

Automated Segmentation of White Matter Lesions in 3D Brain MR Images, Using Multivariate Pattern Classification

Zhiqiang Lao^a, Dinggang Shen^a, Abbas Jawad^b, Bilge Karacali^a, Dengfeng Liu^a, Elias R. Melhem^a, R. Nick Bryan^a, Christos Davatzikos^a

^a Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104, USA

^b Department of Biostatistics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
{Zhiqiang.Lao, Dinggang.Shen, Abbas.Jawad, Bilge.Karacali, Dengfeng.Liu, Elias.Melhem, Nick.Bryan, Christos.Davatzikos}@uphs.upenn.edu

Abstract

This paper presents a fully automatic white matter lesion (WML) segmentation method, based on local features determined by combining multiple MR acquisition protocols, including T1-weighted, T2-weighted, proton density (PD)-weighted and fluid attenuation inversion recovery (FLAIR) scans. Support vector machines (SVMs) are used to integrate features from these 4 acquisition types, thereby identifying nonlinear imaging profiles that distinguish and classify WMLs from normal brain tissue. Validation on a population of 45 diabetes patients with diverse spatial and size distribution of WMLs shows the robustness and accuracy of the proposed segmentation method, compared to the manual segmentation results from two experienced neuroradiologists.

1. Introduction

WMLs in the brain are common in relatively healthy older individuals, as well as in patients with a variety of clinic diseases, including vascular disease, multiple sclerosis (MS), diabetes, stroke and head injury.

Several methods have previously been published for segmentation of WMLs, primarily using MRI data from population with MS. [1-3].

We propose a novel fully automatic WML segmentation approach which uses a combination of image analysis and pattern recognition methods. There are three main steps in our approach, as summarized in Figure 1. *First*, a preprocessing step is designed to include co-registration, skull-stripping, as well as intensity normalization. *Second*, a set of training samples were manually segmented by two experienced neuroradiologists, and then a classification model is built based on the manual segmentation results via SVM and AdaBoost. *Third*, the SVM model, constructed from training samples, is used to perform the voxel-wise segmentation. *Finally*, false positive voxels are further eliminated via a technique described later, thereby producing relative accurate WMLs segmentation results.

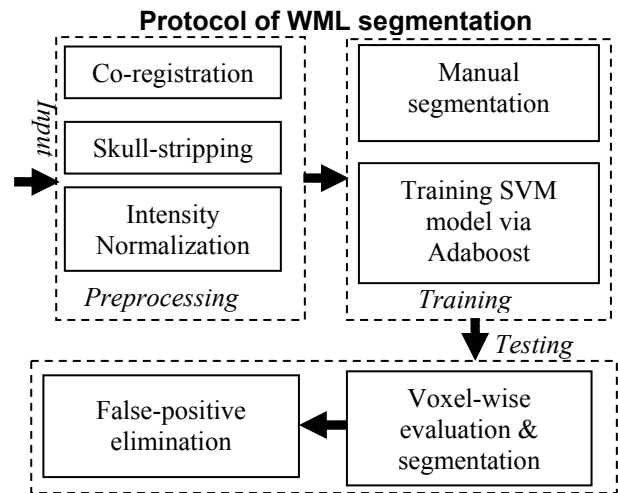


Figure 1 Summary of our automatic WMLs segmentation protocol.

2. Methods

2.1. Preprocessing

The multi-modality images acquired from the same subject were co-registered, in order to compensate for possible motion between scans. Mutual-information-based registration [4], provided by ITK [5], is employed for co-registration of multi-modality image, and the FLAIR image of each subject is used as a reference space, to which all other modality images will be transformed. After co-registration, BET [6] is used to generate an initial brain tissue mask from the co-registered T1 image, and then this brain tissue mask is used to extract brain region from other modality images. Finally, intensity inhomogeneity of same modality images across different subjects is corrected by a global histogram matching method.

2.2 Training

2.2.1 Manual segmentation

A set of training subjects were randomly selected from participants of ACCORD-MIND study, men and women

aged 55 or older with diabetes. WMLs in these subjects were then manually segmented by two neuroradiologists (RNB and ERM). The segmentation of the 1st rater was regarded as the gold standard for training our classifier, while the segmentation result of the 2nd rater was used for evaluating inter-rater agreement and for comparing it against computer-rater agreement.

2.2.2 Feature vector

In general, the amount of intensity overlap between WMLs and normal tissue varies greatly across different modalities. It is thus important to integrate information from different modalities, in order to minimize the ambiguity in identifying WMLs using only a single modality image.

A feature vector is computed for each non-background voxel of each subject in FLAIR space. In order to include spatial information from the vicinity of each voxel and make feature vectors distinctive in identifying WMLs, each feature vector includes not only local intensities of a corresponding voxel in four modality images, but also intensities of neighboring voxels in four modality images. Moreover, to make feature vectors robust to noise, each modality image of same subject is smoothed by a Gaussian filter with a very small kernel, before the feature vector is calculated. Mathematically, for a point v in domain Ω , the feature vector is defined as $F(v)=\{I^m(\mu)\}$: $\forall m \in \{T_1, T_2, PD, \text{ or } FLAIR\}, \forall \mu \in \mathbf{B}(v)$, where $\mathbf{B}(v)$ is a small neighborhood of v in Ω . Figure 2 shows the discrimination ability of the feature vector defined above. For a voxel indicated by a white cross in the left image, the distances (in Hilbert space, as described in the following section) between features of this voxel and features of all other voxels in this image are computed and color-coded as shown in the right. It can be observed that only lesion tissue shows high similarity to the selected WML voxel.

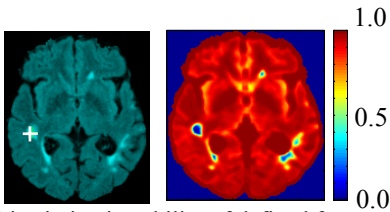


Figure 2 Discrimination ability of defined feature vector.

2.2.3 Training SVM via AdaBoost

SVM [7] [8] is used here as a classifier for WML segmentation. It is essential to select typical non-lesion samples to train SVM, instead of using the whole set of non-lesion samples, so that the numbers of WML and healthy tissue samples are about the same. To achieve this, AdaBoost [9] is used to focus on "difficult" samples until a certain degree of training errors is achieved. During Adaboosting procedure, each sample receives a weight that determines its probability of being selected in a training set for the next iteration. If a training sample is accurately classified, then its chance of being used again

in subsequent iteration is reduced; conversely, if a training sample is inaccurately classified, then the likelihood of this sample being selected again is increased.

2.3 Testing

2.3.1 Voxel-wise segmentation of WML by SVM

In the testing stage, preprocessing steps described in section 2.1 need to be completed first, and then the pseudo-likelihood of each voxel being WML is measured by a generated SVM classifier, as described above. The output of SVM is a scalar value for each brain tissue location (as shown with different color in the left of Figure 3), which is further segmented by an optimal threshold to produce initial WML segmentation results (as shown in the right of Figure 3). The false positive segmentations in these initial WML segmentations will be screened out by the method proposed next.

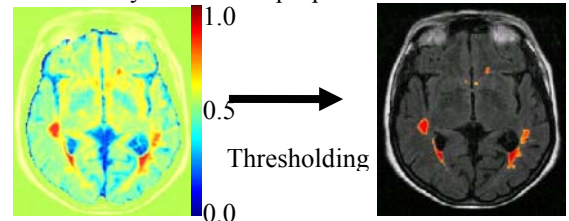


Figure 3 Illustration of voxel-wise classification by SVM.

2.3.2 Elimination of false positive segmentations

Misregistration usually results in a significant number of false positive segmentations around the cortex. This is because intensities around the cortex vary greatly, and thus the extracted features for those locations are more sensitive to mis-registration. On the other hand, intensities in the deep brain regions are more homogeneous, thus the extracted features are not as sensitive to mis-registration.

By analyzing the spatial distribution of feature vectors from different samples, we found that the feature vectors of false positive voxels actually form a third class in the Hilbert space associated with the SVM training samples, which is far away from both classes of lesion and non-lesion training samples, and thus those false positive voxels can be eliminated to a large extent by computing the distance of their feature vectors to the training samples in the Hilbert space. In other words, these are voxels whose multivariate imaging profiles don't look like the imaging profiles that were observed during training. The distance measure in Hilbert space between two vectors v_1 and v_2 can be calculated as

$$D_h^2(v_1, v_2) = K(v_1, v_1) + K(v_2, v_2) - 2K(v_1, v_2),$$

where K is the Gaussian kernel function used by the SVM.

Suppose $L = \{v_i^l, 1 \leq i \leq m\}$ is the set of feature vectors of WMLs in training samples, m is the total number of feature vectors in L ; $N = \{v_i^n, 1 \leq i \leq p\}$ is the set of feature vectors of normal tissues in training samples, p is the total number of feature vectors in N ; $F = \{v_i^f, 1 \leq i \leq q\}$ is the set of feature vectors of false positive voxels obtained by

initial SVM classification, q is the total number of feature vectors in F . For each v_i^l , its distance to L in Hilbert space is defined as $d_{v_i^l} = \min_{j=1}^m D_h^2(v_i^l, v_j^l)$, where $j \neq i$; similarly, for each v_i^n , its distance to N is defined as $d_{v_i^n} = \min_{j=1}^p D_h^2(v_i^n, v_j^n)$, where $j \neq i$. For each v_i^f , its distance to L is defined as $d_{v_i^f}^L = \min_{j=1}^m D_h^2(v_i^f, v_j^f)$; similarly its distance to N is defined as $d_{v_i^f}^N = \min_{j=1}^p D_h^2(v_i^f, v_j^f)$. Figure 4 shows the distributions of these distances, which suggests that we can simply use this minimal distance measure to eliminate the false positive voxels by selecting a suitable threshold.

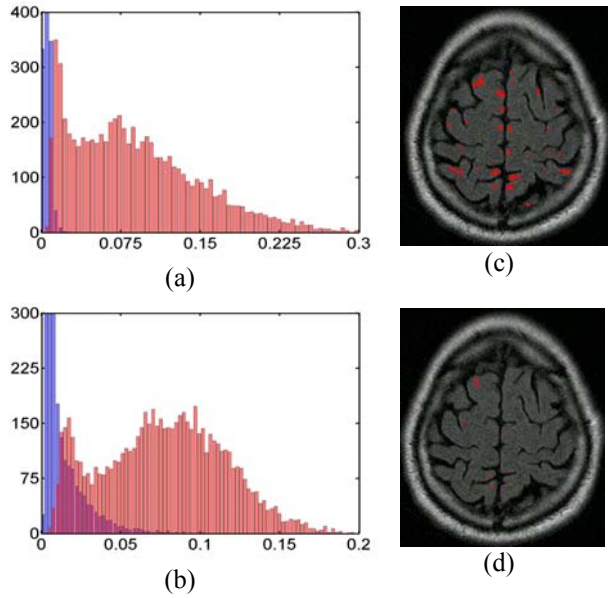


Figure 4 Demonstration of false positive elimination via vector distance in Hilbert space. (a) Distance distribution of $\{d_{v_i^l}^L\}$ (blue) and $\{d_{v_i^f}^L\}$ (red). (b) Distance distribution of $\{d_{v_i^n}^N\}$ (blue) and $\{d_{v_i^f}^N\}$ (red). WML segmentation results (c) before false positive elimination, and (d) after false positive elimination.

Orbital hyperintense regions like Eye and fat, can not be completely removed by the automatic skull-stripping algorithm used in preprocessing stage, but feature vectors belonging to these regions are more similar to WMLs instead of normal tissue. Elimination of this kind of false positive is done by morphological operations combined by adaptive thresholding in skull-stripped FLAIR image I . Suppose that I 's histogram is $H(I)$, $M(I)$ is a binary mask of nonzero voxels in I , C is a closing operation, and τ is an adaptive intensity threshold. Then, the algorithm of eliminating orbital artifacts is given as follows.

Algorithm of eliminating orbital hyperintense regions:

- Step 1: compute intensity threshold τ for hyperintense region that satisfies $\sum_{i=1}^{\tau} H(i) / \sum_{i=1}^{\max} H(i) \geq v_t$, where v_t is a predefined volume ratio threshold.
- Step 2: compute $M_c(I) = C(M(I))$, which is a binary mask after a closing operation over $M(I)$.
- Step 3: compute the mask difference $D_c(I) = M(I) - M_c(I)$.
- Step 4: modify $D_c(I)$ by eliminating voxels whose intensity value in FLAIR is larger than τ . If $D_c(I)$ has no change before and after modification, then save current $M(I)$ and go to Step 6.
- Step 5: modify $M(I) = M_c(I) \cup D_c(I)$, go to Step 2.
- Step 6: use $M(I)$ as a new brain mask to eliminate false positive voxels around orbit, and stop.

In practice, the majority of orbital false positives around orbit can be removed only after a few iterations.

3. Result

Figure 5 shows the comparison of our segmentation results with those of "gold standard", for two out of the 45 testing subjects.

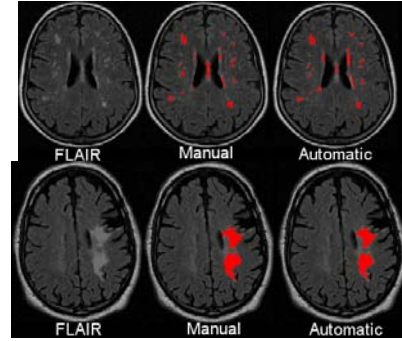


Figure 5 Comparison of WML segmentation results between gold standard and automatic segmentation for two individual subjects.

Figure 6 shows the ROC curve of our segmentation algorithm obtained by varying the classification thresholds. It is zoomed to show interesting details. Different symbols on the ROC curve show different thresholds used. Additionally, "*" shows the 2nd rater's manual segmentation result, compared to the gold standard.

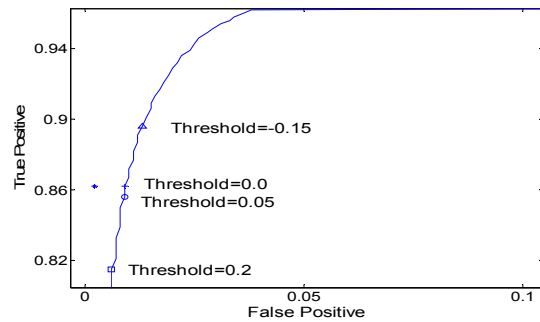


Figure 6 A zoomed part of ROC curve of our segmentation algorithm. The "*" indicates the result of the 2nd rater compared to gold standard. Other symbols on the curve denote different thresholds, i.e., 'Δ' threshold = -0.15, '+' threshold = 0.0, 'o' threshold = 0.05, '□' threshold = 0.2.

We have also performed statistical comparisons between the lesion volume obtained by manual and automatic segmentation in terms of Pearson correlation (PC), Spearman correlation (SC), coefficient of variation (CV) and reliability coefficient (RC). PC and SC measure the degree to which segmented WML lesion volume by the automatic method and the gold standard are systematically related. CV measures intra-rater variation. Table 1 and Table 2 show PC and SC measures between our method and the two manual segmentation results, respectively. From a statistical point of view, a p-value < 0.06 is regarded as highly correlated. CV values of 1st rater, 2nd rater and our algorithm are 2.44, 1.96 and 1.27, respectively. RC values of 1st rater, 2nd rater and our algorithm are 1.0, 0.6236 and 0.9969, respectively. These results indicate that the proposed automatic lesion segmentation algorithm is reasonably highly correlated to the manual raters, while being consistent and reliable relative to the inter-rater agreement.

Table 1 PC comparison of automatic and manual segmentation method

| | 1 st rater | 2 nd rater | Auto |
|-----------------------|-----------------------|-----------------------|--------|
| 1 st rater | 1.0 | 0.9703 | 0.8532 |
| 2 nd rater | 0.9703 | 1.0 | 0.8844 |
| Auto | 0.8532 | 0.8844 | 1.0 |

Table 2 SC comparison of automatic and manual segmentation method

| | | 1 st rater | 2 nd rater | Auto |
|-----------------------|---------|-----------------------|-----------------------|--------|
| 1 st rater | Roh | 1.0 | 0.9554 | 0.7753 |
| | p-value | | <.0001 | <.0001 |
| 2 nd rater | Roh | 0.96 | 1.0 | 0.7318 |
| | p-value | <.0001 | | <.0001 |
| Auto | Roh | 0.7753 | 0.7318 | 1.0 |
| | p-value | <.0001 | <.0001 | |

4. Discussion

Both the ROC analysis and statistical comparisons with the manual raters indicate that the proposed algorithm is both accurate and reliable. The proposed algorithm can be further improved by registering all individual brains into a template space through a relatively accurate registration [10]. Thus, the distribution of manually-segmented WMLs in the template space can be used as additional information for WML segmentation and also for elimination of false positive voxels.

It is important to note that, although the two raters were very highly correlated, their actual measurements differed significantly in magnitude. In particular, Rater 2 consistently over-segmented relative to Rater 1, or conversely, Rater 1 consistently under-segmented relative to Rater 2. This emphasizes a common problem encountered especially in longitudinal studies: although raters can be consistent in their delineation, their results cannot be mixed together, especially if one rater succeeds

the other, because this would introduce artificial longitudinal changes in the data. The effect can also be significant in cross-sectional studies, if the two raters are not completely randomly assigned to subjects in two or more groups. In any case, mixing measurements from different raters can significantly increase measurement variance, and therefore reduce statistical power. In contrast, despite its limitations, the computer-based delineation is very consistent.

In conclusion, the proposed method offers highly accurate and generalized WML segmentation. In its generality, it is also applicable to many other segmentation problems, such as tissue segmentation or segmentation of atrophy. 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, longitudinal population studies.

Acknowledgments: We would like to thank Drs Lenore J. Launer, Jeff D. Williamson and Ms. Lisa Desiderio for providing the datasets, valuable comments and giving us permissions to publish this paper. We also would like to thank patients recruited by ACORD-MIND project, which is funded by the NIA through an intra-agency agreement with NIH/NIH [Y3-HC-3065]. This research was supported (in part) by the Intramural Research Program of the NIH, National Institute on Aging contract N01-HC-95178.

References

1. Kamber, M., et al., *Model-Based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images*. IEEE Transactions on Medical Imaging, 1995. **14**(3): p. 442-453.
2. Udupa, J., Wei, L., Samarasekera, S., Miki, Y., van Buchem, MA, Grossman, RI, *Multiple Sclerosis Lesion Quantification Using Fuzzy-Connectedness Principles*. IEEE Transactions on Medical Imaging, 1997. **16**(5): p. 598-609.
3. Warfield, S.K., et al., *Adaptive, template moderated, spatially varying statistical classification*. Medical Image Analysis, 2000. **4**(1): p. 43-55.
4. Viola, P. and W.M. Wells III. *Alignment by maximization of mutual information*. in *Proc. Int. Conf. Computer Vision*. 1995. Los Alamitos, CA.
5. ITK, S.G., *ITK Software Guide*, <http://www.itk.org/HTML/Documentation.htm>.
6. Smith, S.M., *BET: Brain Extraction Tool*. *FMRIB technical report TR00SMS26*.
7. Vapnik, V.N., *Statistical Learning Theory*. 1998, New York: Wiley. 736.
8. Lao, Z., et al., *Morphological classification of brains via high-dimensional shape transformations and machine learning methods*. Neuroimage, 2004. **21**(1): p. 46-57.
9. Richard, O.D., E.H. Peter, and G.S. David, *Pattern Classification*. 2nd edition ed. Vol. 1. 2000: Wiley-Interscience. 654.
10. Shen, D. and C. Davatzikos, *HAMMER: Hierarchical attribute matching mechanism for elastic registration*. IEEE Transactions on Medical Imaging, 2002. **21**(11): p. 1421-1439.