NeuroImage: Clinical 22 (2019) 101709

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Brain structure segmentation in the presence of multiple sclerosis lesions

Sandra González-Villà[a,b,*], Arnau Oliver[a], Yuankai Huo[b], Xavier Lladó[a], Bennett A. Landman[b]

[a] Institute of Computer Vision and Robotics, University of Girona, Ed. P-IV, Campus Montilivi, 17003 Girona, Spain
[b] Electrical Engineering, Vanderbilt University, Nashville, TN 37235, USA

ARTICLE INFO

Keywords:
Brain structures
Parcellation
Multiple sclerosis lesions
Segmentation
MRI
Multi-atlas
Label fusion

ABSTRACT

Intensity-based multi-atlas segmentation strategies have shown to be particularly successful in segmenting brain images of healthy subjects. However, in the same way as most of the methods in the state of the art, their performance tends to be affected by the presence of MRI visible lesions, such as those found in multiple sclerosis (MS) patients. Here, we present an approach to minimize the effect of the abnormal lesion intensities on multi-atlas segmentation. We propose a new voxel/patch correspondence model for intensity-based multi-atlas label fusion strategies that leads to more accurate similarity measures, having a key role in the final brain segmentation. We present the theory of this model and integrate it into two well-known fusion strategies: Non-local Spatial STAPLE (NLSS) and Joint Label Fusion (JLF). The experiments performed show that our proposal improves the segmentation performance of the lesion areas. The results indicate a mean Dice Similarity Coefficient (DSC) improvement of 1.96% for NLSS (3.29% inside and 0.79% around the lesion masks) and, an improvement of 2.06% for JLF (2.31% inside and 1.42% around lesions). Furthermore, we show that, with the proposed strategy, the well-established preprocessing step of lesion filling can be disregarded, obtaining similar or even more accurate segmentation results.

1. Introduction

Multiple sclerosis (MS) is an inflammatory, demyelinating and neurodegenerative disease of the central nervous system involving immune-mediated destruction of myelin and axonal damage that affects both white matter (WM) and gray matter (GM). MS is characterized by the formation of focal inflammatory lesions, also called plaques. Besides these demyelinating lesions, the disease also causes degeneration which makes patients experiment a consequent brain volume loss, known as atrophy, as the disease progresses.

GM atrophy in the brain has been shown to be associated with cognitive impairment in MS (Amiri et al., 2018; Nocentini et al., 2014; Eijlers et al., 2018), being a better predictor than WM lesion volume (Tillema et al., 2016), and also relevant to disease progression (Jacobsen et al., 2014). Furthermore, deep GM atrophy has been shown to be associated with the development of definite MS and disability progression in early relapsing remitting MS (Zivadinov et al., 2013; Debernard et al., 2015). The effect of the disease on isolated structures has also been studied, concluding that the thalamus atrophy is a clinically relevant biomarker of the neurodegenerative disease process (Houtchens et al., 2007).

Both MS lesions and brain atrophy, are usually measured in-vivo from magnetic resonance images (MRI) by means of automatic or semi-automatic algorithms. The most frequent modalities to segment WM lesions include PD-w, FLAIR and T2-w, since lesions appear hyper-intense in these sequences which makes them easier to detect. However, cortical lesions are rarely visualized in these modalities, which makes other sequences such as DIR, PSIR or MP-RAGE also useful for finding MS lesions. On the other hand, brain atrophy is measured from T1-w images (where lesions appear hypo-intense), due to the high contrast between tissues shown in this modality. To measure the atrophy of different brain structures, the most common procedure is to compute structure volumes from a previous segmentation of the brain.

Several automatic methods have been proposed in the literature to segment the brain on its structures (González-Villà et al., 2016; Iglesias and Sabuncu, 2015), that can be classified based on the strategy followed, including, learning-based (Kushibar et al., 2018; Fischl et al., 2002), atlas-based (Warfield et al., 2004; Iglesias et al., 2012; Artaechevarría et al., 2009), deformable-based (Patenaude et al., 2011), etc. Among all the approaches proposed in the literature, multi-atlas methods have been demonstrated to be robust and provide good segmentation results on healthy subjects (González-Villà et al., 2016; Heckemann et al., 2010; Asman and Landman, 2013). In this strategy, a set of MR images with available manual segmentation, i.e. atlases, are

[*] Corresponding author at: Institute of Computer Vision and Robotics, University of Girona, Ed. P-IV, Campus Montilivi, 17003 Girona, Spain.
E-mail address: sgonzalez@eia.udg.edu (S. González-Villà).

https://doi.org/10.1016/j.nicl.2019.101709
Received 24 December 2018; Accepted 3 February 2019
Available online 14 February 2019

non-rigidly registered to the target MR image. After that, the deformation fields obtained from these registrations are applied to the corresponding segmentations in such a way that new pairs of images (structural image and segmentation) are obtained, which are similar to target. Then, these candidate segmentations of the target are fused (i.e., label conflicts between the candidate segmentations are resolved voxel-wise) to obtain the final segmentation. Several fusion strategies have been proposed in recent years (Artaechevarría et al., 2009; Huo et al., 2017; Wang et al., 2013), being the ones helped by the structural image intensities the ones that provide the best results (González-Villà et al., 2016; Akhondi-Asl and Warfield, 2013; Landman and Warfield, 2012).

Multi-atlas label fusion strategies based on intensities exploit the target-atlas similarity under the assumption that images with similar appearance are more likely to have similar segmentations. A successful approach, inspired by the non-local means method (Buades et al., 2005) and first introduced by Coupé et al. (2011) to multi-atlas segmentation, utilizes a patch based search strategy, i.e. search strategy that aims at matching cubic/squared blocks of the image (patches) with the most similar patch in a second image, to identify correspondences with the atlases. This technique assumes that registration errors are inherent in multi-atlas segmentation due to several facts such as the regularization constraints involved in that process or the failure to reach a global optimum of the objective function. In order to overcome the registration errors, these methods relax the one-to-one mapping constraint existing in traditional weighting methods and re-compute the correspondences for every voxel/patch of the target image and the atlases before segmentation, usually based on intensity similarity. However, in the same way as most of the proposed brain structure algorithms in the state of the art (González-Villà et al., 2017), these strategies are designed to segment healthy subjects and, their performance tends to be affected by the presence of MS lesions.

Recently, different lesion filling approaches (Valverde et al., 2014; Chard et al., 2010; Battaglini et al., 2012; Prados et al., 2016) have been successfully proposed in order to minimize the effect of the abnormal MS lesion intensities on the segmentation. Those techniques have been demonstrated to improve tissue measurements (Valverde et al., 2014; Chard et al., 2010; Battaglini et al., 2012; Prados et al., 2016), and have been used ad-hoc also for brain structure segmentation (Gelineau-Morel et al., 2012; Batista et al., 2012). While the effect of lesion filling has been analyzed on several tissue segmentation strategies, in which intensity distributions of different tissue classes are modelled, it has not been studied the effects on patch based segmentation strategies, i.e. strategies that are based on image patches instead of using the whole image at once, in which patch intensities are independent to the global intensity distributions.

In this work, we propose a new correspondence search approach for multi-atlas label fusion, that can be applied to segment either healthy subjects or patients with MRI visible lesions. We introduce a new correspondence model, able to deal with brain irregularities, such as MS lesions. We assume that the abnormal lesion intensities may affect the correspondence finding on the healthy atlases, obtaining more inaccurate matches than the ones obtained after a masked registration to the atlas, i.e. ignoring the lesion voxels. For this reason, we force the correspondence imposed by the registration result on the lesion areas, whereas we redefine the patch shape on the surroundings of the lesion to prevent these abnormal intensities from interfering in the correspondence search. We integrate this model into two well-known label fusion strategies (Non-local Spatial STAPLE (Huo et al., 2017) and Joint Label Fusion (Wang et al., 2013)), reformulating the original methods to improve the correspondences in the lesion areas, while maintaining the original search model in the rest of the brain.

2. Materials and methods

2.1. Data

Public databases of patients with lesions and including both brain parcellation and lesion annotations are uncommon. Therefore, to test our approach in a large number of cases, we need to run the experiments with simulated data. Specifically, we have simulated artificial multiple sclerosis lesions on a database of 45 healthy patients that included brain parcellation annotations. To simulate the lesions, 140 MS patients from five different databases (including MICCAI'08 (Styner et al., 2008), MICCAI'16 (Commowick et al., 2018), ISBI'15 (Carass et al., 2017), and two in-house databases) were analyzed, and the 45 patients with larger lesion volume were selected as basis for simulation. The selected 45 patients were paired with the annotated healthy images based on their lateral ventricle size in order to pair similarly atrophied brains. Once the couples were assigned, each MS patient image was non-rigidly registered to its corresponding atlas (Avants et al., 2008) (previous initial affine registration (Ourselin et al., 2001)), masking out the lesion areas for more adjusted registration. Then, the normalized intensities of the registered lesions were copied to the atlases, obtaining a new database of simulated patients with lesion volumes ranging from 1.16–68.43 ml, and voxel spacing 1 × 1 × 1 mm.

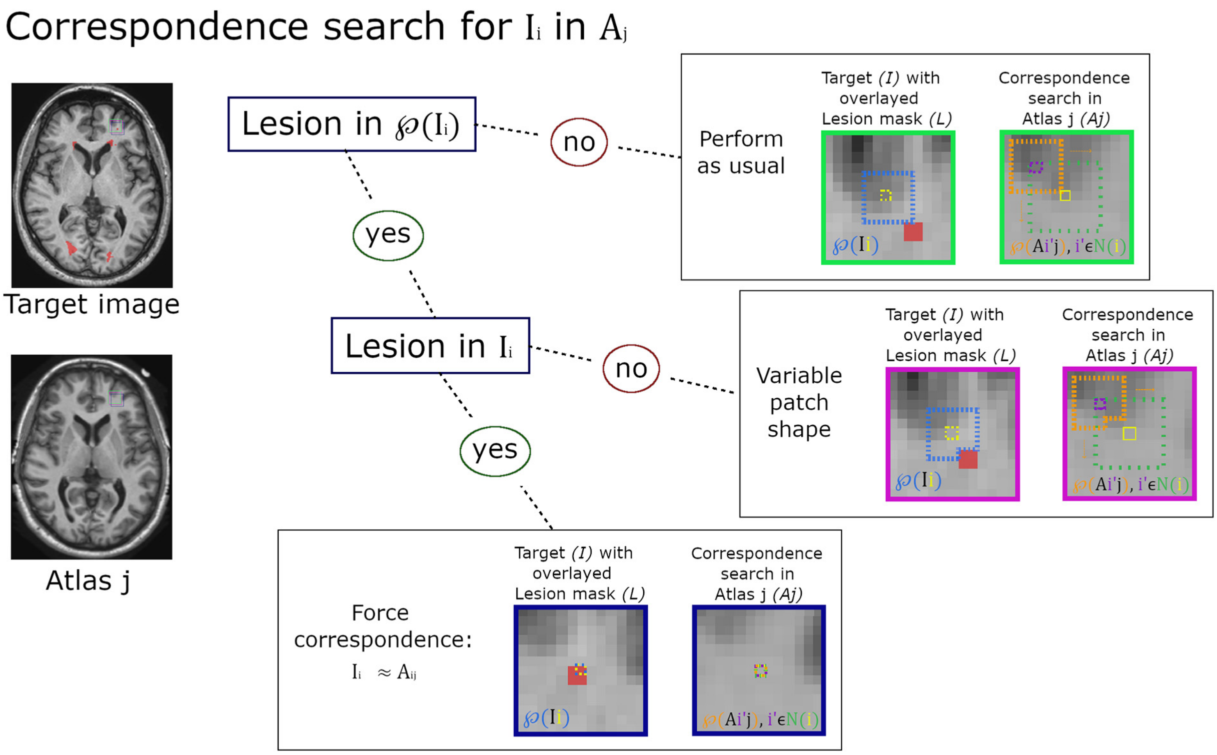
The healthy subject dataset consists of 45 T1-w MR images obtained from the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database (Landman and Warfield, 2012). The images were obtained from Open Access Series on Imaging Studies (OASIS) dataset (Marcus et al., 2007) and labeled according to BrainCOLOR protocol (Klein et al., 2010), including 133 labels that cover the whole brain: subcortical structures, ventricles, cerebral WM, cerebellum, brainstem and 98 regions in the cortex (see supplementary material for more information).

Nevertheless, the recent MRBrains 2018 challenge dataset (MICCAI, 2018) allows us to test our approach also with real data. Although this dataset consists of 30 MRI images obtained from patients with varying degrees of atrophy and white matter lesions, only 7 cases have been released with both lesion and tissue segmentation available. For these 7 patients, lesion load ranges from 0.06 to 70.00 ml. As we use the atlases from the previous dataset, the 133 brain structures labels have been combined to obtain an 8-class segmentation: background, cortical gray matter, basal ganglia, white matter, cerebrospinal fluid in the extra-cerebral space, ventricles, cerebellum and brainstem. Voxel spacing for the images in this dataset is 0.9583 × 0.9583 × 3 mm.

Severely atrophied brains were present in both databases. In the first one, in spite of lesions were simulated on a cohort of healthy subjects, their ages range from 18 up to 90 years old, and therefore, some of them present age-related atrophy.

2.2. Preprocessing

The atlases used in our experiments include the 45 images from the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database (Landman and Warfield, 2012). For the atlas registration, PCA atlas selection is performed and only the 15 most similar atlases are used for segmentation. In the experiments performed on the simulated dataset, since lesions are simulated in images of the same database, the one being analyzed is excluded from the atlas selection, keeping as candidate atlases the remaining 44 images. All the images are histogram normalized and N4 (Tustison et al., 2010) bias field corrected before registration. All the pair-wise registrations are performed using an initial affine registration (Ourselin et al., 2001) followed by a non-rigid (Avants et al., 2008) procedure. In all the registrations performed, the lesions are masked-out to avoid their intensities to interfere in the similarity metric calculation. For a fair comparison of the fusion methods, the same registration results are used for all the strategies.



Correspondence search for Ii in Aj
Target image
Atlas j
Lesion in ℘(Ii)
yes
no
Lesion in Ii
yes
no
Perform as usual
Variable patch shape
Force correspondence: Ii ≈ Aij
Target (I) with overlayed Lesion mask (L)
Correspondence search in Atlas j (Aj)
℘(Ii)
℘(Ai'j), i'ϵN(i)

Fig. 1. Correspondence search scheme. Search for the correspondence of voxel i of the target image I (I_i) on the atlas j (A_j). When there are not lesions in the patch of voxel i, our model performs as the original method (best correspondence is found comparing the target patch of voxel i, i.e. $\wp(I_i)$, to all the atlas patches of the voxels $\acute{i}$ that belong to the neighborhood of i, i.e.$\wp(A_{i'j})$, $i' \in \mathcal{N}(i)$. On the other hand, when there are lesions in the patch, we modify the patch shape to exclude the lesions from the search and use the same patch shape to find the correspondence in the atlas in the same way as before. Finally, when the target voxel i is part of a lesion, we trust the masked registration result and force the correspondence to be A_{ij}. Note that this example is shown in 2D for simplification, where the patch size is set to 5 × 5 (blue and orange striped squares) and the search neighborhood to 7 × 7 (green striped square). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.3. Voxel and patch correspondences

A substantial source of error in multi-atlas label fusion is registration inaccuracy. Registration is a very complex task, which is considered one of the fundamental problems in medical image processing, and may not always give maximum local similarity between image patches. Because of that, some fusion strategies have successfully tried to reduce these registration errors by means of local search windows, that relax the traditional one-to-one mapping constraint. That is, given the patch centered at voxel i of the target image I ($\wp(I_i)$), it is often possible to find a patch $\wp(A_{i'j})$ centered at voxel $\acute{i}$ of the atlas image j which is more similar to $\wp(I_i)$ than to the corresponding patch $\wp(A_{ij})$ centered at voxel i of the atlas j. This similarity is often computed based on image intensities, which constitutes a challenge when the target image presents visible lesions. The lesions show different intensity profiles to that of the atlases healthy tissues, making tough the correspondence search problem.

In this section, we present a new correspondence search approach for patch-based multi-atlas segmentation (label fusion), that can be applied to segment either healthy subjects or patients with MRI visible lesions. A graphical representation of this idea is depicted in Fig. 1. In order to avoid the interference of the lesions on the correspondence search, we assume that such correspondence cannot be further improved inside the lesions based on intensities and enforce them on that areas in such a way we give more weight to the masked registration result. Note that this premise could be applied to any intensity-based multi-atlas label fusion strategy, not only to solve for voxel/patch correspondences but also to estimate the final voting weights, on weighted voting strategies, that lead to the segmentation result, as we will see in Section 2.3.3.

In the following, we reformulate two well-known label fusion strategies to include this idea: Non-local Spatial STAPLE (Huo et al., 2017), and Joint Label Fusion (Wang et al., 2013).

2.3.1. Problem definition

Consider a target gray-level image (with lesions) represented as a vector $I \in \mathbb{R}^{N\times1}$. Let $L \in \{0,1\}^{N\times1}$be a binary lesion mask indicating whether a given voxel i of the target image contains or is part of a lesion, hence $L_i = f(I_i \in lesion)$. Note that the lesion mask is optional and can be neglected if all voxels in it are set to 0, making the modified algorithms behave as its originals. Consider also a set R of registered healthy atlases with associated gray level images, $A \in \mathbb{R}^{N\times R}$.

2.3.2. Masked Non-local Spatial STAPLE (m-NLSS)

Non-local Spatial STAPLE (NLSS) (Huo et al., 2017) is a variant of the STAPLE (Warfield et al., 2004) algorithm from a non-local means perspective. In NLSS, the labels of all the atlas voxels in the neighborhood of the target voxel have a weight in its label assignment based on their intensity similarity. That is, they provide a model in which they learn which label each atlas would have observed given the perfect correspondence with the target and integrate this model into the STAPLE framework.

As stated before, lesion intensities may affect the result of this non-local correspondence model, obtaining wrong voxel correspondences and sometimes even worse than the ones obtained by the one-to-one mapping resulting from masked registration. For this reason, following the previously stated assumption, we define the probability of correspondence between voxel i of the target image and voxel $\acute{i}$ of the j-th atlas ($\alpha_{ji'i}$), i.e. the non-local correspondence model, as follows:

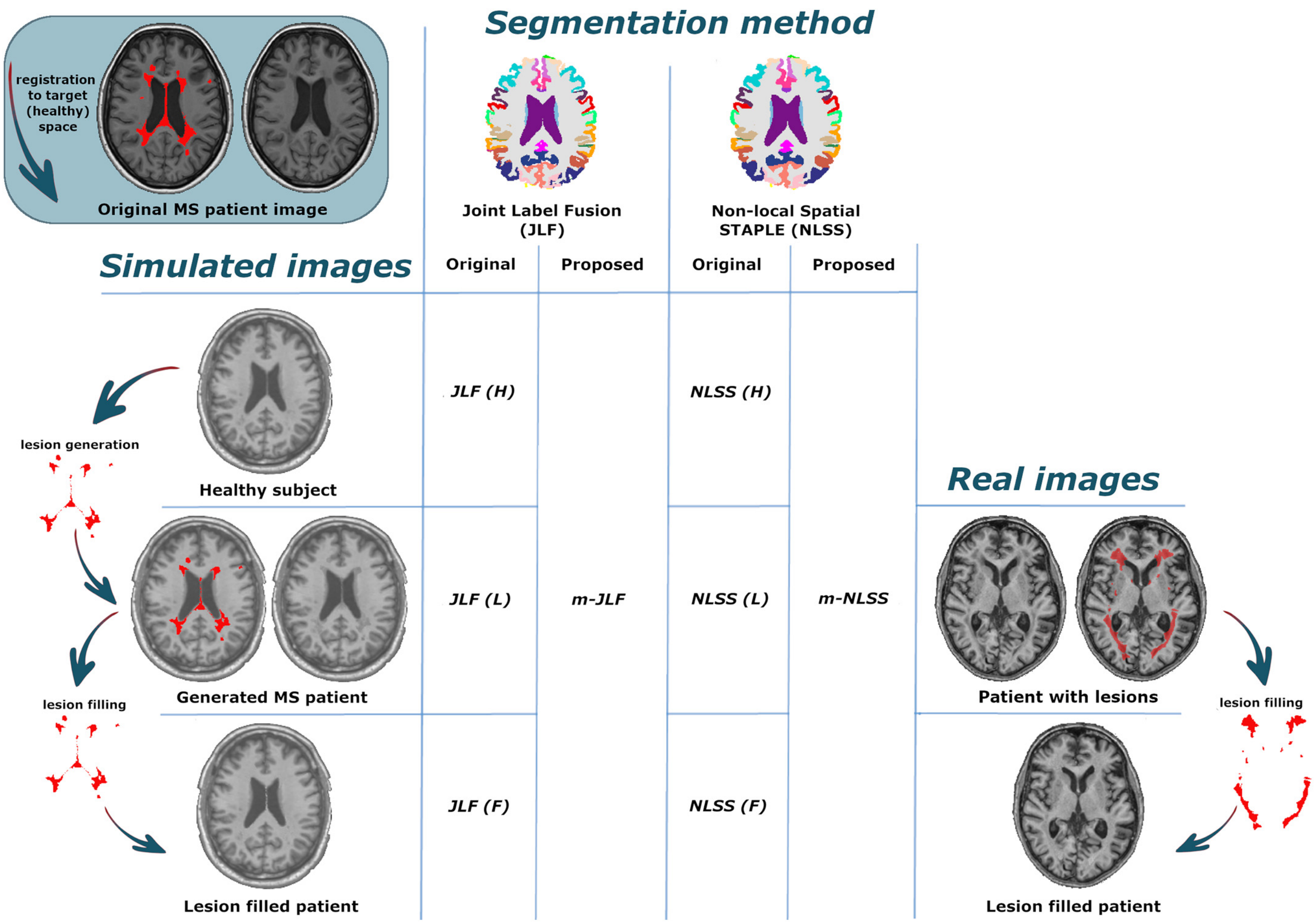


Fig. 2. Evaluation procedure for both the simulated and the real databases. In the simulated database, the original MS patient is registered to the healthy space and the intensities of the registered lesions are copied to the healthy subject. For evaluation, each of the images shown is segmented with the original methods (JLF and NLSS) and the proposed ones (m-JLF and m-NLSS). The performance of the methods is assessed individually for each segmented image, i.e. healthy/lesioned/filled for the simulated database and lesioned/filled for the real database, based on the DSC difference of the proposed method and the original one. Note that the segmentation result when using the proposed method on the healthy, lesioned and filled images will be the same, i.e. the intensities inside the lesion mask are irrelevant for our method and the atlas registration results are the same for all the images analyzed, thus, we only segment the image with lesions.

$$\alpha_{ji\prime i} = \left(\frac{1}{Z_\alpha} \cdot e^{-\frac{\| \wp_{L_i} \circ (\wp(A_{i\prime j}) - \wp(I_i)) \|_2^2}{2\sigma_i^2 \cdot \| \wp_{L_i} \|}} \cdot e^{-\frac{\varepsilon_{i\prime i}^2}{2\sigma_d^2}} \right) \cdot (1 - L_i) + \delta(i\prime = i) \cdot L_i \tag{1}$$

where $\wp(\cdot)$ is the set of intensities in the patch neighborhood of a given intensity location. In this definition,

$\wp_{L_i} = \wp(1 - L_i)$ is the masking term that excludes lesion voxels from the patch calculation and enforces the same patch neighborhood size/shape in both the atlas and the target, $\|\wp_{L_i} \circ (\wp(A_{i'j}) - \wp(I_i))\|_2{}^2$ is the L2-norm between the atlas patch centered at $i^{'}$ and the target patch centered at i, $\varepsilon_{i'i}{}^2$ is the Euclidean distance in physical space between i and $i^{'}$, σ_i and σ_d are the standard deviations for the intensity and distance weights, and Z_α is a partition function that enforces the constraint that $\sum_{i' \in N(\boldsymbol{i})} \alpha_{ji\prime i} = 1$, where $\mathscr{N}(i)$ is the set of voxels in the search neighborhood of a given target voxel. $\delta(i^{'} = i)$ is the Dirac delta function, and $\|\wp_{L_i}\|$ is the number of voxels in the patch neighborhood.

A more detailed and extended formulation of this method can be found in the paper (González-Villà et al., 2018).

2.3.3. Masked Joint Label Fusion (m-JLF)

Joint Label Fusion (JLF) (Wang et al., 2013) is based on the idea that different atlases may produce similar label errors. They assume that the errors produced by the atlases are not independent and address this issue by computing the intensity similarity between the target and each pair of atlases, which allow them to estimate the probability that a pair of atlases produce the same segmentation error.

In order to estimate this pairwise dependency matrix, the authors first solve for the patch correspondences between the target and each atlas, as they also assume registration errors. As these correspondences are computed based on patch intensities, we know that the lesion intensities may interfere on this local patch search, and therefore we redefine the local search correspondence map between the atlas j and the target as follows:

$$i'_{ij} = arg \min_{i' \in N(\boldsymbol{i})} [\|\wp_{L_i} \circ (\wp(A_{i,j}) - \wp(I_i))\|^2 \cdot (1 - L_i) + \delta(i\prime = i) \cdot L_i] \tag{2}$$

where $\wp_{L_i}$ is the same masking term than in Section 2.3.2, $\|\wp_{L_i} \circ (\wp(A_{i'j}) - \wp(I_i))\|^2$is the (masked) sum of squared differences of the non-lesion voxels, $\mathscr{N}(i)$ is the set of voxels in the search neighborhood of a given target voxel, and $\delta(i^{'} = i)$ the Dirac delta function.

In the same way lesions interfere in the local patch search, they also have to be modelled in the pairwise dependency matrix, which estimates how likely two atlases are both to produce wrong segmentation for the target image, given the observed joint patch intensity differences. Following our assumption, we reformulate the matrix of expected pairwise joint label differences between the j-th and k-th atlases, $M_i(j,k)$, as follows:

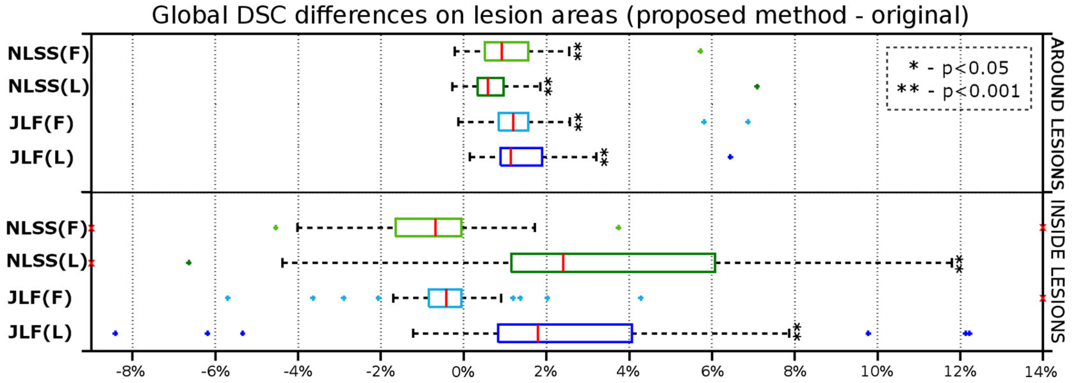


Fig. 3. Global DSC differences on the MS simulated database. Differences between the proposed strategies (m-JLF/m-NLSS) and their corresponding original methods on the lesion areas: inside the lesion masks and on a mask that includes three voxels of the lesion mask contour. Segmentation differences performed for (1) the simulated patients (m-JLF– JLF(L) and m-NLSS– NLSS(L)) and, (2) the lesion filled generated patients (m-JLF– JLF(F) and m-NLSS– NLSS(F)). Statistical significance assessed independently for each boxplot in the figure, which represents the relationship between the proposed strategy and the original method. By means of paired *t*-tests, we test the null hypothesis that the true mean DSC difference between both methods (proposed and depicted) is zero.

$$M_i(j,k) = \left[\frac{(\wp_{L_i} \circ |\wp(I_i) - \wp(A_{i'^{ij}j})|) \cdot (\wp_{L_i} \circ |\wp(I_i) - \wp(A_{i'^{ik}k})|)}{\|\wp_{L_i}\|} \right]^{\beta} \tag{3}$$

where β is a model parameter controlling the weight distribution, $\|\wp_{L_i}\|$ is the number of non-lesion voxels in the patch neighborhood, $A_{i'_{ij}j}$ and $A_{i'_{ik}k}$ are the corresponding voxels of target image voxel i, i.e. I_i, on atlases j and k, respectively.

Finally, the voting weights that lead to the final segmentation, can be estimated from $M_i(j,k)$, in the same way as in the original formulation (Wang et al., 2013).

2.4. Evaluation

To evaluate the usefulness of our strategy, in the first experiment, we compare the performances of our approaches (m-NLSS and m-JLF) with respect to the ones obtained by the original algorithms (NLSS and JLF) when testing the following cases: (i) original MRI images of healthy subjects (H), (ii) images with synthetic MS lesions (L), and (iii) images with synthetic MS lesions after applying a recent lesion filling algorithm (Valverde et al., 2014) (F). Notice that using our proposed strategy, the intensities inside the lesion mask are not relevant and, thus, in the three cases, we obtained the same result. In contrast, the performance of the original algorithms varied in each case (obtaining in what follows, NLSS(H), NLSS(L), and NLSS(F), respectively, and similarly for the JLF algorithm). Fig. 2 shows the evaluated cases and the corresponding nomenclature.

The lesion filling technique used here (Valverde et al., 2014) replaces the lesion voxel intensities by random values of a normal distribution generated from the mean WM signal intensity of each two-dimensional slice. As stated by their authors, this technique is a compromise between global and local methods, reducing the bias caused by refilled voxels on GM and WM tissue distributions by means of global information from the whole slice, whereas aims to reproduce more precisely the signal variability between slices by means of re-computing the mean signal intensity of the normal appearing WM at each slice.

In a second experiment, we tested our approaches in the 7 cases of the MRBrains18 dataset. In this case, we compared the proposed correspondence algorithms (m-NLSS and m-JLF) when segmenting the original images (including lesions) to their originals when segmenting: (i) the original images (NLSS(L) and JLF(L)) and (ii) the images after applying the lesion filling algorithm (Valverde et al., 2014) (NLSS(F) and JLF(F)). See Fig. 2 for details.

We quantitatively evaluate the segmentation results using the global Dice Similarity Coefficient (DSC) across all the structures affected by lesions:

$$\text{DSC}_{global} = \frac{2 \cdot \sum_{l \in \mathscr{L}} |T_l \cap E_l|}{\sum_{l \in \mathscr{L}} |T_l| + \sum_{l \in \mathscr{L}} |E_l|} \tag{4}$$

where T_l is the ground truth segmentation for label l, E_l is the estimate for the label l, and $\mathscr{L}$ is the set of all the available labels.

As the presence of lesions not necessarily affect only the lesion area segmentation itself, but also the surrounding tissues, two measures are calculated: (1) DSC inside the lesion mask and, (2) DSC inside a mask that includes three voxels of the lesion contour. Note that $\mathscr{N}(i)$ was set to 7 × 7 × 7. Besides, to give an overview of the global performance of the strategies, and to evaluate how the lesions affect the segmentation of the whole brain, we also compute the mean DSC across all the structures, and independently for white matter and cortical and subcortical gray matter.

To provide a better comparison between the methods, DSC differences are shown instead of the DSC itself, being the difference of the DSC obtained using our strategy and the original method. Notice that these differences are computed subject by subject, which relates the performance for the same subject with respect to the proposed methods and their original ones.

Statistical analysis is performed using the Matlab software package. Differences in the performance of the analyzed methods are computed using paired-sample ttests. Moreover, the Pearson's linear correlation coefficient is used to compute the correlation between the total lesion volume and the changes in mean DSC.

For a fair comparison of the fusion methods all the parameters are set to the same values in all the original and proposed methods. The search neighborhood is set to 7 × 7 x 7, patch dimensions to 5 × 5 × 5 and σ_i, σ_d and β are set to 0.25, 1.5 and 2, respectively.

3. Results

3.1. Simulated MS lesions

First, we perform an analysis of the segmentation results on the lesion areas. Fig. 3 shows the global DSC differences between the proposed strategies when segmenting the images with lesions (m-JLF and m-NLSS) and the original methods when segmenting: (1) the images with lesions (JLF(L) and NLSS(L)) and, (2) the lesion filled images (JLF (F) and NLSS(F)) (see Fig. 2 for setup details). Notice that each boxplot represents the subtraction of the original method performance from our method's. Hence, positive values indicate an improvement of our proposal with respect to the depicted method. Furthermore, each boxplot significance was assessed independently and represents the relationship with the proposed strategy, and therefore, they are independent to each other. Differences in performances were assessed by means of paired *t*-

tests between the original strategies (shown in Fig. 3) and the proposed ones. The results have been analyzed separately inside the lesion mask and on a region that includes three voxels of the lesion mask contour. As observed in the figure, inside the lesion masks the proposed correspondence models (m-JLF and m-NLSS) performed significantly better than their originals (JLF(L) and NLSS(L)) when the lesions were present (both inside and around the lesion masks). On the other hand, analyzing the segmentation of the filled images with the original methods (JLF(F) and NLSS(F)), we observed that inside the lesion masks the performance was similar to that of our proposal, while around the lesion areas our strategy performed significantly better.

We also compared our proposals to the best possible segmentation the original algorithms could reach, i.e. segmenting the corresponding healthy subjects. We observed from that experiment that, inside the lesion areas, either segmenting the simulated images (with lesions) with the proposed methods (m-JLF and m-NLSS) or filling the lesions before segmentation (JLF(F) and NLSS(F)) did not reach the healthy segmentation performance. This behavior was expected, since the intensities of that areas were “corrupted” both in the image with lesions and in the filled one. However, analyzing the performance around the lesion areas, the proposed methods (m-JLF and m-NLSS) reached similar performance to that of the healthy segmentations (JLF(H) and NLSS(H)). On the other hand, when it comes to the lesion filled images, both strategies (JLF(F) and NLSS(F)) significantly underperformed the healthy segmentation (JLF(H) and NLSS(H)). Besides, both original methods (JLF(L) and NLSS(L)) significantly underperformed the best possible segmentations (JLF(H) and NLSS(H)).

In terms of whole brain segmentation performance (not only restricted to the lesion areas), our proposals (m-JLF and m-NLSS) provided significantly better segmentation results than the original methods when segmenting the simulated images (JLF(L) and NLSS(L)) and the filled images (JLF(F) and NLSS(F)), as observed in Fig. 4. However, whereas the whole brain segmentation performance of our NLSS nonlocal model (m-NLSS) was similar to that of the original method when segmenting the healthy subjects (NLSS(H)), our proposal for JLF (m-JLF) did not reach the best possible performance of the original method (JLF(H)), which is comprehensible since the real intensity information of the lesions is missing in the simulated images.

To give an overview of the whole brain segmentation performance, the mean DSC achieved by the analyzed methods was: 85.97 ± 1.47 (JLF(L)), 86.03 ± 1.46 (m-JLF), 86.05 ± 1.47 (JLF(H)), 85.99 ± 1.46 (JLF(F)), 79.27 ± 1.35 (NLSS(L)), 79.35 ± 1.34 (m-NLSS), 79.35 ± 1.33 (NLSS(H)), and 79.29 ± 1.34 (NLSS(F)).

In order to see where the bigger DSC changes, due to lesions, occurred within the brain, we performed the same analysis on the subcortical and cortical GM, and the WM separately. To do such analysis, the resulting labels were merged before computing the DSC in three groups: (1) cortical labels, (2) left and right cerebral WM, and (3) subcortical structures (both thalamus, putamens, pallidums, caudates, amygdalas, hippocampus and accumbens).

This second experiment, showed that in both methods, i.e. JLF and NLSS, the structure which experimented more performance variance was the WM, when comparing between our proposal (m-JLF/m-NLSS) and the original method segmenting: (1) the healthy images (JLF (H)/NLSS (H)) and, (2) the simulated images (JLF(L)/NLSS(L)). On the other hand, when checking for differences between our strategy (m-JLF/m-NLSS) and the original method segmenting the filled images (JLF (F)/NLSS (F)), we observed that the GM was more affected than the WM, in particular the subcortical structures, where more DSC variance was appreciated. In light of these findings, we see that lesion filling helps in achieving more accurate results than just segmenting the un-preprocessed image, however, the improvement is more visible on the WM than on the GM structures. The results of this analysis are depicted in Fig. 5.

Fig. 6 shows some qualitative results obtained with the analyzed methods. As can be observed from this figure, when the lesions are close to the GM, the original methods ((f) and (j)) tend to segment them as part of this tissue. On the other hand, if the lesions are filled before segmentation ((h) and (l)), GM structures tend to be underestimated. These two issues seem to be handled correctly by our proposals ((g) and (k)), which results look more similar to the healthy subjects segmentation ((e) and (i)) and the ground truth.

Lastly, we analyzed the extent to which total lesion volume affected the observed changes in DSC for the evaluated methods. Significant correlations were found on the DSC differences of the whole brain between JLF(L) and m-JLF ($r = 0.72$, $p < .001$), JLF(H) and m-JLF ($r = -0.50$, p $< .001$), JLF(F) and m-JLF ($r = 0.57$, p $< .001$), NLSS (L) and m-NLSS ($r = 0.86$, p $< .001$), and between NLSS(F) and m-NLSS ($r = 0.63$, p $< .001$). However, no correlation with the lesion load was found on the DSC changes between NLSS(H) and m-NLSS. On the other hand, when analyzing the connection between the total lesion load and the performance differences of segmenting the simulated (L) and the lesion filled (F) images, correlations were found for JLF ($r = 0.52$, p $< .001$), and NLSS ($r = 0.67$, p $< .001$). A more exhausted analysis, separated by tissue, i.e. subcortical structures, cortical GM and WM, is presented in Table 1.

3.2. MRBrainS 2018 challenge

On this database, the experiments performed showed that, in terms of global DSC differences, the modified correspondence models provided, in average, better segmentation results on the lesion areas, for

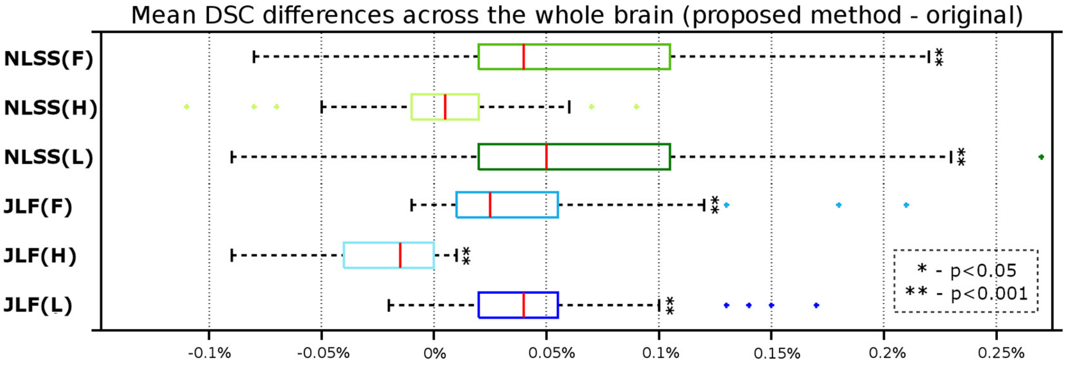


Fig. 4. Mean DSC differences on the MS simulated database. Differences for the whole brain between the proposed strategies (m-JLF/m-NLSS) and their corresponding original method. Segmentation performed for (1) the simulated patients (JLF(L)/NLSS(L)), (2) the healthy subjects (JLF(H)/NLSS(H)) and, (3) the lesion filled generated patients (JLF(F) /NLSS(F)). Statistical significance assessed independently for each boxplot in the figure, which represents the relationship between the proposed strategy and the original method. By means of paired t-tests, we test the null hypothesis that the true mean DSC difference between both methods (proposed and depicted) is zero.

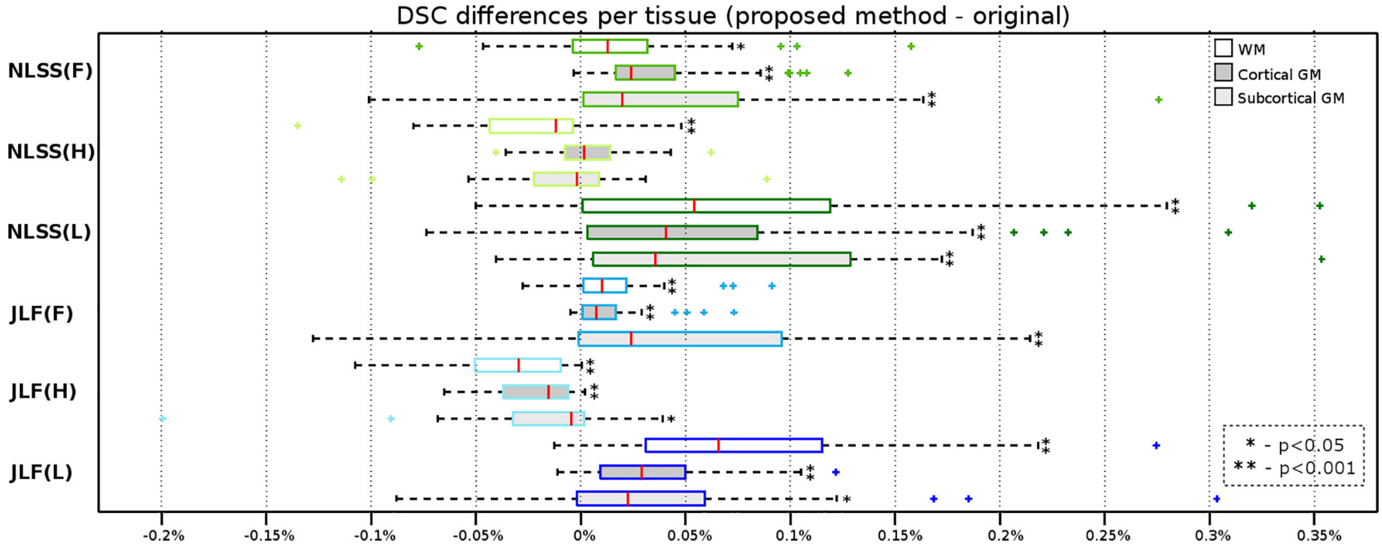


Fig. 5. Dice differences on the MS simulated database. Differences for the gray matter (GM) (cortical and subcortical) and the white matter (WM) between the proposed strategies (m-JLF/m-NLSS) and their corresponding original method. Segmentation performed for (1) the simulated patients (JLF(L)/NLSS(L)), (2) the healthy subjects (JLF(H)/NLSS(H)) and, (3) the lesion filled generated patients (JLF(F) /NLSS(F)). Statistical significance assessed independently for each boxplot in the figure, which represents the relationship between the proposed strategy and the original method. By means of paired *t*-tests, we test the null hypothesis that the true mean DSC difference between both methods (proposed and depicted) is zero.

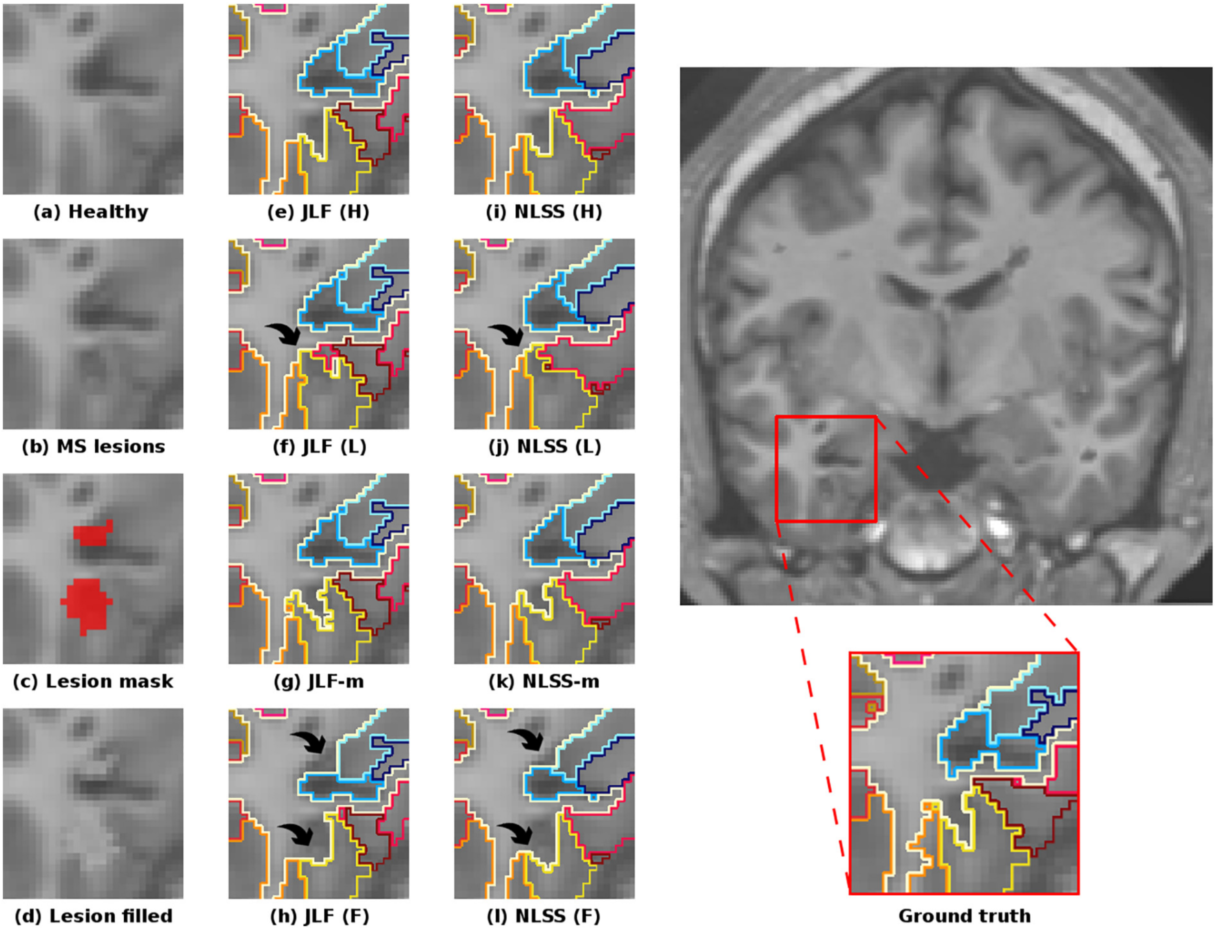

(a) Healthy
(e) JLF (H)
(i) NLSS (H)
(b) MS lesions
(f) JLF (L)
(j) NLSS (L)
(c) Lesion mask
(g) JLF-m
(k) NLSS-m
(d) Lesion filled
(h) JLF (F)
(l) NLSS (F)
Ground truth

Fig. 6. Structural images and segmentation results obtained for the analyzed cases. (a) Original healthy T1-w image, (b) simulated MS lesions on the healthy T1-w image, (c) lesion mask, (d) T1-w image after filling the lesions. Segmentation results of JLF on (e) the healthy subject, (f) simulated MS patient, and (h) lesion filled image; and NLSS on (i) the healthy subject, (j) simulated MS patient, and (l) lesion filled image. Proposed strategies for (g) JLF and, (k) NLSS.

Table 1
Pearson's correlation between the total lesion load and the DSC differences seen between pairs of methods (ref– other). Values calculated independently for the subcortical structures (subcort.), cortical gray matter (cortical) and white matter (WM).

		Lesions (L)			Healthy (H)			Filled (F)			Proposed		
		Subcort.	*Cortical*	*WM*	*Subcort.*	*Cortical*	*WM*	*Subcort.*	*Cortical*	*WM*	*Subcort.*	*Cortical*	*WM*
JLF	*p-val*	0.0024	< 0.0001	< 0.0001	0.0531	0.0003	< 0.0001	0.5967	0.0099	0.2573	ref	ref	ref
	R	0.4580	0.6746	0.8405	−0.2936	−0.5234	−0.6243	−0.0820	0.3847	0.1745	ref	ref	ref
	p-val	0.0004	0.0005	< 0.0001	0.7262	0.0008	0.0009	ref	ref	ref	–	–	–
	R	0.5099	0.5058	0.8354	−0.0543	−0.4854	−0.4834	ref	ref	ref	–	–	–
NLSS	*p-val*	< 0.0001	< 0.0001	< 0.0001	0.0054	0.0001	0.3220	0.1070	< 0.0001	0.2900	ref	ref	ref
	R	0.7772	0.6041	0.7861	−0.4125	0.5622	−0.1528	−0.2463	0.7717	0.1631	ref	ref	ref
	p-val	< 0.0001	0.0150	< 0.0001	0.8540	0.0013	0.0410	ref	ref	ref	–	–	–
	R	0.8199	0.3644	0.7999	0.0286	−0.4704	−0.3094	ref	ref	ref	–	–	–

both JLF and NLSS methods. On the other hand, analyzing the performance of the methods inside and around the lesion masks separately, we observed that inside the lesion masks, m-JLF over-performed JLF(L) on six cases out of seven, whereas m-NLSS showed better performance than NLSS(L) on five of the seven cases. When it comes to the surroundings of the lesions (around lesions), the proposed correspondence models always provided better segmentation results than their corresponding originals.

In terms of the effect of the lesions on the overall performance of the brain, we observed that, with the proposed correspondence models, m-JLF improved its original, in mean, over the 0.1% (77.72 ± 2.15 vs: 77.82 ± 2.12), whereas m-NLSS improved a 0.09% (74.42 ± 1.95 vs: 74.51 ± 1.94). On the other hand, filling the lesions before segmentation improved the original results on 0.21% for JLF(F) (77.93 ± 2.11) and 0.03% for NLSS(F) (74.45 ± 1.90).

Fig. 7 shows some qualitative segmentation results obtained with the original and the proposed correspondence models for both JLF and NLSS methods. From this figure we can observe that the original methods (JLF(L) and NLSS(L)) tend to segment the white matter lesions as part of the lateral ventricles, whereas the proposed non-local models (m-JLF and m-NLSS) as well as the original methods when segmenting the filled images (JLF(F) and NLSS(F)) tend to adjust better the edge between the two structures.

4. Discussion

In this work, we have presented an approach to solve for voxel/patch correspondences on intensity based multi-atlas label fusion segmentation when MRI visible lesions are present. We have presented the theory to apply this approach to two well-known label fusion strategies: Joint Label Fusion (JLF) and Non-local Spatial STAPLE (NLSS). Our proposal performs as well as the original strategy when segmenting healthy subjects, whereas the experiments performed showed that minimizes the effect of the lesions when segmenting lesioned brains, obtaining significantly better results than the original method.

Furthermore, when comparing our approach to the common lesion filling (Valverde et al., 2014) technique, the results obtained for the MS simulated database showed that, masking out the lesions with our approach leads to significantly better segmentation results than filling them before segmentation. On the other hand, when the analysis was performed on the MRBrainS18 dataset, the results showed that filling the lesions outperformed the segmentation result of our proposal for JLF method, whereas the proposal for NLSS achieved better results than lesion filling. However, the results on this second database have to be interpreted carefully, since the amount of analyzed data (only seven images) is small.

WM lesions are usually segmented on FLAIR, T2-w or PD sequences, where they appear larger than in T1-w and, sometimes, lesions that are

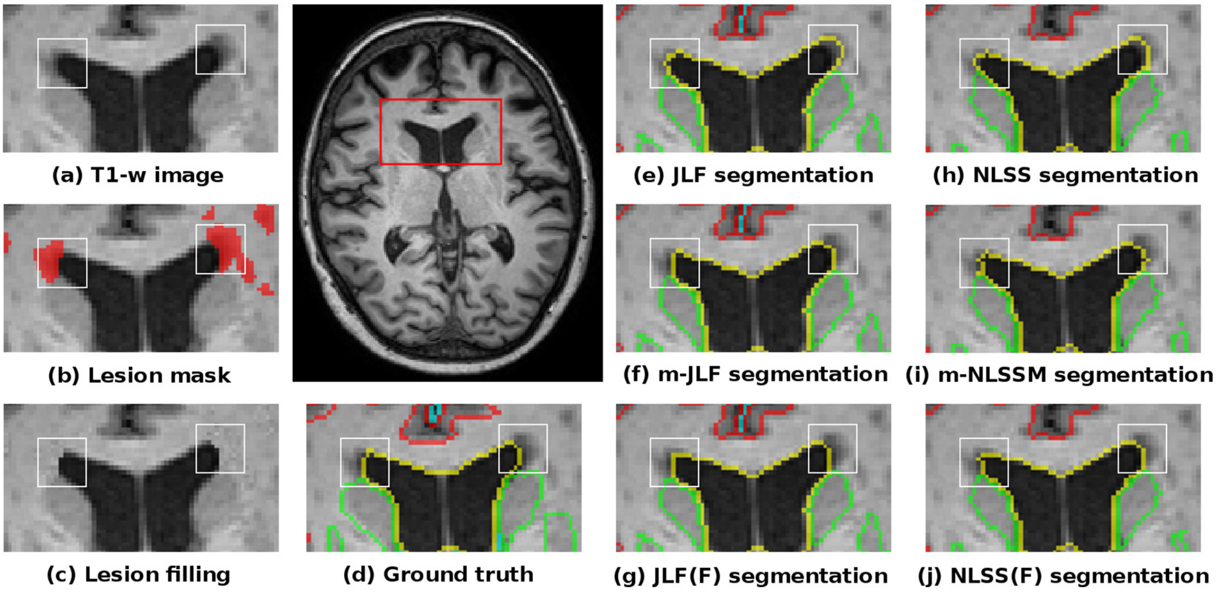


Fig. 7. Qualitative segmentation results of patient "5" from the MRBrainS18 database obtained with the analyzed methods. The image shows (a) the original T1-w image, (b) the superimposed lesion mask, (c) the lesion filled (Valverde et al., 2014) T1-w image, (d) the segmentation ground truth, and the segmentation results for the original T1-w image segmented with (e) JLF, (h) NLSS, (f) our proposal for JLF (m-JLF), and (i) our proposal for NLSS (m-NLSS); and the segmentation results for the lesion filled T1-w image segmented with (g) JLF, and (j) NLSS.

visible in those sequences are not perceptible in T1-w images. When using filling techniques, one has to be careful, since we can be corrupting image intensities that were correct, due to inaccurate segmentations. While it has been demonstrated to improve the results for segmentation techniques in which intensity distributions of different tissue classes are modelled, it has not been studied how it affects patch based segmentation strategies, in which patch intensities are independent to the global intensity distributions.

Differences in the trend of the results seen in both databases could be explained by the nature of the lesions. In the simulated dataset, lesions were located with no restriction on the affected structures, whereas in the seven images of the MRBrainS18 database, lesions did not affect any other structure than the WM. Because lesion filling techniques tend to fill all lesions with white matter-like intensities, these wrong intensities may be affecting the segmentation result of non-white matter lesions on the simulated database. We believe that, for this reason, lesion filling underperformed our proposal on the simulated database, whereas it worked better on the MRBrains18.

The small improvements seen on the whole brain segmentation are due to the lesion volumes being very small compared to the rest of the brain. Thus, a big improvement on small lesion areas will never have a big impact on the whole brain segmentation result. Furthermore, low values seen in the MRBrainS18 database for the whole brain mean DSC in all the analyzed strategies compared to the state of the art could be caused by the low resolution of the analyzed images and different labeling protocol of the atlases used with respect to the ground truth. However, the purpose of using this database was to compare the modified methods to their originals, which are equally affected by the low resolution and labeling protocols.

The analysis performed on the simulated data, showed that the lesion load has an important effect on the performance differences seen between the proposed methods and the rest of the strategies analyzed. In short, larger improvement was seen on the proposed strategies compared to the corresponding original methods when segmenting images with larger lesion loads. This makes us believe that either our proposal works better with larger lesion loads or either the original strategies are strongly affected by large lesion volumes. However, given that strong correlations were also found for healthy-lesions differences when applying the original segmentation strategies, we may conclude that large lesion loads have a bigger effect on the whole brain segmentation results of the original methods. This effect has shown to be mitigated with our proposal for NLSS, achieving similar results than the original strategy when segmenting the corresponding healthy subjects. On the other hand, although the results obtained with our proposal for JLF are better than those obtained with the original method, still do not reach the performance obtained for the healthy images.

In all the experiments performed, we used lesion masks that were manually annotated. However, expert annotated masks are not always available in practice, which generates a need of automatic methods able to come up with it. In this regard, automatic lesion segmentation has become a well-studied field, in the medical imaging community (Valverde et al., 2017; Roura et al., 2015; Tomas-Fernandez and Warfield, 2015). As a proof, several lesion segmentation challenges (Styner et al., 2008; Commowick et al., 2018; Carass et al., 2017; White Matter Hyperintensities Segmentation Challenge, 2018) have been conducted in recent years, in which successful strategies (Valverde et al., 2017), able to achieve segmentation results that are close to human expert inter-rater variability, have been presented. For this reason, we believe that feeding the proposed strategy with automatically segmented lesion masks, instead of manually annotated ones, would not have a significant impact on the final brain parcellation.

Integrated segmentation algorithms by which not only brain structures but also lesions can be segmented could be beneficial for the community (Amiri et al., 2018). The nature of multi-atlas strategies makes them flexible to label edits integration. For this reason, extending the theory of the analyzed methods to force the label decision inside the lesion mask would be straightforward, as we did in our previous work (González-Villà et al., 2018), where we extended the theory of NLSS to include a second mask that forced the label decision in case it was beforehand known and enabled seamless integration of manual edits.

The present study is not free of limitations. The most important one is the lack of public databases with both annotated lesions and brain structures ground truth. The limited amount of real patient images used in this study is not enough to extract statistically significant conclusions on real data, and thus, we extracted them from simulated images.

In conclusion, the results of this study show that the proposed correspondence models improve the segmentation results when MRI visible lesions are present, whereas they behave as the original method when they are not. When comparing to lesion filling, simulated data show that our proposals perform overall better, while on the seven-case real dataset, our correspondence model only outperforms lesion filling in one of the two analyzed methods (NLSS). Besides, using the proposed method we eliminate the preprocessing steps required by lesion filling and make the results more robust, since it is indifferent to the quality of filling method. Although our proposals obtain better segmentation results on the lesion areas (inside + around) and, on the whole brain, the experiments performed demonstrate that voxel/patch correspondences inside the lesion itself could be further improved. Thus, as a future work, we plan to improve our model, relaxing the one-to-one correspondence imposed inside the lesions and estimate them by means of interpolation of the neighboring ones (outside the lesions).

Acknowledgements

S. González-Villà holds a UdG-BRGR2015 grant. This work has been supported by "La Fundació la Marató de TV3", by Retos de Investigación TIN2014-55710-R and TIN2015-73563-JIN, by DPI2017-86696-R and by UdG mobility grant MOB17. This research was supported by NSF CAREER 1452485, NIH R01-EB017230, and the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06. We appreciate the NIH S10 Shared Instrumentation Grant 1S10OD020154-01 (Smith), Vanderbilt IDEAS grant (Holly-Bockelmann, Walker, Meliler, Palmeri, Weller), and ACCRE's Big Data TIPs grant from Vanderbilt University.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, National Science Foundation, or other sponsors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2019.101709.

References

Akhondi-Asl, A., Warfield, S.K., 2013. Simultaneous truth and performance level estimation through fusion of probabilistic segmentations. IEEE Trans. Med. Imaging 32 (10), 1840–1852.

Amiri, H., de Sitter, A., Bendfeldt, K., Battaglini, M., Wheeler-Kingshott, C.A.M. Gandini, Calabrese, M., Geurts, J.J.G., Rocca, M.A., Sastre-Garriga, J., Enzinger, C., de Stefano, N., Filippi, M., Rovira, À., Barkhof, F., Vrenken, H., MAGNIMS Study Group, 2018. Urgent challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI. NeuroImage: Clin. 19, 466–475.

Artaechevarría, X., Muñoz Barrutia, A., Ortiz-de Solorzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imag. 28 (8), 1266–1277.

Asman, A.J., Landman, B.A., 2013. Non-local statistical label fusion for multi-atlas segmentation. Med. Image Anal. 17 (2), 194–208.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41.

Batista, S., Zivadinov, R., Hoogs, M., Bergsland, N., Heininen-Brown, M., Dwyer, M.G., Weinstock-Guttman, B., Benedict, R.H.B., 2012. Basal ganglia, thalamus and neocortical atrophy predicting slowed cognitive processing in multiple sclerosis. J. Neurol. 259 (1), 139–146.

Battaglini, M., Jenkinson, M., De Stefano, N., 2012. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Hum. Brain Mapp. 33 (9), 2062–2071.

Buades, A., Coll, B., Morel, J.M., 2005. A non-local algorithm for image denoising. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2005). vol. 2. pp. 60–65.

Carass, A., et al., 2017. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage 148, 77–102.

Chard, D.T., Jackson, J.S., Miller, D.H., Wheeler-Kingshott, C.A.M., 2010. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. J. Magn. Reson. Imaging 32 (1), 223–228.

Commowick, O., et al., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci. Rep. 8 (1), 13650.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954.

Debernard, L., Melzer, T.R., Alla, S., Eagle, J., Van Stockum, S., Graham, C., Osborne, J.R., Dalrymple-Alford, J.C., Miller, D.H., Mason, D.F., 2015. Deep grey matter MRI abnormalities and cognitive function in relapsing-remitting multiple sclerosis. Psychiatry Res. Neuroimaging 234 (3), 352–361.

Eijlers, A.J.C., van Geest, Q., Dekker, I., Steenwijk, M.D., Meijer, K.A., Hulst, H.E., Barkhof, F., Uitdehaag, B.M.J., Schoonheim, M.M., Geurts, J.J.G., 2018. Predicting cognitive decline in multiple sclerosis: a 5-year follow-up study. Brain 141 (9), 2605–2618.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355.

Gelineau-Morel, R., Tomassini, V., Jenkinson, M., Johansen-Berg, H., Matthews, P.M., Palace, J., 2012. The effect of hypointense white matter lesions on automated gray matter segmentation in multiple sclerosis. Hum. Brain Mapp. 33 (12), 2802–2814.

González-Villà, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., Lladó, X., 2016. A review on brain structures segmentation in magnetic resonance imaging. Artif. Intell. Med. 73, 45–69.

González-Villà, S., Valverde, S., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, Ll., Rovira, À., Oliver, A., Lladó, X., 2017. Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation. NeuroImage: Clin. 15, 228–238.

González-Villà, S., Huo, Y., Oliver, A., Lladó, X., Landman, B.A., 2018. Multi-atlas parcellation in the presence of lesions: application to multiple sclerosis. In: Patch-based Techniques in Medical Imaging. Lecture Notes in Computer Science Vol. 11075. pp. 104–113 (Patch-MI 2018).

Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., Hammers, A., 2010. Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. NeuroImage 51 (1), 221–227.

Houtchens, M., Benedict, R., Killiany, R., Sharma, J., Jaisani, Z., Singh, B., Weinstock-Guttman, B., Guttmann, C., Bakshi, R., 2007. Thalamic atrophy and cognition in multiple sclerosis. Neurology 69 (12), 1213–1223.

Huo, Y., Asman, A.J., Plassard, A.J., Landman, B.A., 2017. Simultaneous total intracranial volume and posterior fossa volume estimation using multi-atlas label fusion. Hum. Brain Mapp. 38 (2), 599–616.

Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: a survey. Med. Image Anal. 24 (1), 205–219.

Iglesias, J., Sabuncu, M., Van Leemput, K., 2012. A generative model for probabilistic label fusion of multimodal data. Multimodal Brain Image Anal. 7509, 115–133.

Jacobsen, C., Hagemeier, J., Myhr, K.-M., Nyland, H., Lode, K., Bergsland, N., Ramasamy, D.P., Dalaker, T.O., Larsen, J.P., Farbu, E., et al., 2014. Brain atrophy and disability progression in multiple sclerosis patients: a 10-year follow-up study. J. Neurol., Neurosurg. Psychiatry 85, 1109–1115.

Klein, A., Dal Canton, T., Ghosh, S.S, Landman, B.A., Lee, J., Worth, A., 2010. Open labels: online feedback for a public resource of manually labeled brain images. In: 16th Annual Meeting for the Organization of Human Brain Mapping.

Kushibar, K., Valverde, S., González-Villà, S., Bernal, J., Cabezas, M., Oliver, A., Lladó, X., 2018. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. Med. Image Anal. 48, 177–186.

Landman, B.A., Warfield, S., 2012. MICCAI 2012 workshop on multiatlas labeling. In: MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling, Nice, France.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507.

MICCAI Grand Challenge on MR Brain Segmentation (MRBrainS18): http://mrbrains18.isi.uu.nl/, Granada, Spain, 2018.

MICCAI White Matter Hyperintensities Segmentation Challenge: http://wmh.isi.uu.nl/, Granada, Spain, 2018.

Nocentini, U., Bozzali, M., Spano, B., Cercignani, M., Serra, L., Basile, B., Mannu, R., Caltagirone, C., De Luca, J., 2014. Exploration of the relationships between regional grey matter atrophy and cognition in multiple sclerosis. Brain Imaging Behav. 8, 378–386.

Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N., 2001. Reconstructing a 3D structure from serial histological sections. Image Vis. Comput. 19 (1), 25–31.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56 (3), 907–922.

Prados, F., Cardoso, M.J., Kanber, B., Ciccarelli, O., Kapoor, R., Gandini Wheeler-Kingshott, C.A.M., Ourselin, S., 2016. A multi-timepoint modality-agnostic patch-based method for lesion filling in multiple sclerosis. NeuroImage 139, 376–384.

Roura, E., Oliver, A., Cabezas, M., Valverde, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, Ll., Rovira, À., Lladó, X., 2015. A toolbox for multiple sclerosis lesion segmentation. Neuroradiology 57 (10), 1031–1043.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S., 2008. 3D Segmentation in the clinic: a grand challenge II: MS lesion segmentation. MIDAS J.. http://hdl.handle.net/10380/1509.

Tillema, J.M., Hulst, H.E., Rocca, M.A., Vrenken, H., Steenwijk, M.D., Damjanovic, D., Enzinger, C., Ropele, S., Tedeschi, G., Gallo, A., Ciccarelli, O., Rovira, À., Montalban, X., de Stefano, N., Stromillo, M.L., Filippi, M., Barkhof, F., 2016. On behalf of the MAGNIMS Study Group, regional cortical thinning in multiple sclerosis and its relation with cognitive impairment: a multicenter study. Mult. Scler. J. 22 (7), 901–909.

Tomas-Fernandez, X., Warfield, S., 2015. A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion Segmentation. IEEE Trans. Med. Imag. 34 (6), 1349–1361.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imag. 29 (6) 1310–20.

Valverde, S., Oliver, A., Lladó, X., 2014. A white matter lesion-filling approach to improve brain tissue volume measurements. NeuroImage: Clin. 6, 86–92.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, Ll., Rovira, À., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage 155, 159–168.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35 (3), 611–623.

Warfield, S., Zou, K., Wells, W., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23 (7), 903–921.

Zivadinov, R., Bergsland, N., Dolezal, O., Hussein, S., Seidl, Z., Dwyer, M.G., Vaneckova, M., Krasensky, J., Potts, J.A., Kalincik, T., Havrdova, E., Horakova, D., 2013. Evolution of cortical and thalamus atrophy and disability progression in early relapsing-remitting MS during 5 years. AJNR Am. J. Neuroradiol. 34, 1931–1939.