Automated White Matter Lesion Segmentation by Voxel Probability Estimation

Petronella Anbeek, Koen Vincken, Matthias van Osch, Bob Bisschops, Max Viergever, and Jeroen van der Grond

Department of Radiology, Image Sciences Institute, University Medical Center, Heidelberglaan 100, rm E01.335, Utrecht, The Netherlands
{nelly,koen,thijs,bob,max, jeroen}@isi.uu.nl

Abstract. A new method for fully automated segmentation of white matter lesions (WMLs) on cranial MR imaging is presented. The algorithm uses five types of regular MRI-scans. It is based on a k-Nearest Neighbor (KNN) classification technique, which 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 are produced. ROC-curves show that the segmentations achieve high sensitivity and specificity. The similarity index (SI) is used for further analysis and for determination of the optimal threshold. The probabilistic equivalent of the SI allows direct evaluation of the probability maps, which provides a strong tool for comparison of different classification results. This method for automated WML segmentation reaches an accuracy that is comparable to methods for multiple sclerosis lesion segmentation.

1 Introduction

In the last decade much attention has been paid on cerebral white matter lesions in the elderly or in patients with cardiovascular risk factors. In these patients WMLs are a common finding on cranial MR imaging [1]. WMLs are associated with age, clinically silent stroke, higher systolic blood pressure lower forced expiratory volume in one second, hypertension, atrial fibrillation, carotid and peripheral arterioscleroses, impaired cognition or depression [2,3]. Furthermore, it has been shown that stroke patients with a high WML load have an increased risk of hemorrhagic transformation, higher preoperative risk of a disabling or fatal stroke during endartectomy or intercerebral hemorrhage during anticoagulation therapy [4]. The increased interest in WML research, may improve diagnosis and prognosis possibilities for patients with cardiovascular symptoms. In this respect it would be highly advantageous to use an automated segmentation method that detects WMLs with a high sensitivity and specificity. Such methods have been developed for the detection of multiple sclerosis (MS) lesions but not for WMLs in general, which is more complicated because of the more heterogeneous nature of WMLs.

The aim of the present study is to develop an automated WML segmentation method, based on a supervised KNN-classification technique using multi-spectral information from T1-weighted (TI-w), T1-weighted inversion recovery (IR), proton

R.E. Ellis and T.M. Peters (Eds.): MICCAI 2003, LNCS 2878, pp. 610–617, 2003.

density-weighted (PD), T2-weighted (T2-w) and Fluid Attenuation Inversion Recovery (FLAIR) scans by voxel probability estimation, that is suitable for large population studies.

2 Methods

2.1 MR Imaging and Patients

MRI studies were performed on a Philips Gyroscan ACS-NT 15 whole body system operating at 1.5 Tesla (Philips Medical Systems, Best, The Netherlands). All patients had the same MR protocol of the brain consisting T1-w, IR, T2-w, PD and FLAIR scans. All scans were performed with a 4 mm slice thickness, no slice gap, 38 slices, a 230 x 230 mm field of view and a 256 x 256 scan matrix. The individual scan parameters were: T1-w (FFE): repetition time (TR)/ echo time (TE) 234/2 ms, IR: TR/ inversion time (TI)/TE 2919/410/22 ms, T2-w: TR/TE 2200/100 ms, PD: TR/TE 2200/11 ms and FLAIR: TR/TI/TE 6000/2000/100 ms.

Twenty patients with arterial vascular disease (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 (65.6 ± 7.7, range 49-75), 17 patients were male.

2.2 Manual Segmentation and Image Preprocessing

The WMLs were manually segmented by the first author. WMLs had to be hyperintense on FLAIR, PD and T2-w images. According to the patterns of WMLs, four patient categories were composed: (1) all patients (n=20), (2) patients with low lesion load (n=8), (3) patients with moderate lesion load (n=7), (4) patients with large lesion load (n=5). The manual segmentations were in two steps independently reviewed and corrected by two investigators. The final manual WML segmentation was reevaluated in a consensus meeting and considered as gold standard.

To correct for MR inhomogeneities a method was used, which resulted in similar gray values of major anatomical structures in different patients per image type [5,6]. To correct for differences due 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 [7]. To reduce the number 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-weighted image of every patient [8].

2.3 Voxel Classification

The aim of the method for automatic segmentation of the WMLs was to determine the lesion probability per voxel. For this purpose a KNN-classification method was used by which a case (in our study a voxel) is classified dependent on its feature values. The learning set for segmentation of one patient was built from the voxels of the other

19 patients. All voxels in the learning set were labeled with the value of 0 (non-lesion) or 1 (lesion) according to the manual segmentations.

The features used in this study can be divided into two categories: Voxel intensities and spatial information. The first group is defined by the gray values of a voxel in the available images: T1-w, IR, PD, T2-w and FLAIR, after the preprocessing steps described above. Using only these features provides a 5-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 by two different methods. The Euclidean coordinates (x and y) and the polar coordinates (ρ and φ), measured from the center of gravity in the FLAIR image, which was the reference image for registration, were used separately. Coordinate ρ was the Euclidean distance from the center of gravity and φ the angle with the horizontal axis. Through-plane a spatial feature denoting the slice number z was included.

All experiments were performed with five different feature sets: (F) only voxel intensities, (Fxy) voxel intensities and spatial features x and y, (Fxyz) voxel intensities and x, y and z, (Frp) intensities and ρ and φ, (Frpz) intensities and ρ, φ and z.

Because all features had different ranges, they had to be rescaled in order to achieve similar significance for every feature in the classification. This was achieved by variance scaling: Subtraction of the mean from the feature values and division of the outcome by the standard deviation. This approach provided for every feature a mean of 0 and variance of 1.

The choice of k in KNN-classification depends on the number of features and the number of cases. With a low value of k the result is more influenced by individual cases. A higher value of k smoothens the outcome of the classification. In this study we used a relatively low number of features in combination with an extremely high number of cases. Therefore we choose for a large k. Experimentally it was observed that for this task a higher value than 100 has a marginal influence on the accuracy of the classification. By taking computation time into account it was concluded that 100 was the most appropriate choice for k.

For every voxel the probability that it was lesion was defined as the fraction of lesion voxels within the k neighbors of the examined voxel in the feature space. A new image was constructed from the voxel probabilities, which is called the probability map. By further analysis of the probability map the decision was made whether the voxel was classified as lesion or non-lesion.

2.4 Evaluation

By applying different thresholds on the probability map, binary segmentations of the WMLs were produced. These outcomes were compared with the gold standard, where the amount of correctly classified pixels, i.e. 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 were computed for every threshold of the probability map. The TPF was represented in an ROC-curve as function of the FPF for the category of all patients and all five feature sets.

Furthermore the binary segmentations were evaluated by the SI [9], which is a measure for the correctly classified lesion area, relative to the total area of WML in

the reference (= the gold standard) as well as the area of the segmentation. It is defined by

$$SI = \frac{2 \times TP}{2 \times TP + FP + FN} . \quad (1)$$

This measure was represented in a graph as function of the threshold for all feature sets and all thresholds.

In practice the probability of voxels being lesion might be more useful than the binary segmentations generated by applying thresholds. Also for clarity of the evaluation it is desirable to have a general measure, representing the accuracy of probability map as a whole. Therefore a probabilistic version of the similarity index was also computed. The probabilistic similarity index (PSI) is defined by

$$PSI = \frac{2 \times \sum P_{x,\,gs\,=\,1}}{\sum 1_{x,\,gs\,=\,1} + \sum P_x} . \quad (2)$$

With:

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

3 Results

KNN-classification has been performed on each patient with the five different feature sets. Figure 1 shows an example image of the classification result of a patient with feature set Fxyz. The presented images are: FLAIR, manual segmentation, probability map and the 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. By analysis of the SI the optimal threshold for this situation can be determined.

3.1 ROC-Curves

The ROC-curves were computed for the classifications with the five different feature sets of the category of all patients. The areas under the curves have been computed and are presented in table 1. These areas appear to be relatively high. This result is mainly due to the high specificity, which is caused by the low prior probability of the lesion voxels. From the ROC-curves and the areas can be concluded that the feature sets including spatial features x, y and z or ρ, φ and z perform better than the feature set without spatial features.

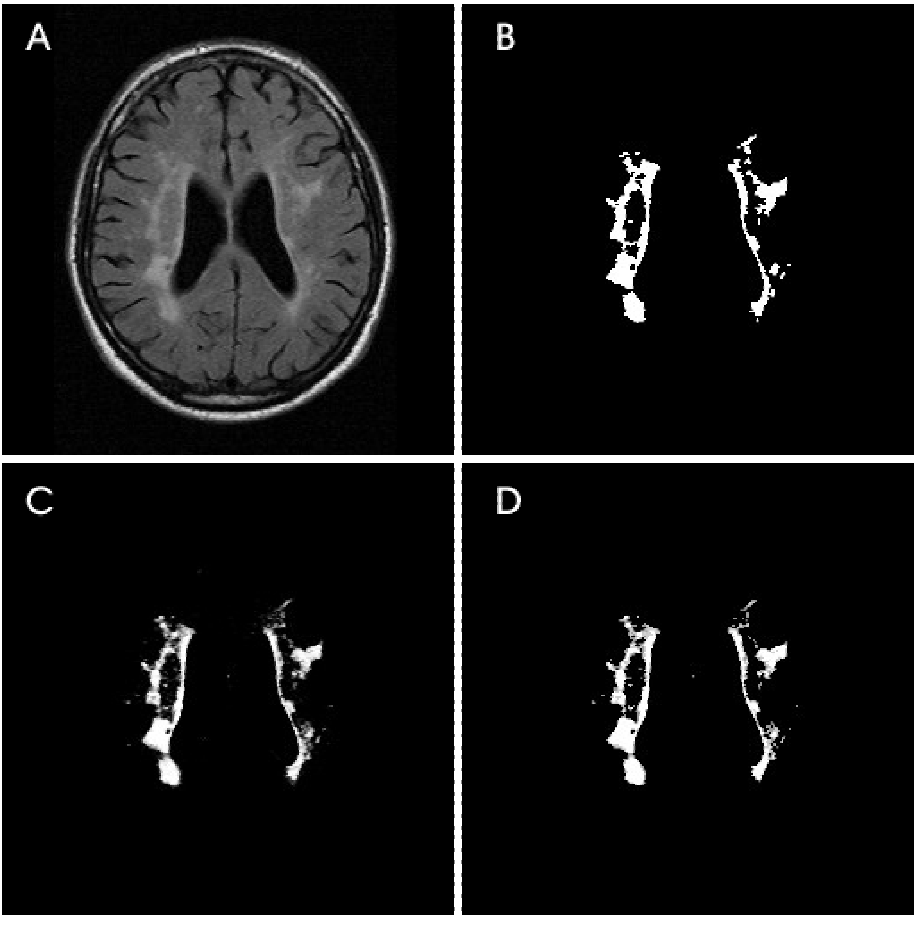

Fig. 1. WML classification and segmentation. (A) FLAIR image, (B) manual segmentation, (C) probability map, (D) segmentations of probability map with thresholds: black: probability (p) < 0.3, dark gray: 0.3 < p <= 0.5, light gray: 0.5 < p <= 0.8, white: 0.8 < p <= 1.

Table 1. Area under the ROC-curve

Feature set	All patients	Few lesion	Moderate lesion	Large lesion
F	0.9832	0.9575	0.9815	0.9845
Frp	0.9871	0.9759	0.9851	0.9874
Frpz	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

3.2 Similarity Index

Figure 2 shows the SIs for the segmentation with thresholds running from 0 to 1. The graph presents the results of classification with the five different feature sets: (1) F: only voxel intensities, (2) Frp: voxel intensities with ρ and φ, (3) Frpz: voxel intensities with ρ, φ an z, (4) Fxy: voxel intensities with x and y, (5) Fxyz: voxel intensities with x, y and z.

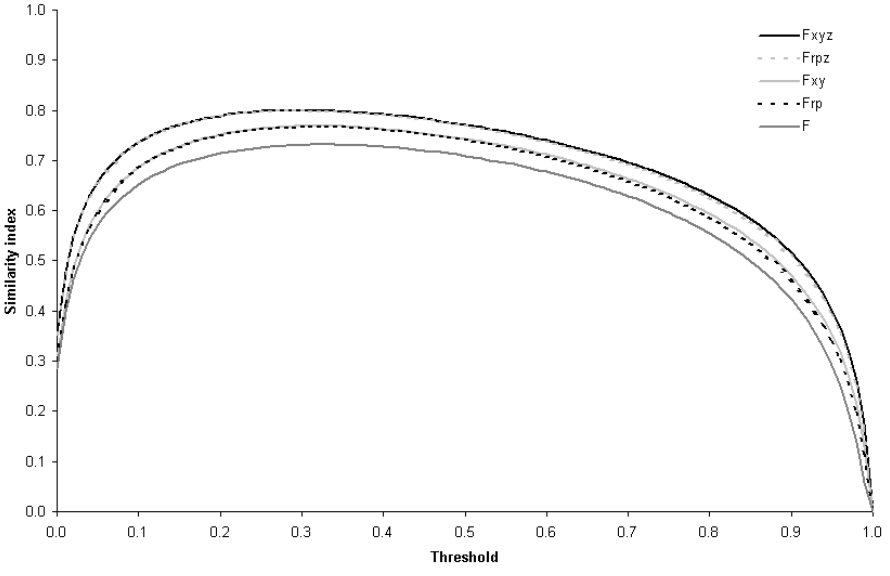

Fig. 2. Similarity indexes of classifications as function of the threshold for different feature sets, F: only voxel intensities; Frp: voxel intensities with ρ and φ; Frpz: voxel intensities with ρ, φ an z; Fxy: voxel intensities with x and y; Fxyz: voxel intensities with x, y and z.

The graph shows that the feature sets, including spatial features x, y and z or ρ, φ and z, have the best performance. x, y and z turn out to be equivalent to ρ, φ and z. As a conclusion from this graph the optimal threshold for the generation of a binary segmentation is set at 0.3. Table 2 shows the SI of the segmentation with the optimal threshold and the PSI for all patient categories and all feature sets.

Table 2. Similarity index with threshold 0.3 (probabilistic similarity index)

Feature set	All patients	Few lesion	Moderate lesion	Large lesion
F	0.73 (0.62)	0.33 (0.25)	0.70 (0.57)	0.80 (0.70)
Frp	0.77 (0.65)	0.39 (0.28)	0.73 (0.60)	0.83 (0.72)
Frpz	0.80 (0.69)	0.49 (0.35)	0.75 (0.63)	0.85 (0.75)
Fxy	0.77 (0.65)	0.40 (0.29)	0.73 (0.60)	0.83 (0.73)
Fxyz	0.80 (0.69)	0.50 (0.36)	0.75 (0.64)	0.85 (0.76)

The SI and the PSI are both optimal for feature sets including spatial features. The in-plane features x and y or ρ and φ, as well as slice number z, give improvement on the results. All patient categories benefit from the addition of spatial features, although in patients with few lesion loads they are most effective. For all feature sets holds that in patients with larger lesion load better results are achieved.

4 Discussion

The combination of spatial information and gray values of MR images in KNN-classification provides a strong technique for WML-segmentation with a high accuracy. The method produces a probability map, which contains more valuable information about the state of the lesions and the total lesion volume than a binary segmentation. By applying thresholds on the probability map different binary segmentations can be obtained, by which the sensitivity and specificity can be varied, dependent on the purpose of the segmentation.

ROC-curves show that this method reaches a high accuracy. For example, this result is shown by the area under the ROC-curve and is achieved for all amounts of lesion load. In a subgroup of patients with an average WML volume of 30.3 ml this method reaches a mean sensitivity of 81.2 % with a mean FPF of 0.0035. This is equivalent to the results of the method described by Alfano et al. [10] for segmentation of MS lesions in a group of patients with an average WML volume of 31.0 ml.

Nevertheless, for a better insight and for comparison with other techniques a different evaluation method is necessary. Investigation of the SI is suitable to evaluate the segmentations in a quantitative and objective way. From the SI can be concluded that the method has better performance for large lesions then for small lesions. The cause of this lies in the amount of false positive classified pixels. Since small errors have a relatively large effect on a small reference area, this method will give poorer evaluation result in small lesions. Furthermore the SI shows that adding features containing spatial information improves the result substantially. In particular the use of spatial features is beneficial in the segmentation of small lesions. The influence of Euclidian coordinates x and y is approximately equivalent to polar coordinates ρ and φ. Moreover, the slice number z has a significant contribution to the accurateness of the classification. Further improvement by spatial features might be achieved by denoting the exact location in the brain by features. Usage of a brain atlas for reference is a possible solution for this.

The SI is also useful for determination of an optimal threshold. The graphs of the SI show that there are different optimal thresholds for different patient categories. For category 2 it is approximately 0.5, for category 3 and 4 approximately 0.3. Therefore in this situation it might be useful to let the threshold for the final segmentation depend on the lesion load measured from the probability map. However, inspection of the SI of the individual patients shows that the optimal threshold is more dependent on the accuracy of the overall classification than on the lesion load.

Furthermore, the necessity arises to evaluate the probability map directly without being dependent on application of a threshold. This is the case when only the probability map is inspected or when the method is compared with other methods. For this purpose the probabilistic similarity index is introduced, analogously to SI. The PSI is a number for comparison of the probability map with the gold standard in which vox-

els classified with a higher chance to be lesion resembles a higher value when the gold standard denotes a 1. The PSI always has lower values than the corresponding SI with optimal threshold. This doesn't indicate a worse result, but is caused by the fact that it gives an overall view over the probability map.

In conclusion, the KNN-approach offers excellent ways to perform automated WML-segmentation. Moreover, since the method has a general basis it is applicable to many other segmentation problems, for instance segmentation of atrophy, white matter, gray matter or CSF.

References

1. Longstreth, W.T., et al.: Clinical Correlates of White Matter Findings on Cranial Magnetic Resonance Imaging of 3301 Elderly People. Stroke **27** (1996) 1274–1282
2. de Groot, J.C., et al.: Cerebral White Matter Lesions and Depressive Symptoms in Elderly Adults. Arch Gen Psychiatry **57** (2000) 1071–1076
3. de Groot, J.C., et al.: Cerebral White Matter Lesions and Cognitive Function: the Rotterdam Scan Study. Ann Neurol **47** (2000) 145–151
4. Briley, D.P., et al.: Does Leukoaraiosis Predict Morbidity and Mortality? Neurology **54** (2000) 90–94
5. Nyúl, L.G., et al.: On Standardizing the MRI Intensity Scale. MRM **42** (1999) 1072-1081
6. Nyúl, L.G., et al.: New Variants of a Method of MRI Scale Standardization. IEEE TMI **19** (2000) 142–150
7. Maes, F., et al.: Multimodality Image Registration by Maximization of Mutual Information. IEEE TMI **16** (1997) 187–198
8. Stokking, R., et al.: Automatic Morphology-Based Segmentation (MBRASE) from MRI-T1 Data. NeuroImage **12** (2000) 726–738
9. Zijdenbos, A.P., et al.: Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Transactions on Medical Imaging **13** (1994) 716–724
10. Alfano, B., et al.: Automated Segmentation and measurement of Global White Matter Lesion Volume in Patients With Multiple Sclerosis. JMRI **12** (2000) 799–807