Automated MS-Lesion Segmentation by K-Nearest Neighbor Classification

Petronella Anbeek, Koen L. Vincken and Max A. Viergever

July 14, 2008

Image Sciences Institute
University Medical Center Utrecht
Heidelberglaan 100, rm QS.459, 3584 CX Utrecht, The Netherlands

Abstract

This paper proposes a new method for fully automated multiple sclerosis (MS) lesion segmentation in cranial magnetic resonance (MR) imaging. The algorithm uses the T1-weighted and the fluid attenuation inversion recovery scans. It is based the K-Nearest Neighbor (KNN) classification technique. The data has been acquired at the Children's Hospital Boston (CHB) and the University of North Carolina (UNC). Manual segmentations, composed by a human expert of the CHB, were used for training of the KNN-classification. The method uses voxel location and signal intensity information for determination of the probability being a lesion per voxel, thus generating probabilistic segmentation images. By applying a threshold on the probabilistic images binary segmentations are derived. Automatic segmentations were performed on a set of testing images, and compared with manual segmentations from a CHB and a UNC expert rater. Furthermore, a combined segmentation was composed from segmentations from different algorithms, and used for evaluation. The proposed method shows good resemblance with the segmentations of the CHB rater. High specificity and lower specificity has been observed in comparison with the combined segmentations. Over- and undersegmentation can be easily corrected in this procedure by varying the threshold on the probabilistic segmentation image. The proposed method offers an automated and fully reproducible approach that accurate and applicable on standard clinical MR images.

1 Introduction

Multiple sclerosis (MS) is a disease of the central nervous system, affecting brain tissue, and leading to various symptoms such as motor impairment and neuropsychological problems. MS can be observed in magnetic resonance imaging (MRI) as visible lesions in the brain. Research to this disease is highly important for better understanding of causes and progression of the disease, and improvement of treatment. MRI plays an important role in MS research, often for determining size and location of affected tissue. An accurate and reproducible method for MS-lesion measurement is highly beneficial in large and longitudinal studies, in order to compare groups. Several methods for MS or other brain lesion segmentation have been developed recently [1-4]. Existing methods are based on different types on MRI scans, different underlying mathematics, and have different methods for evaluation. It would be advantageous for MS research in general, to compare different segmentation methods. The proposed method for MS-lesion segmentation is based on the automatic brain tissue segmentation method that was developed by Anbeek et al. [5]. The technique uses K-Nearest Neighbor (KNN) classification, which is a

statistical pattern classification technique. Similar to the brain tissue segmentation, this method generates probabilistic and binary segmentation images, but has especially been developed for segmentation of MS-lesions.

2 Methods

MS-lesion training and testing data

MR image sets of 45 patients were provided by two separate sources: 25 images sets from Children's Hospital Boston (CHB) and 20 from University of North Carolina (UNC). The UNC cases were acquired on a Siemens 3T Allegra MRI scanner with slice thickness of 1 mm and in-plane resolution of 0.5 mm. No scanner information was provided about the CHB cases. The complete set of images of one patient consisted of a T1-weighted (T1), a T2-weighted (T2), fluid attenuated inversion recovery (FLAIR) image, and diffusion tenser images: the fractional anisotropy map (DTI_FA) and the mean diffusivity map (DTI_MD). All data of one patient were rigidly registered to a common reference frame and resliced to isotropic voxel spacing, with resolution 512x512x512, using B-spline based interpolation. The image sets were randomly divided into training sets for MS-lesion segmentation (10 from each source), and testing sets (15 from CHB, 10 from UNC). One set of the CHB testing data was discarded, because of poor image quality, resulting in 14 CHB testing image sets.

Manual segmentations and gold standard

MS-lesion segmentations of all image sets were created manually by two raters, one from CHB, and one from UNC. Manual segmentations of all training image sets made by the CHB rater were provided, as well as manual segmentations of only the UNC training image sets made by the UNC rater. We decided to focus on the segmentations of only the CHB rater, because of inter rater differences, and since this was a complete set of manual segmentations. The manual segmentations of the testing image sets have not been provided on beforehand to the teams joining the competition. They were used in a later stage as a reference for evaluation of the automatic segmentations of the testing data.

Image preprocessing

Coregistration of the images was already performed before they were provided. Consequently, we performed one preprocessing step, concerning the creation of a brain mask, indication the region of interest for the segmentation. This reduces the amount of voxels being processed, thus saving computer time and memory. The mask was created by applying the brain extraction tool [6] on the T1 image with a relatively high value for the fractional threshold (-f). This procedure resulted in a narrow brain mask, consisting of brain tissue only. We have observed empirically that this narrow mask gave good performance of the segmentation method.

K-Nearest Neighbor classification

The proposed method is based on K-Nearest Neighbor (KNN) classification, and determines for every voxel in the image the probability that it is part of MS-lesion tissue. K-Nearest Neighbor classification is a statistical pattern recognition method, assigning samples (image voxels) to a class (MS-lesion) by searching for samples in a learning set with similar values in some measurable features. A feature space is defined, in which each axis represents one of the voxel features. The learning set consists of preclassified samples, which are entered into the feature space according to their feature values. A new image voxel is classified by comparing it to a number of K learning samples with smallest Euclidian distance to it in the feature space. Commonly, the most frequent class among the K learning samples is assigned to this voxel. However, our method does not assign one class to the voxel, but determines the probability per voxel being part of a lesion.

Previous research has shown that the FLAIR image contains most distinctive information for segmentation of white matter lesions [5]. Since MS-lesion tissue is a kind of white matter lesion, and its signal intensity is comparable with white matter lesion signal intensity in MR images, we have chosen to use only the FLAIR image in our KNN-segmentation method.

The proposed KNN-classification method uses two types of features: spatial and intensity features. The first group represents the voxel location in the brain that is uniquely defined by the x-, y- and z-coordinates in the image. Therefore, these three coordinates are used as three features, providing three dimensions of the feature space. Second, the signal intensity of a voxel in the FLAIR image denotes the last feature, resulting in a four-dimensional feature space.

As different features have different ranges, the feature space was rescaled to define a proper metric to compare distances. This was achieved by variance scaling: for each feature, the mean of the feature values was subtracted from the voxel value, and the outcome was divided by the standard deviation. This approach resulted in a mean of 0 and variance of 1 for all features. Since the spatial features were also normalized by this method, this implicitly corrects for differences in size and location between the patients.

The choice of variable K in KNN-classification is dependent on the relation between the number of features and the number of cases. A small K will cause the result being influenced by individual cases, while a large value of K makes the classification outcome smoother. In general, for this type of problems a large K is favorable [7, 8]. By performing experiments on the training set with different K-values, K = 40 was chosen. The decision for this choice was made by visual inspection of the images in the training set. A larger K did not improve the results appreciably, but had a negative effect on the computational efficiency.

Separate training sets were composed for segmentation of the CHB and the UNC testing sets, from the CHB and UNC training patients respectively. From the CHB training set patients 4, 5 and 9 were excluded, due to image and manual segmentation quality. For the UNC training set only the manual segmentations of the CHB rater were used. Patients 1, 5 and 6 of the UNC training data were excluded for similar reasons. Because of the large number of voxels in the training sets only 5 percent of the samples were randomly selected, and inserted in the feature space. This reduced computation time and computer memory significantly.

The voxel probability for being an MS-lesion was defined as the fraction of lesion voxels amongst the K closed neighbors in the feature space. The outcomes were represented in an image, showing the lesion probability per voxel, called the probability map. Subsequently, by applying thresholds with values between 0 and 1 to the probability maps, binary segmentations of the tissue types were derived. By varying the thresholds, different binary segmentations were generated: a low threshold produces a relatively wide segmentation, which becomes tighter and more specific when the threshold is increased.

For determination of the optimal value for the threshold, probabilistic segmentations were created with the KNN-method from the images in the training set. From these probability maps binary segmentations were derived with different thresholds. These binary segmentations were compared with the provided manual segmentations visually and by calculation of the Tanimoto coefficient. This procedure resulted in an optimal threshold of 0.4, which was applied on all probabilistic segmentations of the testing sets.

3 Results

Probabilistic segmentations of the MS-lesions have been generated for all CHB and UNC testing patients. An example of the segmentation is represented in figure 1, showing the FLAIR image, the probability map, and a binary segmentation that is derived by applying a threshold of 0.4 on the probability map.

All binary segmentations of the testing images have been compared with manual segmentations of the two raters: a UNC rater and a CHB rater. Furthermore, a STAPLE segmentation was composed from all submitted segmentations of the competition [9]. The segmentation of our method was also compare with this combined STAPLE segmentation. The results have been represented in table 1. This table shows as measures for the comparison with the UNC and CHB raters:

- Volume diff. (volume difference): absolute percent volume difference to the expert rater segmentation.
- Avg. Dist. (average distance): the absolute percent volume difference to the expert rater segmentation.
- True Pos. (true positive rate): percentage of the number of lesions in our segmentation that overlap with a lesion in the expert segmentation.
- False Pos. (false positive rate): percentage of the number of lesions in our segmentation that don’t overlap with a lesion in the expert segmentation.

All measures have been scored in relation to how the expert raters compare against each other. A score of 90 for any of the metric indicates a comparable performance with an expert rater.

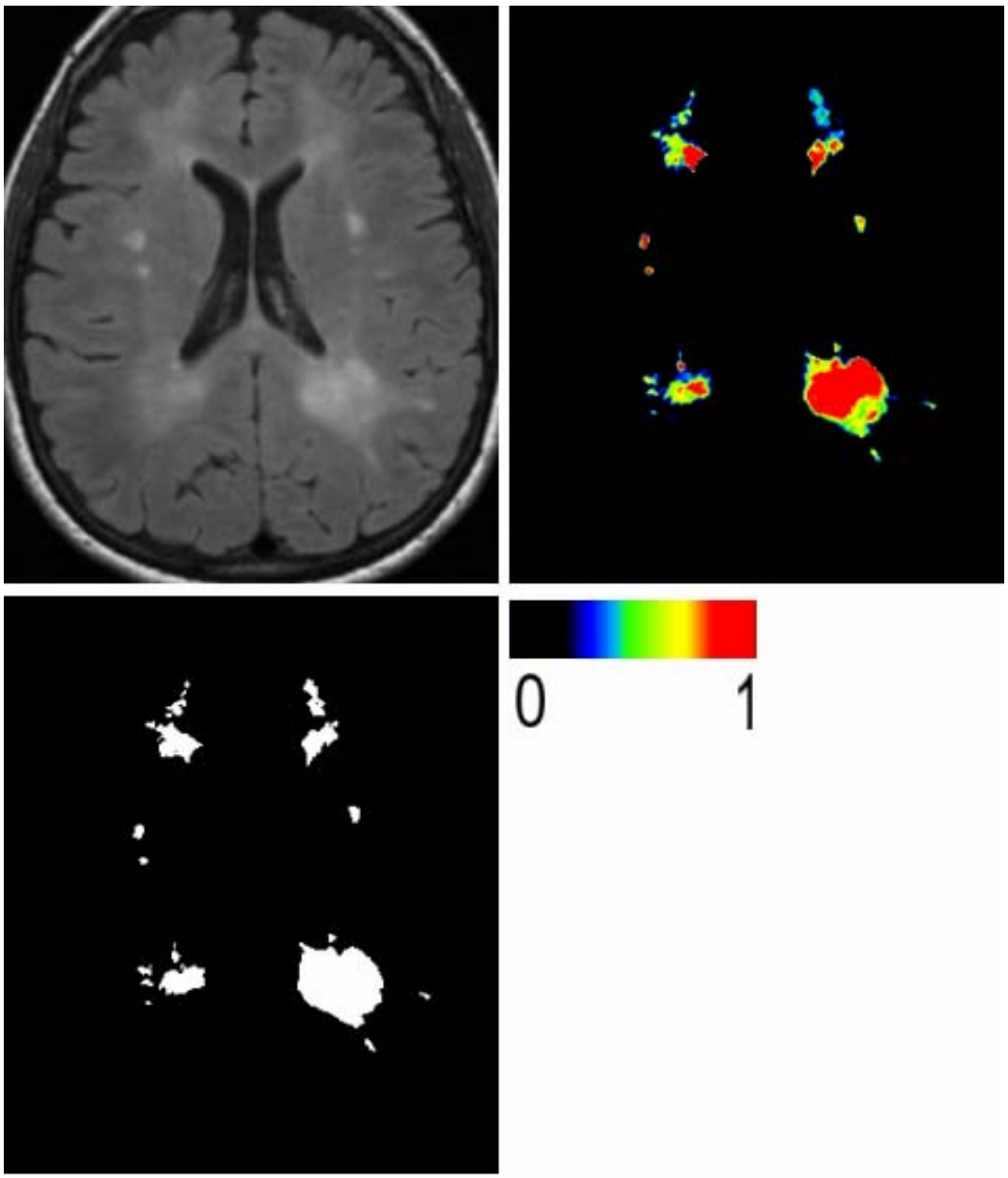

Figure 1 MS-lesion segmentation results. Top left: FLAIR image; top right: probabilistic segmentation, showing probability of lesion per voxel (see color bar); down left: binary segmentation, derived from probabilistic segmentation with threshold 0.4.

Ground Truth	UNC Rater								CHB Rater									STAPLE		
All Dataset	Volume Diff.		Avg. Dist.		True Pos.		False Pos.		Volume Diff.		Avg. Dist.		True Pos.		False Pos.		Total	Specificity	Sensitivity	PPV
	[%]	Score	[mm]	Score	[%]	Score	[%]	Score	[%]	Score	[mm]	Score	[%]	Score	[%]	Score				
UNC test1 Case01	47.7	93	6.4	87	53.5	82	77.4	63	23.6	97	3.7	92	62.5	87	74.5	64	83	0.9977	0.5047	0.9107
UNC test1 Case02	207.9	70	4.4	91	58.8	85	74.1	64	59.3	91	2.0	96	36.4	72	17.2	99	84	0.9969	0.6896	0.9710
UNC test1 Case03	36.9	95	5.0	90	43.7	76	82.7	59	18.4	97	4.8	90	47.1	78	81.3	60	81	0.9829	0.5708	0.6483
UNC test1 Case04	31.4	95	3.3	93	44.7	77	43.2	83	6.5	99	1.7	97	59.3	85	48.6	80	89	0.9984	0.8135	0.9766
UNC test1 Case05	75.3	89	7.5	85	31.0	69	57.1	75	44.2	94	5.4	89	60.9	86	61.2	72	82	0.9998	0.3114	0.9809
UNC test1 Case06	85.4	87	9.6	80	34.5	71	67.5	69	35.0	95	15.5	68	50.0	80	94.0	52	75	0.9976	0.1545	0.8053
UNC test1 Case07	64.0	91	5.5	89	41.0	75	49.4	80	16.4	98	1.9	96	70.0	91	64.9	70	86	0.9970	0.4777	0.8607
UNC test1 Case08	80.1	88	7.6	84	34.0	71	40.7	85	67.5	90	1.7	96	77.8	96	40.7	85	87	1.0000	0.2696	1.0000
UNC test1 Case09	11.3	98	45.0	7	0.0	51	100.0	49	25.2	96	53.7	0	0.0	51	100.0	49	50	0.9999	0.4324	0.9892
UNC test1 Case10	50.2	93	20.9	57	15.0	60	90.6	54	80.0	88	17.8	63	50.0	80	90.6	54	69	1.0000	0.2952	1.0000
CHB test1 Case01	12.0	98	6.2	87	40.0	74	91.2	54	60.0	91	6.2	87	74.2	94	92.4	53	80	0.9891	0.4956	0.5355
CHB test1 Case02	246.3	64	7.4	85	72.7	93	96.7	51	47.6	93	4.8	90	73.7	93	84.9	58	78	0.9355	0.6126	0.3271
CHB test1 Case03	129.8	81	11.6	76	71.4	92	96.0	51	11.0	98	9.1	81	66.7	89	93.3	53	78	0.9726	0.4934	0.2191
CHB test1 Case04	647.8	5	22.9	53	90.9	100	98.9	49	259.7	62	16.0	67	88.9	100	91.2	54	61	0.8731	0.4048	0.1427
CHB test1 Case05	84.6	88	9.3	81	40.7	75	98.1	50	65.0	90	2.5	95	78.3	96	79.5	61	79	0.9916	0.1954	0.5522
CHB test1 Case06	9.0	99	5.2	89	36.1	72	96.8	51	5.0	99	5.1	89	31.8	70	97.2	50	77	0.9696	0.4274	0.4992
CHB test1 Case07	88.3	87	5.7	88	58.3	85	95.5	52	14.5	98	2.7	94	68.4	90	93.6	53	81	0.9658	0.6784	0.6013
CHB test1 Case08	99.7	85	10.1	79	92.6	100	93.4	53	33.7	95	8.8	82	76.5	95	89.2	55	81	0.9530	0.6326	0.4191
CHB test1 Case09	14.4	98	3.8	92	37.6	73	76.0	63	27.8	96	2.6	95	27.8	67	63.2	71	82	0.9907	0.5538	0.8089
CHB test1 Case10	194.9	71	7.8	84	68.4	90	97.7	50	44.2	94	2.5	95	79.3	97	92.7	53	79	0.9791	0.6768	0.6127
CHB test1 Case11	66.4	90	4.9	90	52.3	81	92.9	53	46.2	93	1.5	97	58.6	85	83.3	59	81	0.9902	0.4421	0.7136
CHB test1 Case12	69.6	90	13.6	72	32.5	70	96.6	51	68.6	90	14.0	71	25.6	66	96.1	51	70	0.8597	0.4213	0.1987
CHB test1 Case13	45.1	93	8.4	83	60.0	86	94.3	52	11.1	98	1.9	96	81.0	97	67.2	69	84	0.9781	0.8315	0.6950
CHB test1 Case15	18.3	97	2.5	95	63.0	87	89.8	55	55.9	92	2.4	95	74.5	94	87.5	56	84	0.9544	0.8353	0.6738
All Average	100.7	85	9.8	80	48.9	79	83.2	59	46.9	93	7.8	84	59.1	85	78.5	62	78	0.9739	0.5092	0.6726
All UNC	69.0	90	11.5	76	35.6	72	68.3	68	37.6	94	10.8	79	51.4	81	67.3	69	79	0.9970	0.4519	0.9143
All CHB	123.3	82	8.5	82	58.3	84	93.9	53	53.6	92	5.7	88	64.7	88	86.5	57	78	0.9573	0.5501	0.4999

Table 2 Evaluation results of the automatic segmentations, compared with three different gold standards: manual segmentations of the UNC rater, of the CHB rater, and of the combined STAPLE segmentation of different automatic methods.

The outcomes show that the all average scores of all measures, except the False Positive Rate are between 79 and 93. The scores compared with the UNC rater are between 79 and 85, and compared with the CHB rater are between 84 and 93. This means that the binary segmentations better resemble the reference segmentations of CHB rater, than those of the UNC rater. Compared with the CHB rater, the segmentation method performs comparable with a human expert, with respect to the volume difference, average difference and true positive rate. Only the False Positive Score is lower for both raters. In combination with the relatively high False Positive Rate, this may indicate a slight oversegmentation with respect to the reference segmentations.

For evaluation of the segmentation results compared with the combined STAPLE segmentation the specificity, sensitivity, and positive predictive value (PPV) were calculated. The PPV is the ratio of true positives to the sum of true positives and false positives. This coefficient provides a good measure combining both sensitivity and specificity. The average measures for the STAPLE segmentation evaluation, show a high specificity and a lower sensitivity, indication that our segmentation are more conservative than the combined STAPLE segmentation.

4 Discussion

In this paper, we propose a method for fully automated MS-lesion segmentation with good results, which is very suitable for usage in clinical practice. Lesions are segmented with an accuracy that is comparable to a human rater. Furthermore, the method is fully reproducible, which is highly advantageous in large and longitudinal cohort studies.

The T1 and FLAIR images are the only images used in this method. These images are quite common in clinical practice, which makes this method easy and widely applicable.

We have used the manual segmentations of the CHB rater only for training of the method. This was done to achieve an optimal segmentation procedure based on the manual segmentations of one rater. The result is appreciated from the evaluation table. The automatic segmentations of the testing data show a good resemblance with the CHB rater. The accuracy is comparable with this expert human rater. However, slight oversegmentation of the testing set with respect to the human rater has been observed from the evaluation results. This can be solved by applying a higher threshold on the probability map, making the binary segmentations smaller. On the other hand, compared with the combined STAPLE segmentation the lesions may be undersegmented. We can conclude from this that it is difficult to define one ultimate gold standard. Despite the presence of two manual segmentations as gold standard and a combined automatic result, it is still difficult to identify the most favorable segmentation. Brain abnormalities, such as MS-lesions, have a large partial volume area, since their intensity changes gradually into normal tissue. Therefore, the ultimate goal of the segmentation mostly determines the suitability of the segmentations. In large cohort studies, a structural over- or undersegmentation may not be problematic, as long as it is consequently performed. In this case, reproducibility is highly important, since a proper comparison between groups must be guaranteed. The proposed method is fully automated and reproducible.

The T1 image has been used for the creation of the brain mask, and only the FLAIR image was involved in the KNN-classification stage. Not using all image types may seem unfavorable, since extra features may add knowledge to the system. However, in previous study we have shown that in KNN-classification features with less information that existing features disturb the distances in the feature space, hence influencing the results negatively. Therefore, in KNN classification, reducing the dimensionality can improve the outcome.

The brain mask also has great consequences for the accuracy of the segmentation. A mask that is too wide influences the scaling of x-, y- and z-coordinates, and decreases the benefit of the spatial features. Furthermore, a large amount of surrounding tissue, like skull and skin, can disturb the KNN-classification of other tissue types. We opted for the brain extraction tool, because this generates a very accurate cortex mask in the T1 image. In general, it is important that the mask is similar for all patients, in the training set as well as in the testing set.

The generation of probabilistic in stead of binary segmentations is advantageous, since it provides large flexibility in further processing. According to the goal of the segmentation and the wishes of the user the segmentation can be adjusted to a wider or narrower one easily by the choice of the threshold on the probability map. Furthermore, the calculation of lesion volume can also be performed by using the probabilistic voxel values, sometimes giving more accurate results than using the binary segmentation.

In conclusion, KNN-classification provides a convenient technique for probabilistic segmentation of MS-lesion tissue. The proposed method is straightforward, in the sense that little preprocessing and no postprocessing steps are incorporated, and can be applied to routine diagnostic MRI. Therefore, it is suitable for brain segmentation problems in a large variety of applications.

References

[1] Y. Duan, P.G. Hildenbraind, M.P. Sampat, D.F. Tate, I. Csapo, B. Moraal, R. Bakshi, F. Barkhof, D.S. Meier, C.R. Guttmann. *Segmentation of subtraction images for the measurement of lesion change in multiple sclerosis.* AJNR Am J Neuroradiol 2008; 29(2): 340-6.

[2] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, P. Suetens. *Automated segmentation of multiple sclerosis lesions by model outlier detection.* IEEE Trans Med Imaging 2001; 20(8): 677-88.

[3] F. Kruggel, J.S. Paul, H.J. Gertz. *Texture-based segmentation of diffuse lesions of the brain's white matter.* Neuroimage 2008; 39(3): 987-96.

[4] P. Anbeek, K.L. Vincken, M.J.P van Osch, R.H.C Bisschops, J. van der Grond. *Probabilistic segmentation of white matter lesions in MR imaging.* Neuroimage 2004; 21(3): 1037-44.

[5] P. Anbeek, K.L. Vincken, G.S. van Bochove, M.J.P. van Osch, J. van der Grond. *Probabilistic segmentation of brain tissue in MR imaging.* Neuroimage 2005; 27(4): 795-804.

[6] S.M. Smith. *Fast robust automated brain extraction.* Hum Brain Mapp 2002; 17(3): 143-155

[7] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, Great Britain, 1995.

[8] R.O. Duda, P.E. Hart, D.G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., New York, U.S.A. 2001.

[9] S.K. Warfield, K.H. Zou, W.M. Wells, *Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.* IEEE Trans Med Imaging. 2004; 23(7): 903-21.