White matter lesion extension to automatic brain tissue segmentation on MRI

Renske de Boer a,b,⁎, Henri N. Vrooman a, Fedde van der Lijn a, Meike W. Vernooij b,c, M. Arfan Ikram b, Aad van der Lugt c, Monique M.B. Breteler b, Wiro J. Niessen a,d

a Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC, Rotterdam, the Netherlands
b Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands
c Department of Radiology, Erasmus MC, Rotterdam, the Netherlands
d Imaging Science and Technology, Faculty of Applied Sciences, Delft University of Technology, the Netherlands

A R T I C L E   I N F O

Article history:
Received 26 September 2008
Revised 3 December 2008
Accepted 12 January 2009
Available online 22 January 2009

Keywords:
MRI
Brain tissue segmentation
White matter lesions
White matter hyperintensities

A B S T R A C T

A fully automated brain tissue segmentation method is optimized and extended with white matter lesion segmentation. Cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM) are segmented by an atlas-based k-nearest neighbor classifier on multi-modal magnetic resonance imaging data. This classifier is trained by registering brain atlases to the subject. The resulting GM segmentation is used to automatically find a white matter lesion (WML) threshold in a fluid-attenuated inversion recovery scan. False positive lesions are removed by ensuring that the lesions are within the white matter. The method was visually validated on a set of 209 subjects. No segmentation errors were found in 98% of the brain tissue segmentations and 97% of the WML segmentations. A quantitative evaluation using manual segmentations was performed on a subset of 6 subjects for CSF, GM and WM segmentation and an additional 14 for the WML segmentations. The results indicated that the automatic segmentation accuracy is close to the interobserver variability of manual segmentations.

© 2009 Elsevier Inc. All rights reserved.

Introduction

Brain tissue segmentation on structural magnetic resonance imaging (MRI) has received considerable attention. Quantitative analysis of MR images of the brain is of interest in order to study the aging brain in epidemiological studies, to better understand how diseases affect the brain and to support diagnosis in clinical practice. Manual quantitative analysis of brain imaging data is a tedious and time-consuming procedure, prone to observer variability. Therefore, there is a large interest in automatic analysis of MR brain imaging data, especially segmentation of cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM). In the last decade several automatic brain tissue segmentation methods have been proposed, often based on T1-, T2- or proton density-weighted MR images. Some use a fixed set of labeled samples, that were derived from manual segmentations, to train the classifier (Amato et al., 2003; Anbeek et al., 2005). This has, however, some disadvantages as it is dependent on the MRI sequence, requires a laborious training stage and is limited to the MRI intensity variations captured in the training set. Therefore, some studies developed methods to obtain subject-specific training samples labeled by, for example, clustering (Harris et al., 1999; Barra and Boire, 2000), Gaussian mixture models (Ruan et al., 2000; Lemieux et al., 2003) or atlas registration (Cocosco et al., 2003; Song et al., 2006). These methods are independent of intersubject intensity variations and MRI sequence. Another option, often used nowadays, is updating both the classification and the model parameters in an iterative process (Van Leemput et al., 1999; Zhang et al., 2001; Kovacevic et al., 2002; Ashburner and Friston, 2005; Ruf et al., 2005; Awan et al., 2006). This type of method is more complicated but it is also independent of intersubject intensity variations and MRI sequence. Several of these studies have evaluated their method on a reasonably large group varying between 34 to 71 subjects (Harris et al., 1999; Kovacevic et al., 2002; Cocosco et al., 2003; Lemieux et al., 2003; Song et al., 2006; Vrooman et al., 2007). The method developed by Zhang et al. (2001) is incorporated as FSL’s brain tissue segmentation method, FAST, and is used in multiple studies, similar to the SPM brain tissue segmentation method by Ashburner and Friston (2005).

Besides automatic brain tissue segmentation, automatic WML segmentation has also received considerable interest. White matter lesions (WML) are commonly found in elderly subjects and are associated with cognitive decline (de Groot et al., 2002) and increased risk of stroke (Vermeer et al., 2003) and dementia (Prins et al., 2004). Recent studies often use T2-weighted or fluid-attenuated inversion recovery (FLAIR) scans in which white matter lesions are hyperintense. Several automatic segmentation methods have been developed based on intensity alone (Jack et al., 2001; Admiraal-Behloul et al., 2005; DeCarli et al., 2005) or including also spatial- (Anbeek et al., 2004), texture- (Kruggel et al., 2008) or shape-information (Alfano et al., 2000). It is difficult to compare the reported accuracies of these WML segmentation methods. Often

⁎ Corresponding author. Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands.
E-mail address: reniske.deboer@erasmusmc.nl (R. de Boer).

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.01.011
different evaluation measures are used and some of these measures depend on the WML load of the subject (Admiral-Behloul et al., 2005). Furthermore, automatic WML segmentations are often evaluated by comparison to manual segmentations and the evaluation is therefore influenced by the manual segmentation protocol. The robustness of an automatic segmentation method can be demonstrated by applying the method to a large dataset. Only some studies evaluated their WML segmentation method on datasets of 100 or more subjects (Admiral-Behloul et al., 2005; Kruggel et al., 2008; Maillard et al., 2008).

In this paper a fully automated method for CSF, GM and WM segmentation based on multimodal MRI data is optimized and extended with WML segmentation. The contribution of this paper to the existing literature is threefold. Firstly, we evaluate different atlas registration methods for a brain tissue segmentation method presented by Cocosco et al. (2003) and Vrooman et al. (2007) where atlas registration is used to automatically train a k-nearest neighbor classifier. Different types of registration are compared: single- versus multiple-atlas registration; affine versus B-spline based non-rigid registration at different control point spacings; and registration of a varying number of atlases. Secondly, the method is extended with an automatic WML segmentation. This segmentation method uses the GM classification to determine a white matter lesion intensity threshold value in the FLAIR scan. Thirdly, the method is qualitatively validated on a large dataset of 209 elderly subjects. A quantitative evaluation is performed on a small subset using manual segmentations.

Materials and methods

Atlas data

Twelve atlases have been obtained by manual segmentation of scans from the Rotterdam Scan Study (de Leeuw et al., 2001), that were acquired in 1995–1996. This population-based imaging study is aimed at investigating determinants of age-related neurologic diseases among elderly persons. The twelve subjects were female and had a mean age (±standard deviation) of 64 (±1.8) years. MR brain imaging was performed on a 1.5 T Siemens scanner using a quadrature head coil. An inversion recovery double contrast, 3D half-Fourier acquisition single-shot turbo spin echo (HASTE) sequence was performed (TR=2800 ms, TI=440 ms, matrix=192×256, FOV=25×25 mm², 128 contiguous sagittal slices of 1.25 mm). The voxel dimensions were 1×1×1.25 mm³. Two HASTE modules were sequentially acquired after the inversion pulse (effective TE of 29 ms and 440 ms). Each HASTE module combined non-selective radio frequency excitations to provide a short interecho spacing of 3.9 ms. We used the first HASTE module (HASTE-Odd), with contrast similar to an inverted T1-weighted image, for subsequent processing. The datasets were manually segmented by two trained physicians, using a paintbrush method in the tool ‘Display’ from the Montreal Neurological Institute (MNI). The labeled output contained four labels, background (BG), CSF, GM and WM.

Test data

Imaging data from the Rotterdam Scan Study (Hofman et al., 2007), acquired in 2005–2006, were used for the evaluation of the method. Scans were obtained on a 1.5 T GE scanner using an 8-channel head coil. The protocol included three high-resolution axial MRI sequences, i.e. a T1-weighted 3D Fast RF Spoiled Gradient Recalled Acquisition in Steady State with an inversion recovery pre-pulse (FASTSPGR-IR) sequence (TR=13.8 ms, TE=2.8 ms, TI=400 ms, FOV=25×25 cm², matrix=416×256 (interpolated to 512×512 resulting in voxel sizes of 0.49×0.49 mm³), flip angle=20°, NEX=1, bandwidth (BW)=12.50 kHz, 96 slices with slice thickness 1.6 mm zero-padded in the frequency domain (0.8 mm), a proton density (PD) weighted sequence (TR=12,300 ms, TE=17.3 ms, FOV=25×25 cm², matrix=416×256, NEX=1, BW=17.86 kHz, 90 slices with slice thickness 1.6 mm), and a FLAIR sequence (TR=8000 ms, TE=120 ms, TI=2000 ms, FOV=25×25 cm², matrix=320×224, NEX=1, BW=31.25 kHz, 64 slices with slice thickness 2.5 mm).

For this study, 215 subjects were used, that were randomly chosen from age- and sex-specific strata from the total Rotterdam Scan Study population. Table 1 shows the subjects per age group and sex. All subjects were non-demented and none had multiple sclerosis, even though the latter was no exclusion criteria. Two independent physicians performed manual segmentations of brain tissues in six datasets, using a paintbrush method in the MNI-tool ‘Display’. Scans were manually segmented into CSF, GM and WM on the T1-weighted volumes and WMLs were segmented on the FLAIR volumes. Another 20 subjects had manual WML segmentations by one expert. Six of these 20 subjects were randomly selected for parameter optimization of the WML segmentation and were therefore not included in the final set of 209 test subjects.

Pre-processing of test data

We use T1-weighted and PD-weighted MR images for the segmentation of CSF, GM and WM by the previously proposed method. White matter lesions are segmented using a FLAIR scan. All scans are registered to the T1-weighted image by rigid registration and resampled by trilinear interpolation to the resolution of the T1-weighted image. A brain mask, to exclude e.g. cerebellum, eyes and skull, is obtained by non-rigid registration of a manual segmented brain mask to the T1-weighted image using Elastix (www.isi.uu.nl/Elastix/). Subsequently, the scans are corrected for intensity non-uniformity using the N3 method described by Sled et al. (1998) within the brain mask. A range matching procedure on the PD-weighted and T1-weighted images ensures unbiased feature weights in the subsequent k-nearest neighbor (kNN) classification. This procedure excludes 4% of the voxels with lowest intensities and 4% of the highest intensity voxels while rescaling the remaining intensities between zero and one.

Brain tissue segmentation

In the first stage of the algorithm, CSF, GM and WM are segmented using an automatically trained kNN classifier which is an

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Subjects per sex and age group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total subjects included</td>
</tr>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>Men</td>
<td>Total</td>
</tr>
<tr>
<td>Women</td>
<td>105</td>
</tr>
<tr>
<td>Total</td>
<td>215</td>
</tr>
</tbody>
</table>

¹ 'Total' is the number of subjects in the corresponding sex/age group.  
² 'TMS' is the number of subjects with total manual segmentation by 2 observers.  
³ 'LMS' states the number of subjects with manual WML segmentation by 1 observer.
extension of the work by Cocosco et al. (2003). Training samples for the kNN classifier are obtained from the subject itself by atlas-based registration. This is accomplished by registration of single- or multiple-atlases to the subject, using the registration method introduced by Rueckert et al. (1999). The centers of mass are aligned first, followed by an affine registration and, in some experiments, a spline-based non-rigid registration using decreasing control point spacings (20 mm, 10 mm, 5 mm, 2.5 mm). All registrations are driven by normalized mutual information. The transformations are obtained by registration of the grayscale images and applied to the labeled images.

The registered labeled images are equally weighted in the averaging to create tissue probability maps (TPMs) of the three tissue classes and a fourth background class. These TPMs give the probability, represented by a value between zero and one, of a certain voxel belonging to a certain (tissue) class. The TPMs are thresholded in order to get candidate training samples with a predefined probability to belong to a specific label. A threshold of 0.7 is chosen, which is shown by Vrooman et al. (2007) to be a good threshold value for these tissue types, especially CSF. When the CSF TPM is thresholded at this value, it includes not only ventricular CSF but also sulcal CSF, contrary to higher thresholds. Inclusion of sulcal CSF will result in more variations in the CSF samples.

For all four classes, 7500 candidate training samples are randomly taken from the spatial locations masked by the thresholded TPMs. The features of the samples consist of the intensity values of the PD- and T1-weighted images at the sample locations. A pruning step is applied to the initial set of samples to remove samples with incorrect labels (Cocosco et al., 2003). First, a minimal spanning tree of the samples in feature space is created. In an iterative process, the pruning algorithm removes connections whose length exceeds a threshold value equal to a constant multiplied with the average length of the other connections of a sample. At every iteration the threshold value decreases. This process is continued until every tissue class has a unique main cluster in feature space. A main cluster is defined as the cluster containing more samples of a certain class than the other clusters. A cluster is a unique main cluster when it is the main cluster for a single class. The final step removes all samples that are not connected or that are not in their main cluster. A k-nearest neighbor classifier performs the final classification based on the pruned sample set. A value of 45 is used for k, similar to Vrooman et al. (2007) and Cocosco et al. (2003). The kNN implementation uses a fast nearest neighbor lookup library (http://www.cs.umd.edu/mount/ANN).

White matter lesion segmentation

Upon completion of the first step of the algorithm in which CSF, GM and WM are segmented, WMLs that are present in the brain are misclassified as GM with a ‘halo’ of WM. Fig. 1 shows a slice of a FLAIR image and an automatic brain tissue segmentation with an outline of a corresponding manual WML segmentation. In the FLAIR image the WMLs are clearly visible as hyperintensities and thus do not resemble GM intensity. Therefore a histogram is created of all voxels in the FLAIR image that are classified as GM. An example of such a FLAIR histogram is shown in Fig. 2. The highest peak in the histogram corresponds to the true gray matter voxels. The FLAIR intensities corresponding to white matter lesion voxels are located to the right of this peak. The histogram is smoothed by a convolution with a Gaussian kernel (σG = 4 FLAIR intensity units). This makes it possible...
to estimate the FLAIR intensity corresponding to the center of the GM peak by the histogram bin containing most true positive gray matter voxels. The peak is approximated by a Gaussian function with the mean (μ) defined as the peak center location and the standard deviation (σ) calculated using the full width at half maximum. The threshold T for the WML is subsequently defined as:

\[ T = \mu + \alpha \sigma, \]  

with α a threshold parameter to be optimized. The threshold for the example in Fig. 2 is shown as a dashed line. The WML segmentation is obtained by thresholding the FLAIR image.

Upon thresholding, a number of regions are wrongly classified as WML. In several segmentation methods, information from neighboring voxels is utilized in order to improve segmentation performance, e.g. through Markov random fields (Van Leemput et al., 1999; Van Leemput et al., 2001; Ruan et al., 2000; Khayati et al., 2008). In our case, most of the false positive WML are clearly located outside the white matter and a relatively simple measure suffices. For every lesion, the following fraction is calculated,

\[ \text{WM fraction} = \frac{\text{number of neighboring WM voxels}}{\text{number of neighboring CSF + GM voxels}}, \]  

within the one voxel wide surroundings, obtained using a 3D 18-neighborhood relation (spherical kernel with a radius of 3). A lesion is defined as a group of connected voxels, using a 3D 18-connectivity, with a WML label. If the fraction mentioned above is smaller than the optimized parameter β the lesion is reclassified as GM.

**Post-processing**

The non-rigidly transformed brain mask sometimes includes parts outside the brain. These parts, mainly dura and skull, are classified by the automatic segmentation method as mixtures of GM, WM and WML depending on their intensities. Especially false positive WML outside the brain can have a large influence on the relatively small total volume of WML. A simple post-processing step is applied to remove brain tissue and WML located outside the brain. This step uses the brain tissue and WML classification to find components of connected voxels, defined by 3D 18-connectivity, with the same label. For every component, the number of neighboring background and non-background voxels are counted. Just like the false positive WML reclassification step, a 3D 18-neighborhood relation is used for defining neighboring voxels. If the ratio of background voxels to non-background voxels is larger than a certain value the component is relabeled as background. Pilot experiments showed that the improvement of the brain mask is not very sensitive to this value. A team of experts chose the value to be 0.4 by visual inspection of several subjects.

**Evaluation measures**

The automatic segmentations are evaluated quantitatively by calculation of the similarity index (SI) or Dice coefficient between the automatic segmentation and the manual segmentations.

\[ \text{SI} = \frac{2(S_1 \cap S_2)}{S_1 + S_2} \]  

where \( S_j \) with \( j = 1, 2 \) is a segmented volume and \( S_1 \cap S_2 \) is the overlap of \( S_1 \) and \( S_2 \). The SI is also used as a measure for the interobserver variability.

The true positive fraction (TPF), or sensitivity, and the extra fraction (EF) are also used for evaluation. The extra fraction is a measure for oversegmentation. EF and TPF are defined as follows:

\[ \text{EF} = \frac{\text{FP}}{\text{TP} + \text{FN}} \]  

\[ \text{TPF} = \frac{\text{TP}}{\text{TP} + \text{FN}} \]  

by false positives (FP), true positives (TP) and false negatives (FN). For the six subjects with total manual segmentation (CSF, GM, WM and WML) the quantitative evaluation is only performed for the voxels that are present in the manual segmentation. This means that the subjects are masked by the total area of the manual CSF, GM, WM and WML segmentations.

In case of small white matter lesions, a slight oversegmentation or undersegmentation will easily lead to low similarity indices, whereas the influence on e.g. the total white matter lesion load is small. Therefore, for evaluation of WML segmentations, we also consider an evaluation measure which describes the maximum and average boundary localization error. This error is computed by calculating a distance map from the border of the lesion in one segmentation to the border of the corresponding lesion in the other segmentation, and subsequently computing the mean and maximum distance per subject. Because the distance measure is not symmetric, each WML segmentation is compared with the other and vice versa and the result is averaged. Of course the distance can only be measured for the lesions that have been identified in both segmentations and therefore we also report the total volume of the FP and FN lesions.

**Experiments**

**Optimization of α and β**

The two WML segmentation parameters α and β were optimized on six of the 20 subjects who had only a manual WML segmentation. These subjects were not used in any of the other experiments. For these six subjects WML segmentations were obtained by the automatic method with non-rigid registration of 11 atlases with 2.5 mm control point spacing and a range of α from 2.0 to 3.1 at intervals of 0.1 and of β from 0.02 to 0.30 at intervals of 0.02. For every subject the SI between the automatic and the manual segmentations was calculated and ranked within the parameter ranges. The parameter values with the minimal summed rank were chosen as the optimal parameter settings. This resulted in α=2.3 and β=0.26. These settings for α and β were used in all other experiments.

**Comparison of different types of atlas registration**

To evaluate the influence of different types of atlas registration on the resulting segmentations, a comparison was made using the six subjects with manual segmentations of the brain tissues and WML performed by two observers. Registration can be time consuming, so reduction of registration time might be desirable. This can e.g. be accomplished by 1) reducing the number of registrations, 2) using affine instead of non-rigid registration or 3) increasing the control point spacing. All these alternatives were tested. First, a single average atlas registration was considered. The average atlas was created by non-rigidly registering 11 atlases to a twelfth. This twelfth atlas was identified as the best target for alignment similar to the method described in Smith et al. (2006). The non-rigid registration of the 11 atlases to the target atlas was performed with a control point spacing of 2.5 mm. The resulting probability atlas was registered to the subject by applying the transformation obtained from registering the alignment target HASTE-Odd image to the subject T1-weighted image. Secondly, multiple-atlas registration was performed by
registering all 12 atlases to the subject and averaging the result to create a probability map. Both the single-atlas and the multiple-atlas methods were tested with affine registration and non-rigid registration at control point spacings of 20 mm, 10 mm, 5 mm and 2.5 mm. In another experiment, the number of atlases was varied. All 12 atlases were used for each phase, consisting of either twelve tests with one atlas, four tests with three randomly picked atlases, or two tests with six randomly picked atlases. All tests were performed using non-rigid registration at a control point spacing of 2.5 mm. The resulting segmentations were evaluated using the similarity index for CSF, GM, WM, WML and brain (GM+WM+WML) averaged over the six subjects. The type of atlas registration yielding the optimal segmentation is determined by summing the average CSF, GM, WM and WML similarity indices and is used for the other experiments.

Quantitative analysis of brain tissue and WML segmentation

The same six subjects were used for quantitative evaluation of the brain tissue and WML segmentation method. TPF, EF and SI were used as evaluation measures. The resulting SIs between automatic and manual segmentations were compared to the interobserver SI of the manual segmentations. For WML segmentations in these six subjects the boundary localization error was assessed by calculating the distance measures. The remaining 14 subjects with white matter lesion segmentation by only one observer were used to provide insight into the amount of over- or under-segmentation depending on total lesion volume. This was accomplished by relating SI and volume differences to the total lesion volume in the segmentations. The differences between automatic and manual segmentation were further investigated by calculating the volume difference per subject. This volume difference per subject was divided into false positive volume and false negative volume and was calculated before and after the reclassification step. Additionally, we investigated whether the performance of the FP WML removal step depended on whether the atlas registration yield the optimal segmentation could be based on either the size of the WML volume as determined with the automatic WML segmentation method. All atlases are non-rigidly registered with a control point spacing of 2.5 mm. SIs are averaged over the tests and the 6 subjects compared to the manual segmentations by two observers.

Qualitative analysis in large dataset

All 209 subjects were segmented using the optimal method from the atlas registration study. The results were visually inspected by a team of two experts. Of every subject, three representative axial slices at 12 mm distance, of both the FLAIR image and the corresponding automatic segmentation were shown and the experts were asked to evaluate the segmentations. The CSF, GM and WM segmentations were rated as ‘good’, ‘reasonable’ or ‘poor’. For the WML segmentations it was specified if there were voxels that were clearly false positive or false negative lesions. This WML over- or under-segmentation specification could be based on either the size of the white matter lesions or on extra or missing lesions. Brain mask errors were not taken into account in the segmentation evaluation because the main focus is on the tissue and WML segmentation.

Association between age and WML volume

Finally, we assessed the association between age and total white matter lesion volume as determined with the automatic WML segmentations. The strength of these associations is compared with previously reported values from population samples in the discussion of this work. Because of small numbers of people over the age of 80 years, we restricted our analyses on the effect of age on white matter lesion volume to persons aged 60 to 80 years. Moreover, we excluded the subjects with suboptimal WML segmentations according to the visual inspection in the previous experiment.

Results

Comparison of different types of atlas registration

The results for the accuracy study, as a function of the atlas registration method used in training, are shown in Fig. 3. The given SIs are averages of twelve SIs obtained by comparing the automatic
segmentations of the six subjects with the manual segmentations by two observers. Fig. 4 shows the average SIs of the experiment using different numbers of atlases. The SIs obtained by registering 12 atlases are given as comparison. Non-rigid registration using 12 atlases and a control point spacing of 2.5 mm gave the highest summed SI and is therefore used for the other experiments. Especially the CSF segmentation is influenced by the type of registration and benefits from registering at a small control point spacing and using 12 atlases.

Quantitative analysis of brain tissue and WML segmentation

Table 2 gives the average TPF, EF, SI, along with the interobserver SI, of the six subjects using the best performing registration method. The SI of the different tissue types is close to the interobserver SI. TPF is overall high and EF is low for all tissue types except WML. The latter is mainly due to the two subjects with lowest WML load. Without these two subjects the EF is 0.16. An example image of segmentation results is shown in Fig. 5. An example segmentation image of a subject with low lesion load is given in Supplementary Fig. 1.

The maximum and mean absolute distance between two WML segmentations, obtained using the distance measure as defined in the Section 'Evaluation measures', is shown in Fig. 6. Subjects are ordered according to their average manual WML segmentation volume. Whereas sometimes a large maximum error occurs (Fig. 6(A)), Fig. 6(B) shows that these large disagreements do not occur frequently. The mean absolute distance between segmentations is smaller than 1 mm in five out of six cases, and on average 0.4 mm. Fig. 6(C) shows the average total volume percentage that was excluded from the distance evaluation because the corresponding lesions are not present in both segmentations.

Fig. 7(A) shows a plot of the similarity index of the WML segmentations versus their volume in the manual segmentation for the 20 subjects with manual WML segmentations by one or more observers. The results for the six subjects with manual segmentations by two observers were averaged. As expected, SI of subjects with small WML volume is lower than the SI of subjects with large WML volume. Similarly the volume differences between the automatic and the manual segmentations are shown in the normalized Bland–Altman plot in Fig. 7(B). The volume differences were normalized by the average segmentation volume to emphasize that sometimes the disagreement is almost as large as the average total WML volume. This is the case for subjects with small WML volume and even the disagreement between the observers is large in these subjects.

An example of white matter lesions reclassified by the FP WML reclassification step can be seen in Fig. 8. This example subject has, compared to observer 1, 4.92 ml FP and 0.272 ml FN WML before and 1.59 ml FP and 0.275 ml FN WML after the reclassification step. The total WML volume according to observer 1 is 9.93 ml. The reclassification step removes the FP WML in the cortical gray matter while preserving e.g. the periventricular white matter lesions. Fig. 9(A) shows the total false positive and false negative WML volume before and after the FP WML reclassification step for 20 subjects. For the six subjects, the two manual volumes and their FP and FN volumes are averaged. It is obvious that the reclassification step decreases the FP volume while keeping the FN volume increase to a minimum. After reclassification, false positive WML volume is less than 2 ml for 18 out of 20 subjects. The graph indicates an FN volume dependence on WML.
load for subjects with WML load up to 6 ml. Fig. 9(B) shows the total false positive WML volume per bin for 20 subjects with a bin size of 10 voxels (except for the last three bins). For the six subjects with manual segmentations by two observers are shown in black. Their SI values and volumes are averaged over the two observers for the SI plot (A).

Fig. 6. Maximum (A) and mean (B) distance between segmentation borders. Percentage of averaged false positive and false negative lesion volumes (C).

Fig. 7. Similarity index (A) and normalized Bland-Altman plot (B) of WML for all subjects with manual segmentation. ‘Aut’ refers to the automatic segmentation, ‘obs’ to observer and ‘segm’ to segmentation. The 6 subjects with manual WML segmentations by two observers are shown in black. Their SI values and volumes are averaged over the two observers for the SI plot (A).

Fig. 8. Example of WML reclassified as GM by the false positive WML reclassification step. From dark to light gray the labels represent CSF, GM, WM and WML. The white labels represent reclassified WML.
size has a range up to 9.3 ml and after this step the maximum FP lesion size is 75 μl. There is no strong indication for lesion size dependence of the FP WML reclassification step.

**Qualitative analysis in large dataset**

Table 3 shows the results for the visual inspection of the 209 subjects. According to the team of experts, 98% of the subjects had ‘good’ brain tissue segmentations and 97% had no obvious oversegmented or undersegmented WML. The six subjects with total manual segmentation and the 14 subjects with manual WML segmentations also had ‘good’ CSF, GM and WM segmentations and no obvious WML over- or undersegmentation. Ten subjects had ‘reasonable’ or ‘poor’ brain tissue segmentations and/or oversegmented or undersegmented WML. The MRI scans of these ten subjects were inspected fully by two experts. Seven of these subjects show motion artifacts in the T1-weighted, PD-weighted and/or FLAIR image. If these subjects are excluded, only one ‘reasonable’ brain tissue segmentation, one ‘oversegmented’ WML segmentation and one ‘undersegmented’ WML segmentation remain. An example slice of the remaining ‘reasonable’ brain tissue segmenta-

**Association between age and WML volume**

This experiment was performed on all subjects between 60 and 80 years old, excluding five subjects in this age category with suboptimal WML segmentations. Study population characteristics, mean WML volume and difference in WML volume per year increase in age are given in Table 4. In order to correct for individual head size, WML volume is expressed as percentage of intra-cranial volume (ICV). The WML volume is natural log transformed for the analysis because of leftward skewness of the untransformed measure. The final two rows in Table 4 list the mean WML volume transformed back to percentages and the exponential WML growth. A scatterplot of age versus WML volume is shown in Fig. 11.

**Discussion and conclusion**

A fully automated method for CSF, GM and WM segmentation has been optimized, extended with WML segmentation, and quantitatively and qualitatively validated. The different brain tissues, CSF, GM and WM, are segmented by an automatically trained kNN classifier using atlas registration. The quantitative evaluation comparing the automatic segmentations to manual segmentations showed similarity indices close to the interobserver similarity index. The quantitative evaluation was performed primarily in the age category 60–70 years, as for these scans manual segmentations were available. There was no noticeable difference between age groups in the qualitative evaluation. We therefore do not expect that the accuracy would be different in different age groups.

The atlases used for the registration-based automatic training were available from an earlier study and were based on data acquired with a different scanner and scanning protocol than the data on which the method was evaluated. This shows that the method can be successfully applied to data acquired with different scanning protocols than the scanning protocol of the atlas datasets. In the study we evaluated (1) single-atlas versus multiple-atlas registration, (2) affine versus non-rigid registration at different control point spacings and (3) registration of different number of atlases. A non-rigidly registered atlas is able to capture more differences between anatomies and this leads to better training samples for classification. Increasing the number of atlases also allows for more anatomical variation to be captured in the training samples. Especially the CSF segmentation benefits from non-rigid registration with a small control point spacing and the use of more atlases. This is most likely caused by the difficulty to obtain training samples of sulcal CSF. The subarachnoid space is very narrow and especially the gyri vary widely between subjects. Non-rigid registration of 12 atlases using a control point spacing of 2.5 mm resulted in the best segmentations with accuracy close to the

**Table 3** Result of visual inspection

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Number of subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSF, GM, WM</td>
<td></td>
</tr>
<tr>
<td>segmentations</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>204</td>
</tr>
<tr>
<td>Reasonable</td>
<td>4</td>
</tr>
<tr>
<td>Poor</td>
<td>1</td>
</tr>
<tr>
<td>WML segmentation</td>
<td></td>
</tr>
<tr>
<td>No FP or FN</td>
<td>202</td>
</tr>
<tr>
<td>Oversegmentation</td>
<td>4</td>
</tr>
<tr>
<td>Undersegmentation</td>
<td>2</td>
</tr>
</tbody>
</table>
interobserver variability. This approach, however, also involves the most computing time. Computation time may be reduced by faster registration methods, e.g. Klein et al. (2007). Further increase of the number of atlases might improve the segmentation even more, but we could not study this due to the limited number of atlases available. The segmentation might also benefit from decreasing the control point spacing beyond 2.5 mm. This will, however, increase the computational costs even further and might give the registration algorithm too many degrees of freedom leading to suboptimal registrations. The atlas data was from elderly subjects, although slightly younger than the average age of the test subjects. Since the atlases are only used for the brain tissue segmentation and since the obtained samples are pruned before they are used, we do not expect any problems to occur when the atlas age is mismatched as long as multiple atlases are used to correct for possible age-related variations or lesions.

White matter lesions are segmented using the brain tissue segmentation and a FLAIR scan. A WML threshold for the FLAIR image is determined by using the FLAIR intensity histogram within a GM mask, obtained from the GM segmentation. It is followed by a simple post-processing step to ensure that the lesions found are within the white matter. This is accomplished by thresholding the WM fraction of neighboring voxels for every lesion. Experiments show that this reclassification step removes a large fraction of false positive white matter lesions while keeping the increase in FN WML to a minimum. The performance of this reclassification does not strongly depend on total white matter lesion load or individual lesion volume. The acquired accuracy of the WML segmentation is not as high as for the brain tissue segmentation but it is still in the range of the interobserver variability. Most false positive WML are very small and a WML size cutoff might therefore be considered. Based on the experiments discarding white matter lesions up to 1.9 μl (10 voxels) would remove a large fraction of false positive WML volume. Distance measures show that in general the average distance between the automatic and the manual lesion segmentation boundaries is small. There is, however, no reliable in vivo method to find out if a suspected lesion area in a MR image is a true lesion and it is therefore impossible to obtain a true golden standard. This results in relatively large disagreements between WML segmentations, both manually and

<table>
<thead>
<tr>
<th>Study population characteristics, mean WML volume and difference in WML volume per year increase of age (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
</tr>
<tr>
<td>Number of subjects</td>
</tr>
<tr>
<td>Age (year)</td>
</tr>
<tr>
<td>Intra-cranial volume (ml)</td>
</tr>
<tr>
<td>White matter lesionsa</td>
</tr>
<tr>
<td>Δ WML volume per yeara (95%CI)</td>
</tr>
<tr>
<td>White matter lesions (%) of ICV</td>
</tr>
<tr>
<td>Exponential WML growth (95%CI)</td>
</tr>
<tr>
<td>(1.043; 1.094)</td>
</tr>
</tbody>
</table>

a Natural log transformed percentage of intra-cranial volume.

b Difference in WML volume per year increase in age.

Fig. 10. Examples of segmentation slices viewed by the experts and labeled as having ‘reasonable’ (A) or ‘poor’ (B) brain tissue segmentation. The corresponding MRI slices are also shown. Segmentation (B) had undersegmented WML according to the experts.

Fig. 11. Scatterplot of natural log transformed white matter lesion volume against age. Volumes are expressed as percentage of intra-cranial volume. Regression lines for linear fit are shown as a solid line for men and a dashed line for women.
automatically obtained. Especially for subjects with low lesion load this has a large influence on similarity index- and extra fraction-values. For these subjects, differences in volume can be almost as large as the average total volume. Other studies (Zijdenbos et al., 1998; Van Leemput et al., 2001) also reported high interobserver and intraobserver variability in WML segmentation and the resulting problem with validating automated segmentation methods.

Visual inspection of 209 segmentations showed no obvious segmentation errors in 98% of the brain tissue segmentations and 97% of the WML segmentations of the subjects. The majority of subjects with segmentation errors had motion artifacts in one or more MR images. The high percentage of correct segmentations shows that the method is robust to intersubject variations. The subject-specific classifier-training by atlas registration prevents segmentation errors due to MR intensity variations between scans and even to a certain extent within an MR image.

We also assessed the association between age and WML volume in subjects aged 60 to 80 years old with no WML segmentation errors. This association was expressed as difference in the natural log transformed WML percentages of ICV per year increase in age (and 95% confidence interval). The associations that we found in our study (0.055 [95% CI 0.018; 0.091] for men and 0.075 [95% CI 0.042; 0.11] for women) were similar to both the estimates reported by Ikram et al. (2008) from their manually corrected automatic WML segmentations in a different subset of the Rotterdam Scan Study that was scanned with a different scanning protocol, and the estimates reported by DeCarli et al. (2005) from semiautomated image segmentation analysis of more than 2200 participants of the Framingham Heart Study.

Partial volume effects may have an impact on segmentation results, as they can induce errors in both the atlas registration and the kNN classification. In order to reduce this error, it is possible to perform a probabilistic brain tissue segmentation by defining the tissue class probability as the fraction of neighbors of this class in the kNN classification. The white matter lesion segmentation method, however, is not designed to give a probabilistic output. Anisotropic voxels have more partial volume artifacts in one (or more) direction(s) than in the other. This will also influence the segmentation results. Furthermore, the WML segmentation and post-processing steps use a symmetric neighborhood definition. Anisotropic voxels will make these processing steps anisotropic too. We did not find this to have any noticeable and undesirable effect given the voxel sizes used in this study. As seen in the experiment evaluated by visual inspection, motion artifacts have a great influence on the segmentation result. Motion does not only induce a possible ‘ringing artifact’ in the final segmentation but it also influences the training samples obtained after atlas registration. This can result in a classification of GM voxels as WM and vice versa. It is therefore important to carefully inspect the MR images used.

In conclusion we introduced and optimized an automated method for CSF, GM, WM and WML segmentation and showed that its accuracy is close to the interobserver variability. Robustness was shown in a large study on 209 subjects. The association between age and WML volume of the subjects younger than 80 years old are comparable to reported associations found by similar studies.

Acknowledgments

The Rotterdam Study is supported by the Erasmus Medical Center and Erasmus University Rotterdam, the Netherlands Organization for Scientific Research (NWO), the Netherlands Organization for Health Research and Development (ZonMw), The Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry of Health, Welfare and Sports, the European Commission (DG XII) and the Municipality of Rotterdam. This study was further financially supported by the Netherlands Organization for Scientific Research (NWO) grants 948-00-010 and 918-46-615. This work benefited from the use of Elastix, software developed by Stefan Klein and Marius Staring, carried out at the Image Sciences Institute (ISI), on funding granted by the Netherlands Organization for Scientific Research (NWO). It is available at lwww.isi.uu.nl/Elastix/. The authors would like to express their gratitude to Mr. Tobias A. Knoch and Mr. Anis Abusereis from the Erasmus Grid Office for their considerable support in performing the computationally intensive registrations for this work. The authors would also like to thank Mr. Marius de Groot for his editorial assistance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2009.01.011.

References


