ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 21 (2004) 1037–1044

Probabilistic segmentation of white matter lesions in MR imaging

Petronella Anbeek,* Koen L. Vincken, Matthias J.P. van Osch, Robertus H.C. Bisschops, and Jeroen van der Grond

Department of Radiology, Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands

Received 17 July 2003; revised 8 October 2003; accepted 8 October 2003

A new method has been developed for fully automated segmentation of white matter lesions (WMLs) in cranial MR imaging. The algorithm uses information from T1-weighted (T1-w), inversion recovery (IR), proton density-weighted (PD), T2-weighted (T2-w) and fluid attenuation inversion recovery (FLAIR) scans. It is based on the *K*-Nearest Neighbor (KNN) classification technique that builds a feature space from voxel intensities and spatial information. The technique generates images representing the probability per voxel being part of a WML. By application of thresholds on these probability maps, binary segmentations can be obtained. ROC curves show that the segmentations achieve both high sensitivity and specificity. A similarity index (SI), overlap fraction (OF) and extra fraction (EF) are calculated for additional quantitative analysis of the result. The SI is also used for determination of the optimal probability threshold for generation of the binary segmentation. Using probabilistic equivalents of the SI, OF and EF, the probability maps can be evaluated directly, providing a powerful tool for comparison of different classification results. This method for automated WML segmentation reaches an accuracy that is comparable to methods for multiple sclerosis (MS) lesion segmentation and is suitable for detection of WMLs in large and longitudinal population studies.

Keywords: Segmentation; White matter lesions; Multiple sclerosis

Introduction

In the last decade, many studies have focused on the prevalence of cerebral white matter lesions (WMLs) in the elderly population or in patients with cardiovascular risk factors. In both patient groups, WMLs are a common finding on cranial MR imaging. Population studies like the Cardiovascular Health Study or the Rotterdam Scan Study have shown that WMLs are associated with age, clinically silent stroke, higher systolic blood pressure, lower forced expiratory volume in 1 s, hypertension, atrial fibrillation, carotid and peripheral arterioscleroses, impaired cognition and depression (De Groot et al., 2000a,b; Longstreth et al., 1996). Furthermore, it has been shown that stroke patients with a large WML load have an increased risk of hemorrhagic transformation, higher preoperative risk of a disabling or fatal stroke during endarterectomy or intercerebral hemorrhage during anticoagulation therapy (Briley et al., 2000). The increased interest in WML research may improve diagnosis and prognosis possibilities for patients with cardiovascular symptoms.

Since WML patterns are very heterogeneous, ranging from punctuate lesions in the deep white matter till large confluent periventricular lesions, the scoring of WMLs is complicated and it has been shown that different visual rating scales lead to inconsistencies between WML studies (Mantyla et al., 1997). Commonly used ordinal WML scoring methods, such as used in the Cardiovascular Health Study or the Rotterdam Scan Study, offer semiquantitative information on the prevalence of WMLs. Exact spatial information is useful since it has been suggested that specific WML patterns are associated with specific symptoms (Benson et al., 2002; Smith et al., 2000). Moreover, for longitudinal studies and to demonstrate relatively small changes in WML patterns, accurate information of WML volume and location is essential. In this respect, the use of an automated segmentation method that detects WMLs with a high sensitivity and specificity, which are demonstrated in a quantitative and objective way, could be advantageous. Successful methods have been developed for the detection of multiple sclerosis (MS) lesions (Alfano et al., 2000; Goldberg-Zimring et al., 1998; Guttman et al., 1999; Kamber et al., 1995; Van Leemput et al., 2001; Warfield et al., 1995a,b; Wei et al., 2002; Zijdenbos et al., 2002). For the more complicated issue of WMLs in general, also some segmentation algorithms exist (Jack et al., 2001; Mohamed et al., 2001; Wei et al., 2002). However, these methods evaluate their results only by visual inspection or measurement of lesion volume. The aim of our research was to develop an automated WML segmentation algorithm, which is fully reproducible and quantitatively validated on a voxel basis.

In this study, we present a method for automatic segmentation of WMLs that is based on a supervised *K*-Nearest Neighbor (KNN) classification technique using information from T1-weighted (T1-w), inversion recovery (IR), proton density-weighted (PD), T2-weighted (T2-w) and fluid attenuation inversion recovery

* Corresponding author. Department of Radiology, Image Sciences Institute, University Medical Center Utrecht, Room E01.335, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands. Fax: +31-30-2513399.
E-mail address: nelly@isi.uu.nl (P. Anbeek).
Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2003.10.012

(FLAIR) scans by probability estimation of voxels being part of a lesion.

Methods

Patients

Twenty patients with arterial vascular disease [transient ischemic attack (TIA), n=4; peripheral arterial disease, n=3; coronary artery disease, n=7; renal artery disease, n = 1; abdominal aorta aneurysm, n=5] were included in this study. The mean age of the patients was 66 years (mean ± SD: 65.6 ± 7.7, range: 49–75). Seventeen patients were male.

MR imaging

MRI studies were performed on a Philips Gyroscan ACS-NT 1.5-T whole body system (Philips Medical Systems, Best, The Netherlands). All patients had the same MR protocol of the brain consisting of transaxial T1-w, IR, T2-w, PD and FLAIR scans. All scans were performed with a 4-mm slice thickness, no slice gap, 38 slices, covering the entire brain, a 230 × 230 mm field of view and a 256 × 256 scan matrix. The individual scan parameters were: T1-w: repetition time (TR)/echo time (TE), 234/2 ms; IR: TR/inversion time (TI)/TE, 2919/410/22 ms; PD: TR/TE, 2200/11 ms; T2-w: TR/TE, 2200/100 ms; and FLAIR: TR/TI/TE, 6000/2000/100 ms. For the IR images, the real images were used and the inversion time was chosen to obtain the best contrast between gray and white matter. The entire acquisition time of all scans was less than 11 min.

Manual segmentation

WMLs were scored independently on hard copies by two investigators (RHCB and JvdG) who were blinded for clinical symptoms of each patient. WMLs had to be hyperintense on FLAIR, PD and T2-w images. WMLs were firstly classified into deep white matter lesions (DWMLs) and periventricular white matter lesions (PVWMLs). The number and size of DWMLs were rated on hard copy according to their largest diameter in categories of: (0) no DWMLs, (1) small (<3 mm), (2) medium (3–10 mm) and (3) large (>10 mm). PVWMLs were rated quantitatively in three regions: adjacent to the frontal horns (frontal capping); adjacent to the lateral wall of the ventricles (bands); and adjacent to the occipital horns (occipital capping) in both hemispheres (De Groot et al., 2000b).

According to the patterns of WMLs, four patient categories were composed of:

1. All patients (n = 20);
2. Patients with small lesion load (n = 8);
3. Patients with moderate lesion load (n = 7);
4. Patients with large lesion load (n = 5).

The patients have been divided into categories 2, 3 and 4 by taking the highest of their DWML and PVWML score (score = 1: small, score = 2: moderate and score = 3: large lesion load).

Secondly, the DWMLs and PVWMLs were manually segmented by the first author (PA). The manual segmentations were independently reviewed and corrected by two investigators (RHCB and JvdG). The final manual WML segmentations were evaluated in a consensus meeting (PA, RHCB and JvdG) and considered as gold standard.

Image preprocessing

The entire image processing protocol started with three preprocessing steps to prepare the data for the KNN classification and analysis. To correct for MR inhomogeneities, a method was used that resulted in similar gray values of major anatomical structures in different patients per image type (Nyúl and Udupa, 1999; Nyúl et al., 2000). To correct for differences owing to patient movement, all images of a patient were registered by rigid registration (translation and rotation), based on normalized mutual information, to the FLAIR image as reference image (Maes et al., 1997). To reduce the amount of data to be investigated and to restrict our analyses to brain tissue only, we isolated the skull and background by applying Mbrase to the T2-w image of every patient (Stokking et al., 2000).

KNN classification

The aim of the method for automatic segmentation of the WMLs was to determine the lesion probability of each voxel. For this purpose, the KNN classification method was used, which is known as a nonparametric procedure for estimation of local class conditional probability density functions from sample patterns (Duda et al., 2001). In general, KNN classification is based on the classification of samples, dependent on their features. In this method, each image voxel is treated as a separate sample. A feature space is defined, in which each axis represents one of the voxel features. A learning set is generated from many preclassified voxels. These learning voxels are entered into the feature space at the coordinates corresponding to their feature values. After this, an image voxel of a new patient is classified by adding it to the feature space and inspection of the K learning voxels that are closest to it. The new case is then classified according to the classes of those K neighbors; for example, the most frequent class of the K neighbors could be assigned to it.

In this application, the learning set for segmentation of one patient was built from the voxels of the other 19 patients (the so-called "leave-one-out" method). All voxels in the learning set were labeled with the value of 0 (non-lesion class) or 1 (lesion class), derived from the manual segmentations. Because of the large number of cases, we decided to randomly select 20% of the voxels for inclusion in the learning set. This reduces computation time and computer memory significantly.

The features used in this study can be divided into two categories: voxel intensities and spatial information. The first group is defined by the signal intensities of a voxel in the acquired images: T1-w, IR, PD, T2-w and FLAIR, which provides a five-dimensional feature space. The second group of features incorporates the spatial location of a voxel in the brain. These were added because in some regions of the brain, lesions are more likely to occur than in others. The spatial features were defined in-plane by two coordinates and through-plane by the z-coordinate. In-plane, the voxel coordinates were measured from the center of gravity in the FLAIR image, which was the reference image for registration, by two different methods. Two types of in-plane coordinates were used separately:

Euclidean coordinates (x and y) as well as the polar coordinates (ρ and φ). Coordinate ρ represented the Euclidean distance from the center of gravity and φ the angle with the horizontal axis.

All experiments were performed with five different features spaces, constructed from different sets of features. Each feature set consisted of all the voxel intensity features and a subset of the spatial features. They were defined by:

F	only voxel intensities
Fxy	voxel intensities and spatial features x and y
Fxyz	voxel intensities and x, y and z
Fρφ	voxel intensities and ρ and φ
Fρφz	voxel intensities and ρ, φ and z

Since different features have different ranges, a rescaling of the feature space was necessary to define a proper metric to compare distances in the feature space, which is essential to justify classification based on KNN. This was achieved by variance scaling: subtraction of the mean of the feature values and division of the outcome by the standard deviation. This approach results in a mean of 0 and variance of 1 for every feature.

The choice of K in KNN classification depends on the number of features and the number of cases. When a small value of K is used, the obtained results are more influenced by individual cases. A larger value of K smoothens the outcome of the classification (Bishop, 1995). If the number of cases goes to infinity, the error rate shows an optimal behavior by approaching the Bayes rate when K increases (Duda et al., 2001). In this study, we used a relatively small number of features in combination with a large number of cases. Therefore, we opted for a relatively large K. It was observed experimentally that for the current purpose a K with a value higher than 100 has a marginal influence on the accuracy of the classification. By taking computation time into account, we concluded that 100 was an acceptable choice for K.

The lesion probability of every voxel was determined by inspection of the K-nearest neighbors of the examined voxel in the feature space. It was defined as the fraction of lesion voxels among those K neighbors. The voxel probabilities were presented in a so-called probability map, which is an image where each voxel intensity value is defined by the lesion probability of that voxel.

Evaluation

By applying different thresholds on the probability map, binary segmentations of the WMLs were produced. These segmentations were compared with the gold standard, where the number of correctly classified voxels, that is, the true positives (TP) and true negatives (TN), was counted as well as the number of false positives (FP) and false negatives (FN). The true positive fraction (TPF), which is the sensitivity, and the false positive fraction (FPF), which is 1-specificity, was calculated for the threshold, running from 0 to 1. They are defined by

$$\mathrm{TPF} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{FPF} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$

The TPF was represented in an ROC curve as function of the FPF for the "all patients" category for all five feature sets.

Furthermore, the binary segmentations were evaluated by three different similarity measures: similarity index (SI) (Dice, 1945; Zijdenbos et al., 1994), overlap fraction (OF) and extra fraction (EF) (Stokking et al., 2000). The SI is a measure for the correctly classified lesion area relative to the total area of WML in both the reference (the gold standard) and the area of the segmented image. The OF measures the correctly classified lesion area relative to the WML area in the reference. The EF measures the area that is falsely classified as lesion relative to the WML area in the reference. The similarity measures are defined by

$$\mathrm{SI} = \frac{2 \times (\mathrm{Ref} \cap \mathrm{Seg})}{\mathrm{Ref} + \mathrm{Seg}},$$

$$\mathrm{OF} = \frac{\mathrm{Ref} \cap \mathrm{Seg}}{\mathrm{Ref}},$$

$$\mathrm{EF} = \frac{\overline{\mathrm{Ref}} \cap \mathrm{Seg}}{\mathrm{Ref}}.$$

In these definitions, Ref denotes the volume of the reference and Seg is the volume of the binary segmentation (Fig. 1). The intersection of Ref and Seg, used in the SI and OF, is similar to the volume of the correctly classified voxels (Overlap). The volume of $\overline{\mathrm{Ref}} \cap \mathrm{Seg}$ corresponds to the false positives (Extra). The SI was represented in a graph as function of the threshold, running from 0 to 1, for all feature sets.

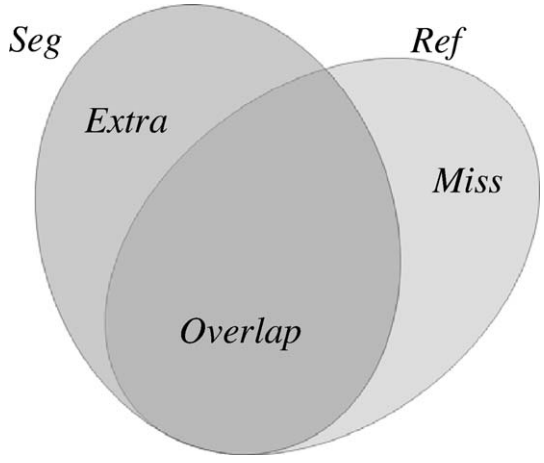

Fig. 1. Comparison of a binary segmentation (Seg) with the reference image (Ref), with (Overlap) the correctly classified voxels, (Extra) the false positives and (Miss) the false negatives.

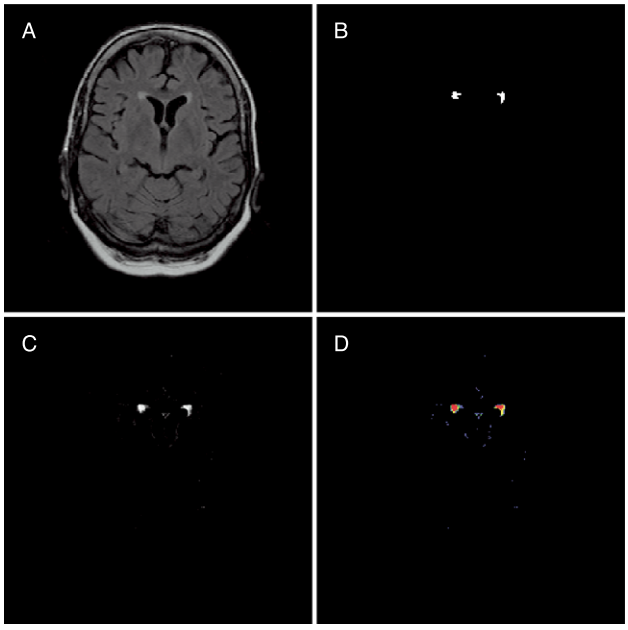

Fig. 2. Classification of a patient with small lesion load. (A) FLAIR image, (B) manual segmentation, (C) probability map, (D) segmentations derived from probability map with different thresholds: black: probability (P)=0, blue: $0<P\leq 0.3$, green: $0.3<P\leq 0.5$, yellow: $0.5<P\leq 0.8$, red: $0.8<P\leq 1$.

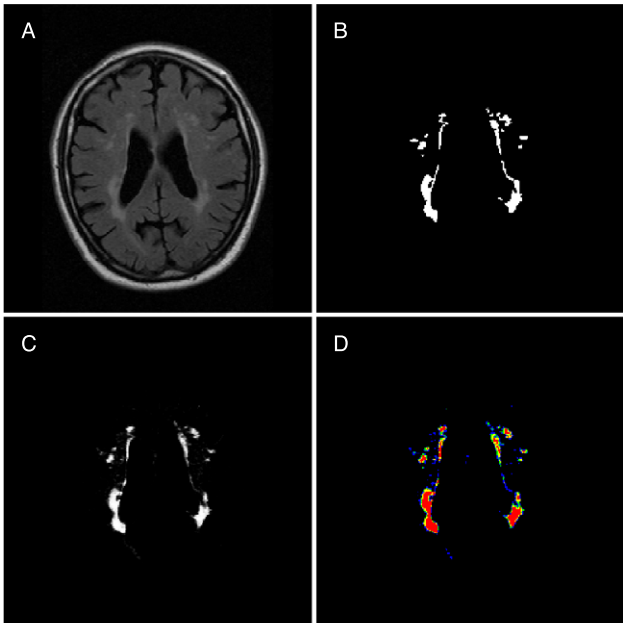

Fig. 3. Classification of a patient with moderate lesion load. (A) FLAIR image, (B) manual segmentation, (C) probability map, (D) segmentations derived from probability map with different thresholds: black: probability (P) = 0, blue: $0<P\leq 0.3$, green: $0.3<P\leq 0.5$, yellow: $0.5<P\leq 0.8$, red: $0.8<P\leq 1$.

In practice, for clarity of the evaluation, it is desirable to have a general measure, representing the accuracy of the probability map as a whole. Therefore, probabilistic versions of the similarity measures are calculated, which provide an opportunity to compare segmentation methods, in which probabilistic outcomes are evaluated by comparison with binary references. The probabilistic

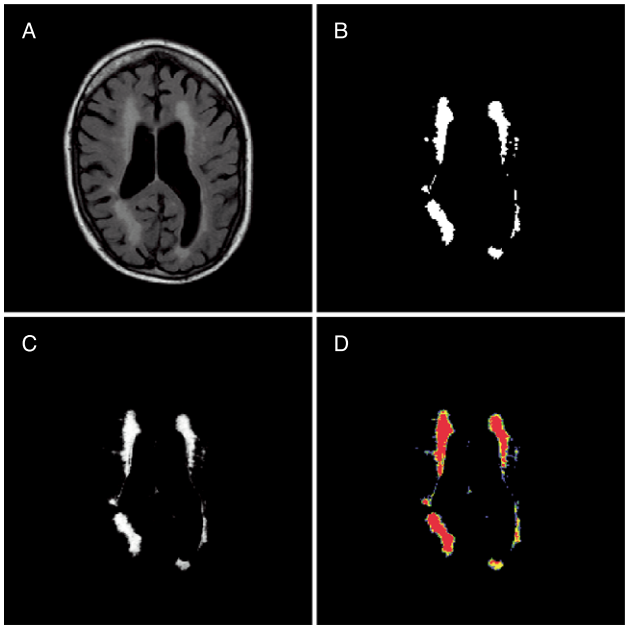

Fig. 4. Classification of a patient with large lesion load. (A) FLAIR image, (B) manual segmentation, (C) probability map, (D) segmentations derived from probability map with different thresholds: black: probability (P)=0, blue: $0<P\leq 0.3$, green: $0.3<P\leq 0.5$, yellow: $0.5<P\leq 0.8$, red: $0.8<P\leq 1$.

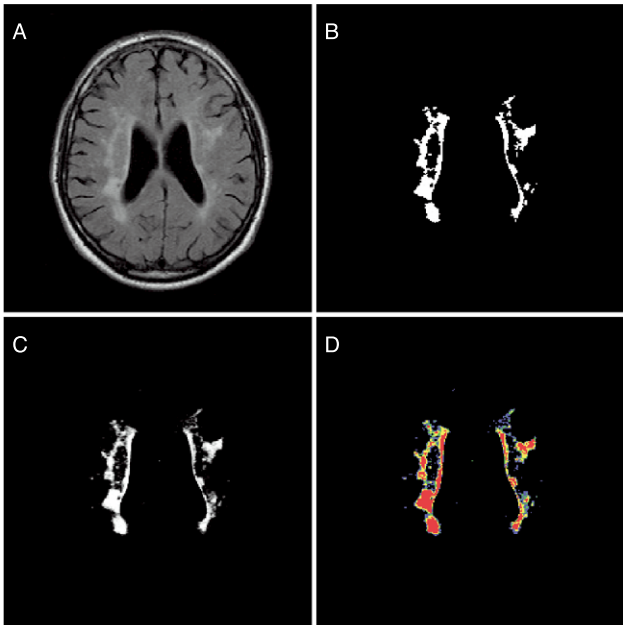

Fig. 7. Classification with probabilistic similarity index: 0.76. (A) FLAIR image, (B) manual segmentation, (C) probability map, (D) segmentations derived from probability map with different thresholds: black: probability (P)=0, blue: $0<P\leq 0.3$, green: $0.3<P\leq 0.5$, yellow: $0.5<P\leq 0.8$, red: $0.8<P\leq 1$.

similarity index (PSI), probabilistic overlap fraction (POF) and probabilistic extra fraction (PEF) are defined by

$$\mathrm{PSI} = \frac{2 \times \sum P_{x,\mathrm{gs}=1}}{\sum 1_{x,\mathrm{gs}=1} + \sum P_x},$$

$$\mathrm{POF} = \frac{\sum P_{x,\mathrm{gs}=1}}{\sum 1_{x,\mathrm{gs}=1}},$$

$$\mathrm{PEF} = \frac{\sum P_{x,\mathrm{gs}=0}}{\sum 1_{x,\mathrm{gs}=1}}.$$

where

- $\Sigma P_{x,\mathrm{gs}=1}$: sum over all voxel probabilities, where in the gold standard (manual segmentation), the intensity value = 1,
- $\Sigma P_{x,\mathrm{gs}=0}$: sum over all voxel probabilities, where in the gold standard, the intensity value = 0,
- $\Sigma 1_{x,\mathrm{gs}=1}$: sum over all voxels in the gold standard,
- ΣP_x: sum over all probabilities in the probability map.

The PSI and POF have values between 0 and 1, in which a high value resembles better correlation with the reference, and 1 denotes that the segmentation equals the gold standard. The PEF has values of 0 and higher and should remain as small as possible for a good segmentation. For example, a POF of 0.6 indicates that in the segmentation the total area of the gold standard has been classified with a value of 0.6, or that 60% of the reference area has been classified with probability 1 or any combination of both cases.

The probabilistic measures were calculated for all patient categories and all feature sets.

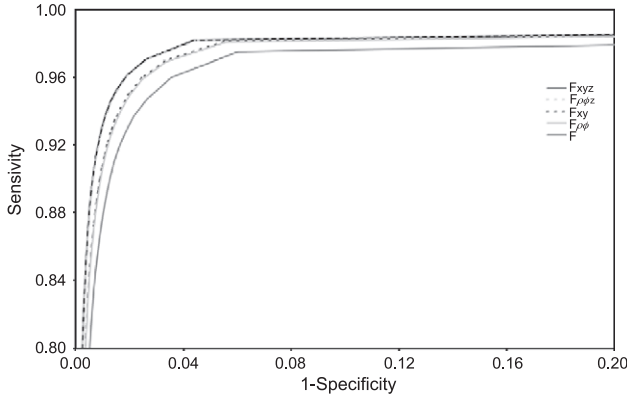

Fig. 5. ROC curves of classifications over all patients with different feature sets: (*Fxyz*) voxel intensity features and spatial features *x*, *y* and *z*; (*Fρφz*) voxel intensities and ρ, φ and *z*; (*Fxy*) voxel intensities and *x* and *y*; (*Fρφ*) voxel intensities and ρ and φ; (*F*) only voxel intensities.

Table 1
Area under the ROC curve

Feature set	All patients	Small lesion	Moderate lesion	Large lesion
F	0.9832	0.9575	0.9815	0.9845
Fρφ	0.9871	0.9759	0.9851	0.9874
Fρφz	0.9885*	0.9870	0.9865	0.9883
Fxy	0.9874	0.9765	0.9855	0.9877
Fxyz	0.9886*	0.9869	0.9868	0.9883

Note. For the all patients category, significance of feature sets including spatial features with respect to feature set *F* has been calculated.
* $P < 0.05$ (paired samples *t* test).

Statistics

To compare differences in feature sets for areas under the ROC curves, similarity measures (SI, OF and EF) and probabilistic similarity measures (PSI, POF and PEF), paired samples *t* tests were used. A $P < 0.05$ was considered as statistically significant. These analyses were only performed for the "all patient" category, and not for the subcategories of patients, because of the limited statistical power.

Results

KNN classification has been performed five times per patient, according to the five different feature sets. In Figs. 2–4, example images are shown of the classification results of patients, with a small, moderate and large lesion load, with feature set *Fxyz*. For each patient category, the following images are shown: FLAIR, manual segmentation, probability map and a color image with segmentations generated by applying thresholds of 0.3, 0.5 and 0.8 to the probability map. The images demonstrate that the choice of the threshold on the probability map has large influence on the binary segmentations. A higher threshold increases the specificity of the result, but has a negative effect on the sensitivity.

ROC analysis

The ROC curves of the five different feature sets were calculated for the classifications of all patient categories. Fig. 5 shows a detail of the ROC curves of all patients with thresholds running from 0 to 1. With a threshold, the sensitivity of the segmentations with feature set *Fxyz* reaches 0.9704, with a specificity of 0.9740. With the same threshold, the sensitivity with feature set *F* is 0.9654, with a specificity of 0.9640. The graph shows that the sensitivity and specificity of the other features sets are between these numbers. The areas under the curves have been calculated and are presented in Table 1. The ROC curves and the areas show that including spatial features by feature set *Fxyz* or *Fρφz* improves the results for all patients significantly with respect to feature set *F*, whereas the feature sets without the *z*-coordinate, *Fxy* and *Fρφ* do not increase the area significantly.

Similarity measures

Fig. 6 presents the SI for the segmentations of the category of all patients with the five different feature sets and the threshold

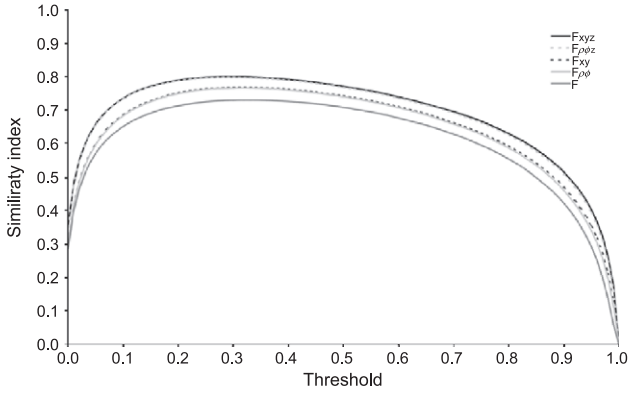

Fig. 6. Similarity index of binary WML segmentations of all patients as function of the threshold with different feature sets: (*Fxyz*) voxel intensity features and spatial features x, y and z; ($F\rho\varphi z$) voxel intensities and ρ, φ and z; (*Fxy*) voxel intensities and x and y; ($F\rho\varphi$) voxel intensities and ρ and φ; (*F*) only voxel intensities.

running from 0 to 1. Similar to the ROC curves, the SI graph shows that feature sets *Fxyz* and $F\rho\varphi z$ have the best performance. Fig. 6 also shows that the optimal threshold for the generation of a binary segmentation is approximately 0.3. Table 2 shows the SI, OF and EF of the binary segmentations with this optimal threshold for all patient categories and all feature sets. The SI improves significantly using feature sets with spatial features, whereas the OF only improves with feature sets *Fxyz* and $F\rho\varphi z$. No significant improvement of the EF was found when adding spatial features. For all feature sets, it was observed that in patients with a large lesion load better results are obtained. Furthermore, the table shows that feature sets *Fxyz* and $F\rho\varphi z$ improve the results to a similar extent.

Probabilistic similarity measures

Table 3 shows the probabilistic equivalents of the SI, OF and EF for all patient categories and all feature sets. Addition of spatial features results in a significantly higher PSI, for all feature sets. The POF is significantly better for feature sets *Fxyz* and $F\rho\varphi z$, with respect to *F*. No significant improvement of the PEF was found adding spatial features. Furthermore, the segmentations are better for patients with a large lesion load and feature sets *Fxyz* and $F\rho\varphi z$ improve the results to a similar extent.

Table 2
Similarity index, overlap fraction and extra fraction with threshold 0.3

Measure	Feature set	All patients	Small lesion	Moderate lesion	Large lesion
SI	*F*	0.73	0.33	0.70	0.80
	$F\rho\varphi$	0.77**	0.39	0.73	0.83
	$F\rho\varphi z$	0.80**	0.49	0.75	0.85
	Fxy	0.77**	0.40	0.73	0.83
	Fxyz	0.80**	0.50	0.75	0.85
OF	*F*	0.75	0.60	0.69	0.79
	$F\rho\varphi$	0.77	0.59	0.71	0.81
	$F\rho\varphi z$	0.79**	0.64	0.73	0.83
	Fxy	0.78	0.60	0.71	0.82
	Fxyz	0.79*	0.64	0.73	0.83
EF	*F*	0.31	1.99	0.28	0.18
	$F\rho\varphi$	0.25	1.44	0.25	0.15
	$F\rho\varphi z$	0.19	0.95	0.22	0.12
	Fxy	0.25	1.42	0.25	0.15
	Fxyz	0.19	0.92	0.22	0.12

Note. For the all patients category, significance of feature sets including spatial features with respect to feature set *F* has been calculated.
* $P < 0.05$ (paired samples t test).
** $P \leq 0.01$ (paired samples t test).

Table 3
Probabilistic similarity index, probabilistic overlap fraction and probabilistic extra fraction

Measure	Feature set	All patients	Small lesion	Moderate lesion	Large lesion
PSI	*F*	0.62	0.25	0.57	0.70
	$F\rho\varphi$	0.65**	0.28	0.60	0.72
	$F\rho\varphi z$	0.69**	0.35	0.63	0.75
	Fxy	0.65**	0.29	0.60	0.73
	Fxyz	0.69**	0.36	0.64	0.76
POF	*F*	0.60	0.48	0.54	0.64
	$F\rho\varphi$	0.63	0.47	0.57	0.66
	$F\rho\varphi z$	0.65**	0.51	0.59	0.69
	Fxy	0.63	0.48	0.57	0.67
	Fxyz	0.66**	0.51	0.59	0.69
PEF	*F*	0.96	2.87	0.90	0.82
	$F\rho\varphi$	0.94	2.37	0.91	0.83
	$F\rho\varphi z$	0.89	1.89	0.87	0.82
	Fxy	0.94	2.35	0.91	0.84
	Fxyz	0.90	1.86	0.87	0.83

Note. For the all patients category, significance of feature sets including spatial features with respect to feature set *F* has been calculated.
** $P \leq 0.01$ (paired samples t test).

Discussion

The combination of spatial information and signal intensities of MR images in KNN classification provides a technique for WML segmentation with a high sensitivity and specificity for all patient categories, which is shown by ROC curves. The method generates a probability map, containing the probabilities of voxels being a lesion. The main advantage of determination of the lesion probabilities over direct classification of voxels into lesion or non-lesion is that it provides an opportunity to obtain different binary segmentations, by which the ratio between sensitivity and specificity can be varied, dependent on the purpose of the segmentation. In this way, segmentations with better agreement with the reference can be produced. For example, the classification according to the *K*-nearest neighbor rule (Duda et al., 2001) would generate the binary segmentation of threshold 0.5, whereas the segmentation with threshold 0.3 has a higher similarity index. Furthermore, the probability map provides more detailed information on the lesion probability per voxel than a binary segmentation.

The proposed algorithm produces segmentations with high sensitivity and specificity. The ROC curve shows that, at a proper point in the curve, the overall sensitivity of the binary segmentations of all patients is 0.9704, with a specificity of 0.9740. However, these high numbers, and particularly the specificity, are not only a consequence of the segmentation, but also due to the small prior probability of the lesions. Therefore, the need arises to use similarity measures, relative to the lesion volume for better information on the quality of the segmentation. In this respect, the use of the similarity measures (SI, OF en EF) provides an opportunity to evaluate the segmentations in a quantitative and objective way. The

optimal similarity indexes of the categories of all patients and of patients with moderate and large lesion load are higher than 0.7 for all feature sets. An SI value of 0.7 resembles an excellent agreement according to Bartko (1991). For the class of patients with large lesion load these values even exceed the value of 0.8.

At present, many studies on MS brain lesion segmentation have been published that describe and evaluate the results in quantitative terms by comparison of lesion volume or by analysis of the number of correctly or misclassified lesions (Alfano et al., 2000; Goldberg-Zimring et al., 1998; Guttman et al., 1999; Jack et al., 2001; Wei et al., 2002). However, only a limited number of these studies evaluate also spatial correspondence of the results with a reference segmentation (Kamber et al., 1995; Van Leemput et al., 2001; Zijdenbos et al., 2002). Kamber calculates an error rate, which at least exceeds 1%. In the proposed method, this error rate is in all cases below 0.5%. Van Leemput and Zijdenbos also use the similarity index for measuring the correspondence with a manual segmentation, which reaches maxima of 0.51 and 0.68, respectively. The similarity index for all patients in the proposed method exceeds both of these numbers. To our knowledge, other methods for general WML segmentation only evaluate the results by visual inspection or by measurement of lesion volume (Jack et al., 2001; Mohamed et al., 2001; Wei et al., 2002).

The results of our study show that adding features containing spatial information to the feature space improves the segmentations significantly. The influence of Euclidian coordinates x and y appears to be similar to polar coordinates ρ and φ. Addition of the z-coordinate to the feature set is also essential for a better classification. Further improvement by spatial features might be achieved by indicating the exact location in the brain, by using a brain atlas as reference (Cocosco et al., 2002; Van Leemput et al., 1999a,b; Warfield et al., 2000; Wells et al., 1996).

The similarity measures also suggest that the proposed method produces better results for patients with a large lesion load than for patients with a small lesion load. This can be explained by the fact that small errors have a relatively larger effect on a smaller reference area.

The SI is not only used as a measure for the accuracy of the segmentation, but also for determination of the optimal threshold to generate a binary segmentation. For the "all patients" category, this threshold is approximately 0.3. The SI graphs are considerably flat, which indicates the robustness of the optimal threshold. This is supported by the fact that for thresholds running from 0.13 to 0.55 the SI is at least 95% of the maximum SI. The behavior of the optimal threshold in other situations and applications is difficult to predict. In general, it can be stated that higher agreement of the probability map with a binary segmentation, that is, more probabilities close to 0 or 1, increases the robustness of the optimal threshold. The location of the optimal threshold can in other applications be determined by performing some tests with the binary reference that always has to be available for the learning set. However, the choice of the threshold may in many cases also be an expert decision that depends on the acceptable ratio between false positive and false negative classified voxels.

Probabilistic similarity measures provide useful tools to evaluate a probabilistic segmentation directly, without being dependent on the generation of a binary segmentation. Valuable study on these measures has been performed by Zou et al. (2002). In our paper we applied the PSI, POF and PEF as probabilistic versions of the SI, OF and EF. The PSI always denotes a lower value than the corresponding SI with optimal threshold. This does not indicate a worse result, but is caused by the fact that the probabilistic classification outcome is compared to the binary gold standard. To illustrate the meaning of the probabilistic measures, and to provide a better intuitive understanding, an example of a classification with a relatively high PSI value (0.76) as outcome is presented in Fig. 7. The POF of this classification is 0.76 and the PEF is 1.01. Classification has been performed with feature set *Fxyz*. The images are: FLAIR, manual segmentation, probability map and the color image with the binary segmentations of the probability map with thresholds 0.3, 0.5 and 0.8. With threshold 0.3, the binary segmentation has a SI of 0.83, an OF of 0.90 and an EF of 0.26.

The gold standard in this method was constructed by consensus of some experts. Although this reference segmentation was satisfactory for our main purpose, other more sophisticated approaches, which estimate a ground truth from a group of expert segmentations, are also available and can be applied (Warfield et al., 2002).

In conclusion, the KNN approach offers valuable ways to perform automated WML segmentation. Moreover, since this probabilistic classification method has a general basis it is applicable to many other segmentation problems, for instance, segmentation of atrophy, white matter, gray matter or CSF. Finally, the method is fully automatic, can be performed on routine MR diagnostic scans, and is therefore suitable for detection and segmentation of WMLs in large and longitudinal population studies.

References

Alfano, B., Brunetti, A., Larobina, M., Quarantelli, M., Tedeschi, E., Ciarmiello, A., Covelli, E.M., Salvatore, M., 2000. Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis. J. Magn. Reson. Imaging 12, 799–807.

Bartko, J.J., 1991. Measurement and reliability: statistical thinking considerations. Schizophr. Bull. 17, 483–489.

Benson, R.R., Guttmann, C.R., Wei, X., Warfield, S.K., Hall, C., Schmidt, J.A., Kikinis, R., Wolfson, L.I., 2002. Older people with impaired mobility have specific loci of periventricular abnormality on MRI. Neurology 58, 48–55.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford Univ. Press, Oxford, UK.

Briley, D.P., Haroon, S., Sergent, S.M., Thomas, S., 2000. Does leukoaraiosis predict morbidity and mortality? Neurology 54, 90–94.

Cocosco, C.A., Zijdenbos, A.P., Evans, A.C., 2002. Automatic generation of training data for brain tissue classification from MRI. MICCAI 2002. 5th International Conference, Sept. 2002, Tokyo, Japan. Springer-Verlag, Berlin, pp. 516–523.

De Groot, J.C., de Leeuw, F.E., Oudkerk, M., Hofman, A., Jolles, J., Breteler, M.M., 2000a. Cerebral white matter lesions and depressive symptoms in elderly adults. Arch. Gen. Psychiatry 57, 1071–1076.

De Groot, J.C., de Leeuw, F.E., Oudkerk, M., van Gijn, J., Hofman, A., Jolles, J., Breteler, M.M., 2000b. Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study. Ann. Neurol. 47, 145–151.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. Wiley, New York, USA.

Goldberg-Zimring, D., Achiron, A., Miron, S., Faibel, M., Azhari, H., 1998. Automated detection and characterization of multiple sclerosis lesions in brain MR images. Magn. Reson. Imaging 16, 311–318.

Guttman, C.R.G., Kikinis, R., Anderson, M.C., Jakab, M., Warfield, S.K., Killiany, R.J., Weiner, H.L., Jolesz, F.A., 1999. Quantitative follow-up of patients with multiple sclerosis using MRI: reproducibility. J. Magn. Reson. Imaging 9, 509–518.

Jack, C.R., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Yuecheng, X., Muthupillai, R., Manduca, A., Avula, R., Erickson, B.J., 2001. FLAIR histogram segmentation for measurement of leukoaraiosis volume. J. Magn. Reson. Imaging 14, 668–676.

Kamber, M., Shinghal, R., Collins, D.L., Francis, G.S., Evans, A.C., 1995. Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. IEEE Trans. Med. Imag. 14, 442–453.

Longstreth, W.T., Manolio, T.A., Arnold, A., Burke, G.L., Bryan, N., Jungreis, C.A., Enright, P.L., O'Leary, D., Fried, L., 1996. Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people. The Cardiovascular Health Study. Stroke 27, 1274–1282.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imag. 16, 187–198.

Mantyla, R., Erkinjuntti, T., Salonen, O., Aronen, H.J., Peltonen, T., Pohjasvaara, T., Standertskjold-Nordenstam, C.G., 1997. Variable agreement between visual rating scales for white matter hyperintensities on MRI. Comparison of 13 rating scales in a poststroke cohort. Stroke 28, 1614–1623.

Mohamed, F.B., Vinitski, S., Gonzalez, C.F., Faro, S.H., Lublin, F.A., Knobler, R., EstebanGutierrez, J., 2001. Increased differentiation of intracranial white matter lesions by multispectral 3D-tissue segmentation: preliminary results. Magn. Reson. Imaging 19, 207–218.

Nyúl, L.G., Udupa, J.K., 1999. On standardizing the MR image intensity scale. Magn. Reson. Med. 42, 1072–1081.

Nyúl, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imag. 19, 143–150.

Smith, C.D., Snowdon, D.A., Wang, H., Markesbery, W.R., 2000. White matter volumes and periventricular white matter hyperintensities in aging and dementia. Neurology 54, 838–842.

Stokking, R., Vincken, K.L., Viergever, M.A., 2000. Automatic morphology-based brain segmentation (MBRASE) from MRI-T1 data. Neuroimage 12, 726–738.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999a. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imag. 18, 885–896.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999b. Automated model-based tissue classification of the brain. IEEE Trans. Med. Imag. 18, 897–908.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imag. 20, 677–688.

Warfield, S., Dengler, J., Zaers, J., Guttmann, C.R.G., Wells III, W.M., Ettinger, G.J., Hiller, J., Kikinis, R., 1995a. Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions. J. Image Guide Surg. 1, 326–338.

Warfield, S., Dengler, J., Zaers, J., Guttmann, C.R.G. , Wells III, W.M., Ettinger, G.J., Hiller, J., Kikinis, R. 1995. Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions. MRCAS'95. Second international Symposium on Medical Robotics and Computer Assisted Surgery, Nov. 1995, Baltimore, USA. John Wiley, Philadelphia, pp. 140–147.

Warfield, S.K., Kaus, M., Jolesz, F.A., Kikinis, R., 2000. Adaptive, template moderated, spatially varying statistical classification. Med. Image Anal. 4, 43–55.

Warfield, S.K., Zou, K.H., Wells, W.M., 2002. Validation of image segmentation and expert quality with an expectation-maximization algorithm. MICCAI 2002, 5th International Conference, Sept. 2002, Tokyo, Japan. Springer-Verlag, Berlin, pp. 298–306.

Wei, X., Warfield, S.K., Zou, K.H., Wu, Y., Li, X., Guimond, A., Mugler III, J.P., Benson, R.R., Wolfson, L., Weiner, H.L., Guttmann, C.R.G., 2002. Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. J. Magn. Reson. Imaging 15, 203–209.

Wells III, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A., 1996. Adaptive segmentation of MRI data. IEEE Trans. Med. Imag. 15, 429–442.

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method an validation. IEEE Trans. Med. Imag. 13, 716–724.

Zijdenbos, A.P., Forghani, R., Evans, A.C., 2002. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Trans. Med. Imag. 21, 1280–1291.

Zou, K.H., Wells III, W.M., Kaus, M.R., Kikinis, R., Jolesz, F.A., Warfield, S.K., 2002. Statistical validation of automated probabilistic segmentation against composite latent expert ground truth in MR imaging of brain tumors. MICCAI 2002, 5th International Conference, Sept. 2002, Tokyo, Japan. Springer-Verlag, Berlin, pp. 315–322.