., Fischl, B., Segonne, F., Shattuck, D.W., Leahy, R.M., Rex, D.E., Toga, A.W., Zou, K.H., Brown, G.G.: Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. Hum Brain Mapping **27**(2) (Feb 2006) 99–113
15. Zijdenbos, A., Dawant, B., Margolin, R., Palmer, A.: Morphometric analysis of white matter lesions in mr images: method and validation. Medical Imaging, IEEE Transactions on **13**(4) (Dec 1994) 716–724
16. Montillo, A., Udupa, J.K., Axel, L., Metaxas, D.N.: Interaction between noise suppression and inhomogeneity correction in MRI. Volume 5032., SPIE (2003) 1025–1036