Magnetic Resonance Imaging xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.mrijournal.com

Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images

Rita Simões [a,*], Christoph Mönninghoff [b], Martha Dlugaj [c], Christian Weimar [c], Isabel Wanke [b], Anne-Marie van Cappellen van Walsum [a,d], Cornelis Slump [a]

[a] MIRA Institute for Biomedical Technology and Technical Medicine, University of Twente, The Netherlands
[b] Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Germany
[c] Department of Neurology, University Hospital Essen, Germany
[d] Department of Anatomy, Radboud University Nijmegen Medical Centre, The Netherlands

ARTICLE INFO

Article history:
Received 20 August 2012
Revised 30 November 2012
Accepted 24 December 2012
Available online xxxx

Keywords:
White matter hyperintensities
Magnetic resonance imaging
Fluid-attenuation inversion recovery
Automatic segmentation

ABSTRACT

Magnetic Resonance (MR) white matter hyperintensities have been shown to predict an increased risk of developing cognitive decline. However, their actual role in the conversion to dementia is still not fully understood. Automatic segmentation methods can help in the screening and monitoring of Mild Cognitive Impairment patients who take part in large population-based studies. Most existing segmentation approaches use multimodal MR images. However, multiple acquisitions represent a limitation in terms of both patient comfort and computational complexity of the algorithms. In this work, we propose an automatic lesion segmentation method that uses only three-dimensional fluid-attenuation inversion recovery (FLAIR) images. We use a modified context-sensitive Gaussian mixture model to determine voxel class probabilities, followed by correction of FLAIR artifacts. We evaluate the method against the manual segmentation performed by an experienced neuroradiologist and compare the results with other unimodal segmentation approaches. Finally, we apply our method to the segmentation of multiple sclerosis lesions by using a publicly available benchmark dataset. Results show a similar performance to other state-of-the-art multimodal methods, as well as to the human rater.

1. Introduction

White matter hyperintensities (WMHs) are diffuse white matter abnormalities that appear with high intensities in T2-weighted magnetic resonance (MR) images. Although the pathogenesis of WMHs is not yet completely understood, these lesions are often associated with chronic cerebral ischemia, in particular with micro-vascular lesions originated by small vessel atherosclerosis [1]. They occur often in the elderly [2–5] and have been shown to predict an increased risk of stroke, cognitive decline and death [6].

The analysis of the real influence of WMHs on the development of dementia requires clinical studies involving large patient cohorts. Also, an accurate description of the location, shape and volume of the WMHs is necessary. Typically, WMHs are classified according to visual scales, such as the Scheltens scale or the Fazekas scale [7]. However, the results obtained by these visual scales are seldom comparable [8]. In addition, they have been shown to be little sensitive to clinical group differences [9]. Finally, they offer only a qualitative description of the WMHs, originating high intra- and inter-subject variabilities [10].

A quantitative and more reliable way of assessing WMHs is by manually determining the lesion volumes. However, for three-dimensional data this typically requires a slice-by-slice analysis, making the whole process cumbersome and time-consuming for the neuroradiologist. Also, the intra- and inter-rater variability have been reported to be high [11]. Clinical studies with hundreds of patients require, therefore, automated and robust segmentation methods.

Several methods have been proposed to automatically segment WMHs from MRI images, most of them using various types of MRI modalities [12–14]. The use of multimodal data presents several disadvantages. Namely, the acquired datasets must be coregistered, making the segmentations computationally intensive and more prone to errors. In particular, motion artifacts are seen frequently in the MRI data from elderly patients, who are often not able to lie still during the whole acquisition period. This represents a serious limitation for the registration algorithms and can negatively influence the outcomes [15,16].

Other methods have been specifically designed to segment multiple sclerosis (MS) lesions [17,18]. Although MS lesions look similar to vascular-related WMHs in MR images, the spatial

* Corresponding author.
E-mail address: a.r.lopessimoes@utwente.nl (R. Simões).

0730-725X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mri.2012.12.004

distribution of the lesions is often very different, with MS lesions occurring commonly in the corpus callosum and being symmetrically distributed in the brain, unlike the vascular WMHs [19].

WMHs are characterized by a larger T2 relaxation rate due to increased tissue water content and degradation of myelin [15]. Fast fluid-attenuated inversion-recovery (FLAIR) is a T2-weighted MR modality in which the cerebrospinal fluid (CSF) signal is attenuated. In FLAIR images, WMHs are characterized by an intensity range that only partially overlaps with that of normal brain regions, making this MRI modality well suited for lesion segmentation purposes [20].

Despite being the preferred imaging modality used by neuroradiologists to assess WMHs in the clinical setting, FLAIR has seldom been used alone in the automatic detection of these lesions [15,16].

In [15], the authors determined an optimal FLAIR intensity threshold to separate WMHs from normal brain tissue, based on the analysis of the image histograms on a training set. More recently, Ong et al. [21] have applied an outlier detection approach to find this optimal threshold, followed by a false positive correction step that uses the co-registered T1-weighted image. Similarly, de Boer et al. [14] determined the optimal intensity threshold on a training set and used the T1-weighted image to ensure the detected lesions were all within the white matter.

Applying a threshold allows only for crisp segmentation and does not account for the Partial Volume Effect (PVE) that is present in MR images. Having that in mind, Khademi et al. have proposed a segmentation method that allows for fuzzy segmentation and is based on a PVE model in FLAIR images [16].

In the methods described above, only the voxel intensity information is considered. However, it has been recognized that this makes methods highly sensitive to noise. In particular, boundary detection becomes problematic in noisy images. Furthermore, the common assumption that the voxel intensities are independent does not hold in practice. In reality, and intuitively, we can expect a certain voxel's value to be affected by those in its neighborhood [22,23].

In this work, we propose a WMH segmentation method that uses solely FLAIR images. It is based on a modified Gaussian mixture model (GMM) that incorporates neighborhood information, followed by a false positive correction step, where common FLAIR artifacts [24] are eliminated from the segmentation.

Gaussian mixture models (GMM), estimated by the expectation-maximization (EM) algorithm, have been widely used in brain image segmentation [25,26]. They provide a statistical description of the voxels' intensities and allow for fuzzy classification [27]. Because the traditional GMM-EM method is based only on intensity information, we use a modified GMM-EM method, initially proposed in [23], that considers additional contextual information. All initialization parameters are derived from the FLAIR image histogram.

We compare the performance of the proposed method with other unimodal approaches. For each method, the optimal parameters are determined using a training set that is retrieved randomly from our patient database. Evaluation is performed using the remaining patient datasets against the manual segmentation performed by an experienced neuroradiologist. Finally, we apply the method to a publicly available dataset of MS patients and compare the obtained performance results with those by multimodal segmentation methods and with the human expert.

2. Methods

Fig. 1 shows the general overview of our method.

The raw FLAIR image is first preprocessed to remove the skull and to correct for bias field inhomogeneities. Subsequently, a context-sensitive GMM is applied to the brain image and the resulting WMH probability class is thresholded. Finally, the existing FLAIR artifacts (located at the interface between the cerebrospinal fluid and the gray matter and inside the ventricles – red pixels in the last figure) are eliminated by morphological processing of the cerebrospinal fluid segmentation mask, resulting in the final segmentation of the WMH (blue pixels in Fig. 1D)). In the following subsections we will describe these steps in detail.

2.1. Gaussian mixture model

Fig. 2 shows the histograms of the FLAIR images of two patients. Two peaks can be easily distinguished: the one at lower intensities corresponds to cerebrospinal fluid voxels; the highest peak refers to white and gray matter voxels. Additionally, in Fig. 2B a low and broad peak is present at the right-end tail of the histogram. This peak is especially prominent in patients with a large lesion load and corresponds to WMH intensities.

We assume that the data can be modelled by a Gaussian mixture model (GMM) and that each voxel belongs to one of three distinct classes – cerebrospinal fluid (CSF), white and gray matter (WM/GM), or white matter hyperintensity (WMH). The probability density function (pdf) of a gray-level x can then be described by:

$$p(x|\pi,\mu,\sigma)=\sum_{k=1}^{3}\pi_k N(x|\mu_k,\sigma_k) \quad (1)$$

with $k=1,2,3$ respectively corresponding to the CSF, WM/GM and WMH classes. Each Gaussian component N is characterized by a mixing weight π_k, a mean value μ_k and a standard deviation σ_k.

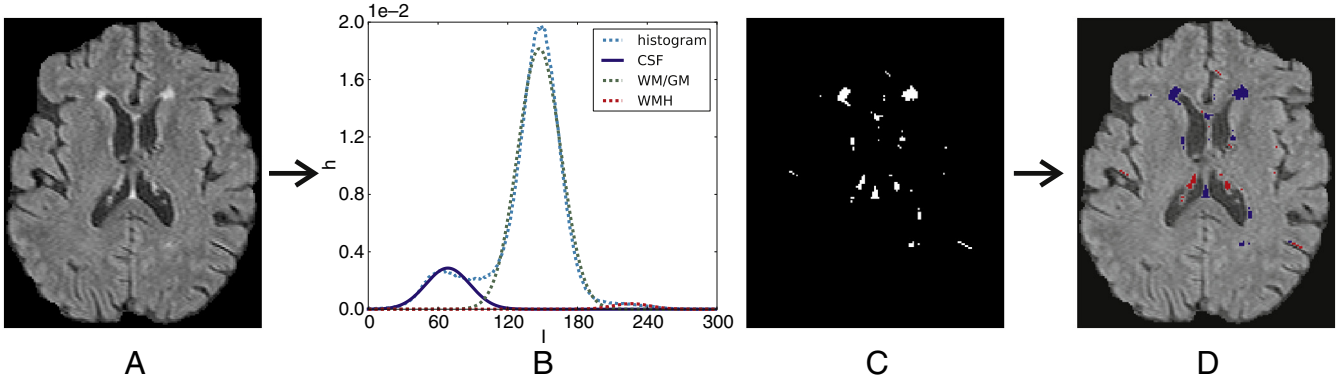

Fig. 1. General overview of the segmentation method: A) we take the histogram of the skull-stripped and bias field-corrected FLAIR image and B) fit a 3-class context-sensitive GMM to it. Subsequently, we apply a threshold to the WMH class probability map, obtaining C) an initial lesion segmentation. Finally, we apply a post-processing step that corrects for artifacts in the initial segmentation; D) in red, the removed artifacts; in blue, the final segmentation. (For interpretation of the color references, we refer the reader to the web version of this article.)

Please cite this article as: Simões R, et al, Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images, Magn Reson Imaging (2013), http://dx.doi.org/10.1016/j.mri.2012.12.004

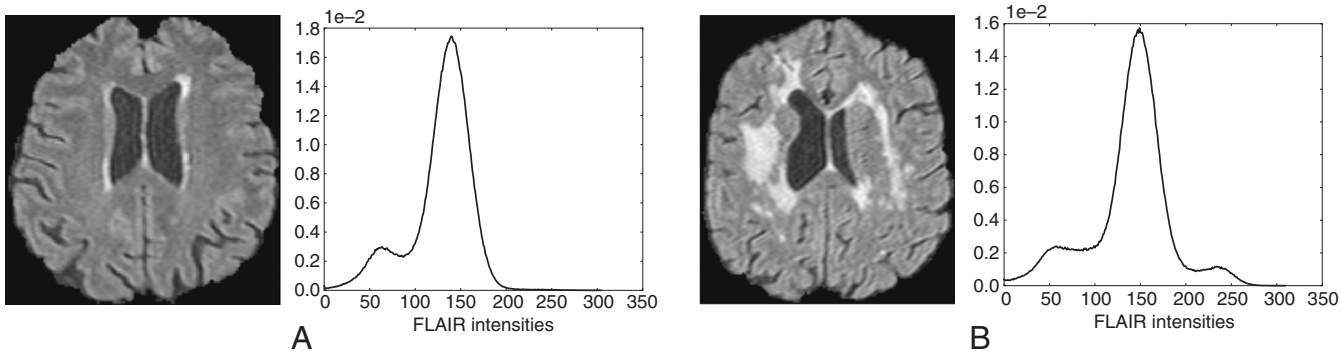

Fig. 2. FLAIR image and respective histogram from a patient: A) with a low WMH load; B) with a high WMH load.

We use the expectation-maximization (EM) algorithm to find these parameters.

2.1.1. Traditional expectation-maximization

The EM algorithm is an iterative procedure that maximizes the log-likelihood of the parameters [28,29]. It alternates between two consecutive steps: the Expectation (E)-step and the Maximization (M)-step. In the E-step, the parameters at the current iteration are used to compute the log-likelihood. In the M-step, the computed log-likelihood is maximized to determine the new parameters.

Assuming that the data, $\mathbf{X}=(x_1,\ldots,x_N)$, are independent and identically distributed variables, the log-likelihood of the parameters given the data is defined as:

$$\ell(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\sigma}|\mathbf{X}) = log\prod_{n=1}^{N} p(x_n|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\sigma}) = \sum_{n=1}^{N} log[p(x_n|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\sigma})] \tag{2}$$

The M-step parameter estimates are derived by maximizing Eq. (2):

$$\begin{aligned}\mu_k^{(i+1)} &= \frac{1}{N}\sum_{n=1}^{N} x_n T_{k,n}^{(i)}\\ \sigma_k^{(i+1)} &= \sqrt{\frac{\sum_{n=1}^{N}\left(x_n-\mu_k^{(i+1)}\right)^2 T_{k,n}^{(i)}}{\sum_{n=1}^{N} T_{k,n}^{(i)}}}\\ \pi_k^{(i+1)} &= \frac{1}{N}\sum_{n=1}^{N} T_{k,n}^{(i)}\end{aligned} \tag{3}$$

where $T_{k,n}^{(i)}$ is determined at the E-step by:

$$T_{k,n}^{(i)} = \frac{\pi_k^{(i)} N(x_n|\mu_k^{(i)},\sigma_k^{(i)})}{p\left(x_n|\boldsymbol{\pi}^{(i)},\boldsymbol{\mu}^{(i)},\boldsymbol{\sigma}^{(i)}\right)} \tag{4}$$

The initial parameters are computed from the histogram as follows: $\mu_{WM/GM}^{(0)}$ and $\mu_{CSF}^{(0)}$ correspond to the first and second highest peaks in the histogram, respectively; $\mu_{WMH}^{(0)}$ is taken as the local histogram maximum between $\mu_{WM/GM}^{(0)}$ and the maximum intensity (if no local maxima are found, we take this value as the average between $\mu_{WM/GM}^{(0)}$ and the maximum intensity); all standard deviations are initialized with the same value: the standard deviation of the voxel intensities in the CSF class (with the threshold for this class being the local minimum between $\mu_{WM/GM}^{(0)}$ and $\mu_{CSF}^{(0)}$); finally, the initial class weights are selected based on the relative ratios between $\mu_{WM/GM}^{(0)}$, $\mu_{CSF}^{(0)}$ and $\mu_{WMH}^{(0)}$. These weights can take values in the interval [0,1]. This means that if there are no lesions in the brain the outcome will be a two-class segmentation (CSF and WM/GM).

The algorithm has converged when the absolute normalized difference between the log-likelihood values at two consecutive iterations is lower than tolerance $T=10^{-3}$.

Although it may be sufficient to obtain a first rough approximation of the voxels' statistical distributions, the traditional GMM-EM algorithm has the disadvantage of taking only intensity information into account. We therefore apply a previously proposed [23] adaptation to the E-step. The difference between the performance of the normal and the modified GMM-EM approaches is particularly significant in images with low WMH loads, as we will show in Section 3.

2.1.2. Context-sensitive expectation-maximization

In [23], the authors introduced contextual information into the traditional GMM-EM method as follows. At each iteration, the posterior probability (Eq. (4)) is substituted by:

$$T_{k,n}^{(i)CC} = \frac{\pi_k^{(i)} C_{k,n}^{(i)} N(x_n|\mu_k^{(i)},\sigma_k^{(i)})}{p\left(x_n|\boldsymbol{\pi}^{(i)},\boldsymbol{\mu}^{(i)},\boldsymbol{\sigma}^{(i)}\right),} \tag{5}$$

which incorporates a context-sensitive penalty term $C_{k,n}^{(i)}$. This term imposes that, at each iteration, the probability that a voxel belongs to class k depends not only on the voxel's intensity, but also on its neighbors' current class probabilities. We define the penalty term as follows:

$$C_{k,n}^{(i)} = \Phi\left\{I_k^{(i)}\right\}(x_n) \tag{6}$$

with $I_k{}^{(i)}$ being the membership image which, at each brain voxel x_n, represents the probability that the voxel belongs to class k. $\Phi\{\cdot\}$ represents the filter used to take the voxel's neighborhood into account.

We initialize the context-sensitive (CS-) EM method with the parameters that result from applying the traditional GMM-EM method to the dataset. After convergence, we apply thresholds t_{WMH} and t_{CSF} to the resulting WMH and CSF membership images, respectively.

2.2. False positive correction

After applying the threshold to the WMH probability map, we still obtain some false positives – voxels that are initially considered to be lesions but are in reality FLAIR artifacts. We apply a postprocessing step that consists of eliminating these voxels from the segmentation.

A common location of false positives is in the interface between the CSF and the cortical gray matter. To eliminate these voxels from our initial segmentation, we use the CSF mask obtained after thresholding the CSF class membership image that results from the segmentation method described above. We perform binary dilation of

Please cite this article as: Simões R, et al, Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images, Magn Reson Imaging (2013), http://dx.doi.org/10.1016/j.mri.2012.12.004

this mask with a three-dimensional cubic structure with size $S \times S \times S$. We mask our first WMH segmentation obtained after applying the EM method with the dilated CSF mask.

Other hyperintense voxels, resulting from flow artifacts (located mainly in the ventricular system) [24] are also eliminated in this step by morphologically "closing the holes" [30] in the dilated CSF mask.

Finally, and because the lesion voxels adjacent to the ventricles are also eliminated after this step, we perform binary propagation [30] to the initial WMH segmentation in order to recover these wrongly eliminated voxels.

2.3. Evaluation metrics

To evaluate the method, we compare our results with the manual segmentation provided by an experienced neuroradiologist. We use the following metrics for comparison: dice similarity coefficient (DSC), overlap fraction (OF) and extra fraction (EF) [12]:

$$\mathrm{DSC} = \frac{2 \times \#\mathrm{TP}}{\#\mathrm{AS} + \#\mathrm{GT}} \tag{7}$$

$$\mathrm{OF} = \frac{\#\mathrm{TP}}{\#\mathrm{GT}} \tag{8}$$

$$\mathrm{EF} = \frac{\#\mathrm{FP}}{\#\mathrm{GT}} \tag{9}$$

with TP and FP being the true and the false positives, respectively, AS the automatic segmentation and GT the ground truth provided by the expert.

Because the lesion load (LL) is often an important measure in clinical studies, we finally determine the correlation coefficient between the obtained LL values with those from the manual segmentations.

3. Experiments and results

3.1. Data

Forty datasets were retrieved from a large database of a cognition study with MCI and control subjects carried out at the University Hospital of Essen, Germany. From these 40 subjects, 15 correspond to stable normal controls, 14 to stable amnestic-MCI subjects, 8 to MCI subjects who have progressed to dementia and 3 to normal subjects who have declined to amnestic-MCI. The age of the subjects is 74.7 ± 4.3 (range 62–82).

Three-dimensional isotropic FLAIR images are utilized in this study (1.5 T Siemens Avanto, Germany); TR = 6000ms; TE = 308ms; TI = 2200ms; voxel size = $1\mathrm{mm}^3$). We apply the following preprocessing steps to the raw FLAIR images:

- brain extraction using BET (FMRIB's brain extraction tool, http://fsl.fmrib.ox.ac.uk/fsl/bet2/) [31];
- bias field correction using FAST (FMRIB's automated segmentation tool, http://fsl.fmrib.ox.ac.uk/fsl/fast4/) [32].

For the evaluation of the method, we use as the ground truth the manual segmentation performed on all 40 FLAIR images by an experienced neuroradiologist using 3D Slicer (www.slicer.org).

The WMH lesion loads are typically divided into three groups: low LL (less than 10 cm^3), medium LL (between 10 and 30 cm^3) and large LL (more than 30 cm^3). After manual labeling, we obtain 18 datasets that are considered to have low LL, 13 datasets with medium LL and only 9 datasets with high LL.

We randomly split our dataset into 30% training and 70% test. That is, we use 12 datasets (four of each LL category) to learn our method's optimal parameters, while the remaining 28 datasets are used as a test set for an independent evaluation of the method.

3.2. Selection of the optimal parameters

3.2.1. First WMH segmentation

Two parameters influence the outcome of the first step of the segmentation method: the threshold which is applied to the WMH class membership to obtain a crisp segmentation and the neighborhood filter type and size ($\Phi\{\cdot\}$ in Eq. (6)).

We use the training set to find the optimal joint parameters. Fig. 3 shows the joint parameter analysis - on the horizontal axes, we plot the threshold values and the filter types. The z-direction shows the corresponding DSC values averaged across the training set. We observe that the DSC index is most sensitive to t_{WMH}, with very little variability across the various neighborhood types. At the optimal threshold (10^{-5}), the average DSC values vary less than 5% across the considered neighborhood types. The exception is the case where no neighborhood information is used. This approach, as we will also show in Section 3, performs considerably worse than the contextual methods.

We then select the first neighborhood (the $3 \times 3 \times 3$ mean filter) for further processing.

For this neighborhood filter, we plot each subject's DSC curve and the average across all training set subjects. The broader curve, with a lower optimal threshold, corresponds to a low LL dataset. On the other hand, the datasets with higher LL have higher optimal thresholds (Fig. 4).

3.2.2. False positive correction

Finally, we correct for the presence of FLAIR artifacts. This step takes also two parameters: the threshold of the CSF membership image and the size of the structuring element used to create the FP mask from the CSF segmentation (Fig. 5).

Similarly to what was done in the previous subsection, we analyze the joint parameters and select the combination that gives the best results on the training set. In this case, we fix the WMH threshold to 10^{-5} and the neighborhood filter to the mean in a $3 \times 3 \times 3$ local window.

As in the previous case, the CSF threshold has the most influence on the DSC value, with the best performance being achieved at $t_{CSF=10^{-2}}$ and with a structuring element size of $5 \times 5 \times 5$. However, for thresholds greater than 10^{-5}, the mean DSC values also vary less than 5%, regardless of the structuring element size.

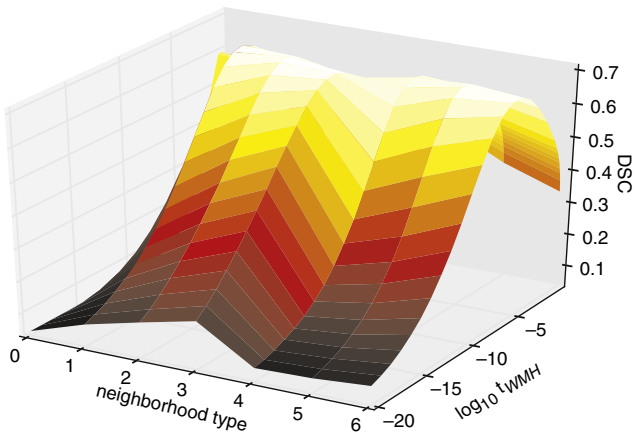

Fig. 3. Search for the optimal parameters of the first step of the segmentation method. The neighborhood filter types are the following: 0: no neighborhood information (traditional GMM-EMM method); 1: mean filter with size $3 \times 3 \times 3$; 2: mean filter with size $5 \times 5 \times 5$; 3: mean filter with size $7 \times 7 \times 7$; 4: isotropic Gaussian filter with $\sigma = 0.7$; 5: isotropic Gaussian filter with $\sigma = 1.5$; 6: isotropic Gaussian filter with $\sigma = 2$.

Please cite this article as: Simões R, et al, Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images, Magn Reson Imaging (2013), http://dx.doi.org/10.1016/j.mri.2012.12.004

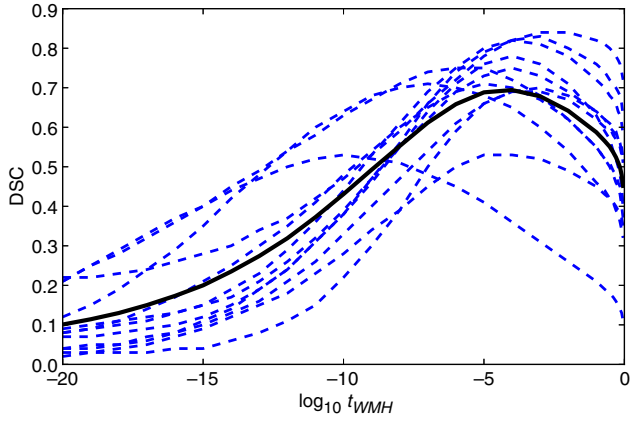

Fig. 4. DSC values for all patients in the training set, using a mean filter with size 3×3×3. The average DSC corresponds to the thicker black line.

3.3. Evaluation on the test set

We evaluate the method against the manual segmentation on the remaining 28 datasets. Table 1 shows the final DSC, EF and OF values, per lesion load, in the test set, as well as the resulting lesion loads in the automatic segmentation (AS) and ground truth (GT).

The average DSC values are 0.51, 0.70 and 0.84 for the low LL, medium LL and high LL, respectively. DSC values above 0.70 are considered to represent a very good agreement between segmentations [33]. The lower similarity values for the low lesion loads are to be expected, since errors in the segmentation have a greater impact on the similarity score when the lesion load is lower. This has also been reported in previous studies [13,12,34].

In Table 1 we can observe a systematic underestimation of the lesion loads in the low LL cases and an overestimation for the high LL datasets. The latter can be visualized on the first example of Fig. 6C) and is also expressed on the relatively high EF values for the high LL datasets (Table 2).

Finally, we plot the automatically obtained LL against the ground truth LL (Fig. 7). The obtained correlation coefficient ($R = 0.9966$) indicates a strong correlation between the two measurements.

3.4. Comparison with other unimodal approaches

To further evaluate the performance of the proposed method, we compare it with four other segmentation approaches which use only FLAIR images. For each of these approaches, we search for the optimal parameters in the training set and evaluate them in the test set. The exception is the first method, in which a threshold is applied to the FLAIR intensities (intensity thresholding, IT). In this case, because the goal is not to evaluate any specific method that searches for an optimal threshold, we take the optimal threshold value for each subject individually. This way we ensure that the obtained DSC is the highest that can be achieved with such approach.

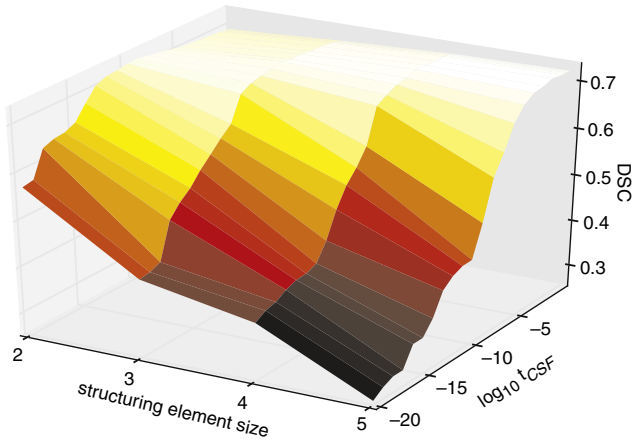

Fig. 5. Search for the optimal parameters in the FP correction step.

Table 1
Performance measures for the 28 patients in the test set.

LL category	Subject ID	DSC	EF	OF	AS (cm^3)	GT (cm^3)
Low	1	0.79	0.19	0.78	5.787	5.973
	2	0.44	0.10	0.31	2.048	4.887
	3	0.70	0.21	0.65	3.781	4.380
	4	0.60	0.11	0.47	3.353	5.742
	5	0.64	0.14	0.54	4.547	6.708
	6	0.21	0.01	0.12	1.177	9.168
	7	0.25	0.03	0.15	0.375	2.160
	8	0.37	0.01	0.23	1.170	4.896
	9	0.70	0.30	0.69	8.100	8.160
	10	0.49	0.03	0.34	1.322	3.583
	11	0.37	0.01	0.23	0.919	3.801
	12	0.40	0.05	0.26	0.698	2.226
	13	0.67	0.15	0.53	4.945	7.062
	14	0.51	0.27	0.43	0.497	0.714
Medium	15	0.72	0.28	0.72	10.291	10.267
	16	0.63	0.15	0.53	8.113	11.917
	17	0.71	0.11	0.61	7.541	10.328
	18	0.74	0.17	0.69	11.471	13.475
	19	0.70	0.28	0.69	11.108	11.497
	20	0.39	0.09	0.26	3.963	11.375
	21	0.77	0.18	0.74	10.877	11.801
	22	0.83	0.17	0.84	13.403	13.313
	23	0.80	0.20	0.79	12.999	13.109
High	24	0.85	0.29	0.96	155.220	124.177
	25	0.86	0.18	0.89	40.293	37.559
	26	0.84	0.23	0.89	56.411	50.679
	27	0.81	0.33	0.90	73.326	59.881
	28	0.83	0.20	0.84	47.226	45.177

The second comparison is with the traditional GMM, with parameters determined by EM (simple GMM, sGMM). This method, unlike the first one, yields a fuzzy segmentation. However, it is also based only on intensity information.

The PVA model introduced in [16] is used for the third comparison. Similarly to the GMM-EM method, its output is a fuzzy segmentation that does not consider any contextual information. However, this method is based not only on the image intensities but also on the gradient magnitudes.

Finally, we compare our approach with an analogous segmentation method – Fuzzy C-Means (FCM), modified in [35] to incorporate neighborhood information (cFCM). Unlike the GMM-EM approach we use here, this method does not assume any probabilistic model for the voxel intensities.

For the proposed method, we show the results obtained after the initial WMH segmentation ("proposed (first)") and after FP correction ("proposed (final)").

The results are shown in Table 2. Fig. 8 shows the average DSC values obtained for the three LL categories.

We observe that the proposed method performs significantly better than the first three context-free approaches. A slight improvement is also observed with respect to the contextual FCM method. However, the FCM method seems to perform considerably less robustly in very low LL cases – particularly with respect to the EF measure.

In all cases, the DSC values are lower for the low LL cases. This is expectable, since errors in these measurements tend to have a larger impact on the final similarity score. Also, the variability is larger in these cases, indicating a lower robustness of the methods.

A criticism that can be made to model-based segmentation methods, such as GMM, is that, for low LL, there may not be enough

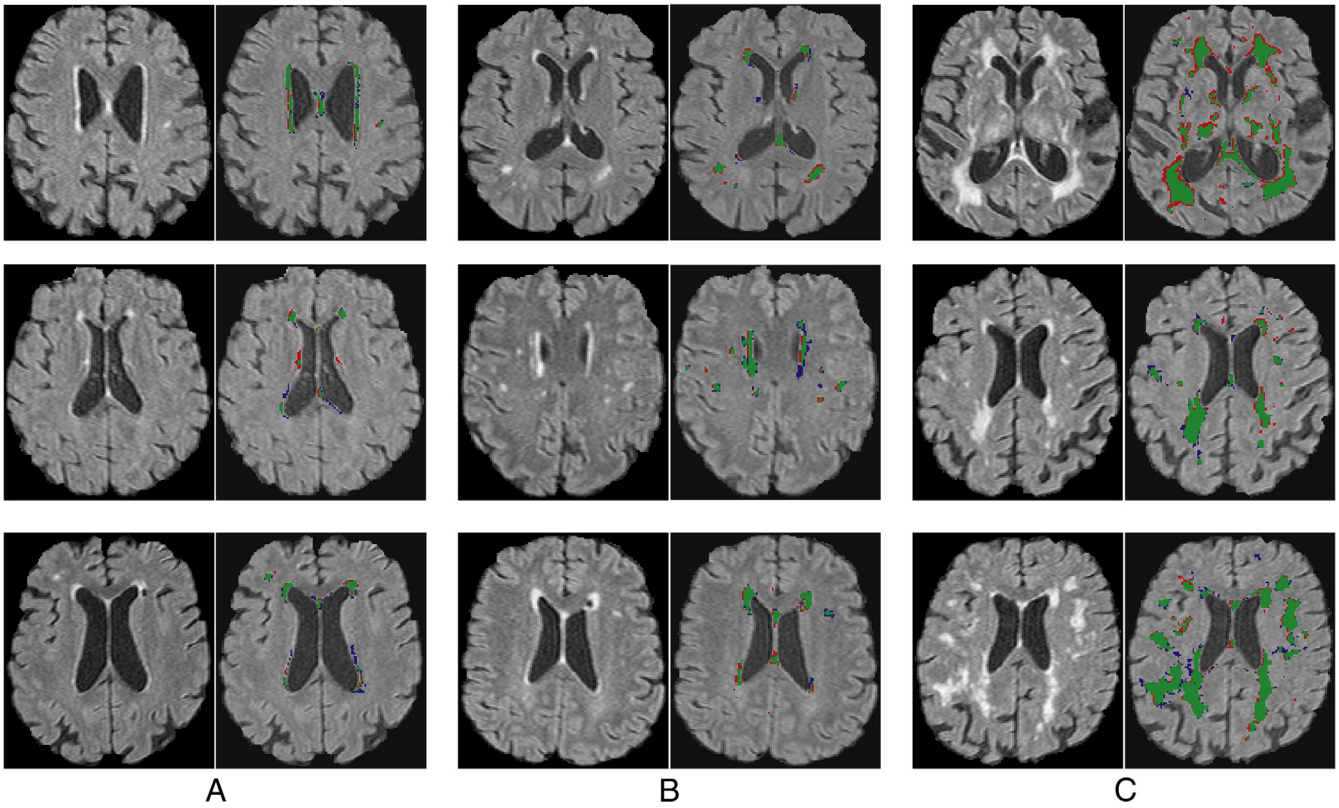

Fig. 6. Segmentation examples for the three lesion load categories: A) low, B) medium and C) high. Green: True positives; Red: False positives; Blue: False negatives. (For interpretation of the color references, we refer the reader to the web version of this article.)

lesion voxels to accurately derive the model's parameters [36]. Although this may be true for the simple GMM (with an average DSC of 0.38 in the low LL case), the problem seems to be overcome by considering contextual information, as in the proposed method, which outperforms the model-free contextual approach (cFCM).

It is worth noting that the performance of the first approach is highly overestimated, since for each patient we take the optimal DSC value (without recurring to a training set). However, results also show that the two other context-free approaches (simple GMM and PVA model) have a similar performance, indicating that adding neighborhood information not only improves the similarity scores but also seems to be a determinant factor in the methods' performance.

Finally, a paired sample t-test on the results of all subjects on the test set shows a significant improvement ($p<0.05$) on the DSC metric with the first step of the proposed method with respect to all other approaches. Furthermore, the second step also accounts for a significant improvement of the performance metrics with respect to the first step, indicating the importance of the artifact elimination step in the segmentation.

Table 3 shows the correlation coefficients between each segmentation approach and the manual measurements.

3.5. Robustness to the initialization parameters

A final evaluation is performed by varying the parameters that initialize the first EM procedure. Converging to local minima is a well-known limitation of the EM method [37]. Therefore, we evaluate the robustness of the proposed method to variations in the three parameters of the Gaussian that describes the WMH class distribution: the mean value μ_{WMH}, the standard deviation σ_{WMH} and the weight π_{WMH}, determined as described in Section 2. Again, we use the dice similarity coefficient as a performance measure.

The results are shown in Fig. 9.

In the horizontal axis we show the parameter values used for comparison. During the evaluation of each parameter, the others remained constant and equal to the values automatically determined by the method, as described in Section 2. The values $\{p_{-2},p_{-1},p,p_1,p_2\}$ correspond to $\{\mu_{WMH}-20,\mu_{WMH}-10,\mu_{WMH},\mu_{WMH}+10,\mu_{WMH}+20\}$ for the WMH mean, to $\{\sigma_{WMH}-10,\sigma_{WMH}-5,\sigma_{WMH},a_{WMH}+5,\sigma_{WMH}+10\}$ for the standard deviation and to $\{\pi_{WMH}/10,\pi_{WMH}/5,\pi_{WMH},\pi_{WMH}\times5,\pi_{WMH}\times10\}$ for the WMH weight.

Even though we select a large range of parameter values, the DSC values remain approximately constant. For the mean value, the variability of the DSC scores (ratio between the range and the

Table 2
Performance average (standard deviation) values for four different approaches and for the proposed method (first step and after FP correction).

Methods	Low LL			Medium LL			High LL		
	DSC	EF	OF	DSC	EF	OF	DSC	EF	OF
IT	0.41 (0.11)	0.40 (0.11)	0.36 (0.10)	0.57 (0.13)	0.31 (0.17)	0.51 (0.12)	0.75 (0.05)	0.23 (0.07)	0.73 (0.07)
sGMM	0.38 (0.15)	1.0 (2.73)	0.34 (0.13)	0.56 (0.14)	0.14 (0.04)	0.46 (0.14)	0.75 (0.05)	0.15 (0.04)	0.70 (0.08)
PVA	0.40 (0.13)	0.40 (0.69)	0.32 (0.11)	0.56 (0.15)	0.11 (0.05)	0.45 (0.15)	0.75 (0.05)	0.19 (0.07)	0.71 (0.07)
cFCM	0.42 (0.19)	1.62 (2.99)	0.47 (0.12)	0.63 (0.13)	0.11 (0.06)	0.52 (0.13)	0.81 (0.04)	0.06 (0.02)	0.73 (0.06)
prop. (first)	0.50 (0.13)	0.36 (0.42)	0.46 (0.19)	0.66 (0.12)	0.34 (0.13)	0.67 (0.16)	0.79 (0.02)	0.37 (0.05)	0.90 (0.04)
prop. (final)	0.51 (0.17)	0.11 (0.09)	0.41 (0.20)	0.70 (0.12)	0.18 (0.07)	0.65 (0.16)	0.84 (0.02)	0.25 (0.06)	0.90 (0.04)

Please cite this article as: Simões R, et al, Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images, Magn Reson Imaging (2013), http://dx.doi.org/10.1016/j.mri.2012.12.004

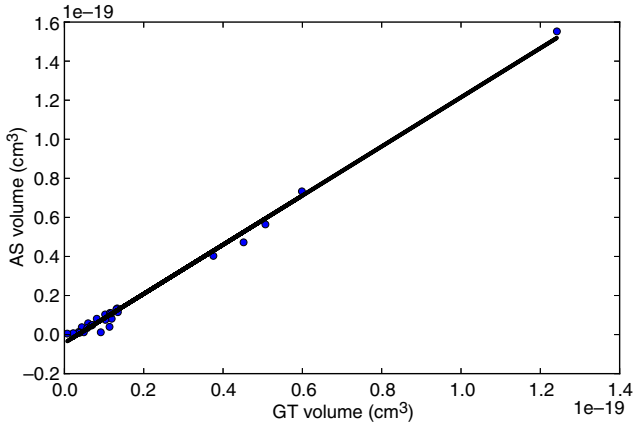

Fig. 7. Ground Truth (GT) and Automatic Segmentation (AS) lesion loads and the fitted linear regression line ($y = 1.28x - 4.19$).

maximum value) is 0.7%. For the standard deviation and the class weight the variabilities are 1.1% and 0.9%, respectively.

3.6. Application in the segmentation of Multiple Sclerosis (MS) lesions

To show the applicability of our method in a different neurological disease, we use a benchmark dataset made available by the Medical Image Computing and Computer Aided Intervention Society's (MICCAI's) MS Lesion Segmentation Challenge 2008 (http://www.ia.unc.edu/MSseg/). The data consist of 23 FLAIR images acquired at the Children's Hospital Boston (CHB) and at the University of North Carolina (UNC), with a dimension of $512 \times 512 \times 512$ voxels, resliced at 0.5 mm × 0.5 mm × 0.5 mm resolution using cubic spline interpolation.

The four error metrics used to evaluate the methods' performance are the following: relative absolute volume difference, average symmetric surface distance, true positive rate and false positive rate. The results were scaled to a range such that a score of 90 points is comparable to the performance of a human expert. For further details on the design of the Challenge, we refer the reader to [38].

The results for all subjects are shown in Table 4.

Our method obtained an overall score of 82.0055 (http://www.ia.unc.edu/MSseg/results_table.php), outperforming other WML segmentation methods in the literature [12,17,21] and reaching similar performance to other methods [18]. It is worth noting that our method performs less than 2 score points worse than the method that is currently at the first position of the Challenge. Also, all other participating methods require at least two MR modalities, while ours uses only FLAIR image data. Finally, some of the methods assume a priori knowledge about the spatial distribution of the MS lesions [18,39]. In contrast, our method has a more general applicability since it uses only intensity information.

Table 3
Correlation coefficients between the lesion loads determined by the automatic and the manual measurements.

IT	sGMM	PVA	cFCM	prop. (first)	prop. (final)
0.9969	0.9901	0.9927	0.9862	0.9957	0.9966

4. Conclusion

In this work, we present a method to automatically segment WMHs using only 3D FLAIR images. It uses a context-sensitive Gaussian mixture model to obtain class probabilities, followed by crisp segmentation and artifact correction. Unlike the majority of the existing approaches (to the best of our knowledge), our method requires no additional MRI modalities nor atlases, thereby shortening the acquisition time, avoiding the need for co-registrations and allowing for near real-time analysis. Results show that the method is suitable for a robust segmentation of WMHs of various loads. Also, a comparison with other segmentation approaches indicates the usefulness of, on the one hand, incorporating contextual information and, on the other hand, considering a model for the lesions (instead of a model-free approach such as FCM). The significant improvements observed on the performance measures after applying the FP correction step (with respect to the initial segmentation) suggest the efficacy of the simple CSF-based mask we have used, without needing additional MR modalities.

We have also demonstrated the applicability of our method in the detection of other lesion types, namely Multiple Sclerosis lesions. In particular, the results on a benchmark dataset show that our method performs comparably to other state-of-the art multimodal methods, with the difference that ours does not need any MR modalities other than FLAIR and does not make assumptions about the spatial distribution of the lesions, therefore having a wider applicability. The final score obtained in this evaluation indicates that the method performs close to the human observer. Because we make no assumptions about the lesion spatial distribution, we believe that this method can be applied to other neurological diseases that have a similar appearance in FLAIR images. Examples include subcortical arteriosclerotic encephalopathy and brain tumors.

A possible drawback of our method is that it requires two preprocessing steps: brain extraction and bias field correction. This is a consequence of the algorithm being fully intensity-based and relying on the brain image histogram. An extension can be considered in which the bias field correction is incorporated into the segmentation framework. Also, a study on the robustness of the method to the presence of field inhomogeneities and wrong brain extractions should

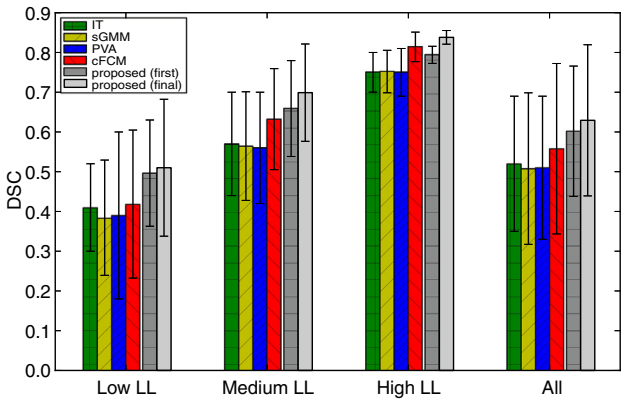

Fig. 8. Average DSC values for the six compared methods, separated by lesion load.

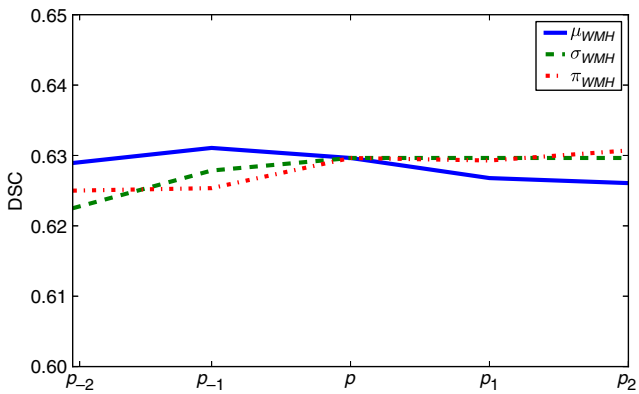

Fig. 9. Variation of the average DSC values with varying initialization parameters.

Table 4
Summary of the performance measures for the 23 patients in the MICCAI Challenge test set.

Ground truth	UNC rater								CHB rater								
All datasets	Volume Diff.		Avg. Dist.		True Pos.		False Pos.		Volume Diff.		Avg. Dist.		True Pos.		False Pos.		Total
	[%]	Score	[mm]	Score	[%]	Score	[%]	Score	[%]	Score	[mm]	Score	[%]	Score	[%]	Score	
Range	4.5–100	85–99	1.2–128	0–97	0–68.4	51–90	0–52	78–100	11.7–142.5	79–98	1.2–128	0–97	0–81.5	51–98	0–60.8	73–100	59–93
Std dev	37.7	5.6	42.8	36.5	21.3	12.2	20.6	9.2	35.3	5.3	42.6	35.4	23.4	13.4	20.0	9.3	11.9
Average	47.4	92.9	24.6	71.3	30.4	68.7	21.3	92.8	62.9	90.8	23.5	73.2	35.2	71.5	18.2	94.6	82.0

be carried out. Finally, it is worth pointing out that we do not perform any registration step, which is typically more time-consuming than the two steps required by our method (particularly when using multimodal data).

Ultimately, we expect that this method can become a useful tool in the evaluation of WMHs in the large patient cohorts required by population-based studies.

References

[1] Pantoni L, Garcia JH. The significance of cerebral white matter abnormalities 100 years after Binswanger's report. A review. Stroke 1995;26:1293–301.
[2] de Leeuw F, de Groot JC, Achten E, Oudkerk M, Ramos L, Heijboer R, et al. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the Rotterdam Scan Study. J Neurol Neurosurg Psychiatry 2001;1:9–14.
[3] Launer LJ, Berger K, Breteler MMB, Dufouil C, Fuhrer R, Giampaoli S, et al. Regional variability in the prevalence of cerebral white matter lesions: an MRI study in 9 European countries (CASCADE). Neuroepidemiology 2006;26:23–9.
[4] Ylikoski A, Erkinjuntti T, Raininko R, Sarna S, Sulkava R, Tilvis R. White matter hyperintensities on MRI in the neurologically nondiseased elderly. Analysis of cohorts of consecutive subjects aged 55 to 85 years living at home. Stroke 1995;26:1171–7.
[5] Hopkins RO, Beck CJ, Burnett DL, Weaver LK, Victoroff J, Bigler ED. Prevalence of white matter hyperintensities in a young healthy population. J Neuroimaging 2006;3:243–51.
[6] Debette S, Markus HS. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ 2010;341:c3666.
[7] Scheltens P, Erkinjunti T, Leys D, Wahlund LO, Inzitari D, del Ser T, et al. White matter changes on CT and MRI: an overview of visual rating scales. European task force on age-related white matter changes. Eur Neurol 1998;39:80–9.
[8] Mäntylä R, Erkinjuntti T, Salonen O, Aronen H, Peltonen T, Pohjasvaara T, et al. Variable agreement between visual rating scales for white matter hyperintensities on MRI: comparison of 13 rating scales in a poststroke cohort. Stroke 1997;28:1614–23.
[9] van Straaten ECW, Fazekas F, Rostrup E, Scheltens P, Schmidt R, Pantoni L, et al. Impact of white matter hyperintensities scoring method on correlations with clinical data: the LADIS study. Stroke 2006;37:836–40.
[10] van den Heuvel DMJ, ten Dam VH, de Craen AJM, Admiraal-Behloul F, van Es ACGM, Palm WM, et al. Measuring longitudinal white matter changes: comparison of a visual rating scale with a volumetric measurement. AJNR Am J Neuroradiol 2006;27:875–8.
[11] Grimaud J, Lai M, Thorpe J, Adeleine P, Wang L, Barker GJ, et al. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn Reson Imaging 1996;14:495–505.
[12] Anbeek P, Vincken KL, van Osch MJP, Bisschops RHC, van der Grond J. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. Med Image Anal 2004;8:205–15.
[13] Admiraal-Behloul F, van den Heuvel DMJ, Olofsen H, van Osch MJP, van der Grond J, van Buchem MA, et al. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. Neuroimage 2005;28:607–17.
[14] de Boer R, Vrooman HA, van der Lijn F, Vernooij MW, Ikram MA, van der Lugt A, et al. White matter lesion extension to automatic brain tissue segmentation on MRI. Neuroimage 2009;45:1151–61.
[15] Jack CR, O'Brien PC, Rettman DW, Shiung MM, Xu Y, Muthupillai R, et al. FLAIR histogram segmentation for measurement of leukoaraiosis volume. J Magn Reson Imaging 2001;14:668–76.
[16] Khademi A, Venetsanopoulos A, Moody AR. Robust white matter lesion segmentation in FLAIR MRI. IEEE Trans Biomed Eng 2012;59:860–71.
[17] Shiee N, Bazin P-L, Ozturk A, Reich DS, Calabresi PA, Pham DL. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage 2010;49:1524–35.
[18] Geremia E, Clatz O, Menze BH, Konukoglu E, Criminisi A, Ayache N. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage 2011;57:378–90.
[19] Barkhof F, Smithuis R. The Radiology Assistant - Multiple Sclerosis, 2007, http://www.radiologyassistant.nl/en/p4556dea65db62/multiple-sclerosis.html
[20] Herskovits EH, Itoh R, Melhem ER. Accuracy for detection of simulated lesions: comparison of fluid-attenuated inversion-recovery, proton density–weighted, and T2-weighted synthetic brain MR imaging. AJR Am J Roentgenol 2001;176:1313–8.
[21] Ong K, Ramachandrama D, Mandavaa R, Shuaibb I. Automatic white matter lesion segmentation using an adaptive outlier detection method. Magn Reson Imaging 2012;30:807–23.
[22] Blekas K, Likas A, Galatsanos NP, Lagaris IE. A spatially constrained mixture model for image segmentation. IEEE Trans Neural Netw 2005;16:494–8.
[23] Tang H, Dillenseger J-L, Bao XD, Luo LM. A vectorial image soft segmentation method based on neighborhood weighted Gaussian mixture model. Comput Med Imaging Graph 2009;33:644–50.
[24] Neema M, Guss ZD, Stankiewicz JM, Arora A, Healy BC, Bakshi R. Normal findings on brain fluid-attenuated inversion recovery MR images at 3T. AJNR Am J Neuroradiol 2009;30:911–6.
[25] Ruan S, Jaggi C, Xue J, Fadili J, Bloyet D. Brain tissue classification of magnetic resonance images using partial volume modeling. IEEE Trans Med Imaging 2000;19:1179–87.
[26] Lemieux L, Hammers A, Mackinnon T, Liu RSN. Automatic segmentation of the brain and intracranial cerebrospinal fluid in T1-weighted volume MRI scans of the head, and its application to serial cerebral and intracranial volumetry. Magn Reson Med 2003;49:872–84.
[27] Caillol H, Pieczynski W, Hillion A. Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation. IEEE Trans Image Process 1997;6:425–40.
[28] Dempster A, Laird N, Rubin D. Maximum Likelihood from incomplete data via the EM algorithm. J R Stat Soc B (Methodological) 1977;39(1):1–38.
[29] Xu L, Jordan MI. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Computation 1996;8:129–51.
[30] Gonzalez RC, Woods RE. Digital Image Processing (3rd International ed., Prentice Hall, 2008). ISBN: 978-0-13-505267-9.
[31] Smith SM. Fast robust automated brain extraction. Hum Brain Mapp 2002;17:143–55.
[32] Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 2001;20:45–57.
[33] Bartko JJ. Measurement and reliability: statistical thinking considerations. Schizophr Bull 1991;17:483–9.
[34] Dyrby TB, Rostrup E, Baaré WF, van Straaten EC, Barkhof F, Vrenken H, et al. Segmentation of age-related white matter changes in a clinical multi-center study. Neuroimage 2008;41(2):335–45.
[35] Chen S, Zhang D. Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. IEEE Trans Syst Man Cybern 2004;34:1907–16.
[36] van Leemput K, Maes F, Vandermeulen D, Colchester A, Suetens P. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans Med Imaging 2001;20:677–88.
[37] Figueiredo MAT, Jain AK. Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Mach Intell 2002;24:381–96.
[38] Styner M, Lee J, Chin B, Chin M, Commowick O, Tran H, Markovic-Plese S, Jewells V, Warfield S. 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation (2008). MIDAS Journal - MICCAI 2008 Workshop, http://hdl.handle.net/10380/1509.
[39] Tomas-Fernandez X, Warfield SK. A new classifier feature space for an improved multiple sclerosis lesion segmentation. Biomedical Imaging: From Nano to Macro, IEEE International Symposium on; 2011.

Please cite this article as: Simões R, et al, Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images, Magn Reson Imaging (2013), http://dx.doi.org/10.1016/j.mri.2012.12.004