Accuracy and execution speed of automatic voxel-based algorithms for segmenting stroke lesions in clinical DWI imaging

S. Mocking[1], P. Garg[2], A. Chutinet[3], W. A. Copen[4], A. Sorensen[5], and O. Wu[6]

[1]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, United States, [2]Athinoula A Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, [3]Department of Neurology, Massachusetts General Hospital, United States, [4]Department of Radiology, Massachusetts General Hospital, [5]Athinoula A Martinos Center for Biomedical Imaging, Massachusetts General Hospital, [6]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital

Introduction: Stroke researchers are increasingly advocating for the use of a "tissue clock" over that of a "ticking clock" for acute stroke patient management. Diffusion-weighted imaging (DWI) is the most useful imaging modality for patients seen within 12 h of symptom onset for stroke diagnosis and prognosis [1]. Baseline DWI lesion volume is being used as an inclusion or exclusion criterion for several clinical trials of acute stroke intervention such as MR RESCUE or EXTEND. Rapid, reproducible assessments of DWI lesion volumes are critical since an average of 1.9 million neurons is lost for every minute of ischemia [2]. We compare four segmentation algorithms' performances for identifying acute DWI lesions against manual outlines on the basis of accuracy and execution speed.

Methods: We retrospectively analyzed MRI data from acute stroke patients imaged within 12 h of stroke onset for whom manual expert outlines are available and who received DWI and perfusion-weighted imaging (N=159). Multisection axial DWI was acquired using echo-planar imaging sequences performed on a 1.5 T imaging instrument (GE Medical Systems) with diffusion-weighting or b-value of 1000 s/mm^2. Apparent diffusion coefficient (ADC) maps were calculated from the slope of the linear regression fit of the log of the DWI and b_0 (b-value=0 s/mm^2) images. The b_0 images were used for the T2-weighted image (T2WI). Four voxel-based segmentation methods were compared with manual outlines in terms of number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). For all the methods, a mask dividing voxels into background and tissue was constructed using Otsu bimodal thresholding [3] of the DWI image; the voxels outside this mask were excluded from further classification. The dataset was randomly divided into 38 cases as a training group and 121 for validation. The first method is based on simple thresholding ADC maps for which voxels with ADC values less than 0.650×10^{-3} mm^2/s [4] are considered lesion voxels. The second method is k-nearest neighbor (k-NN) where the number of neighbors, k = 7. For k-NN, we trained the classifier by inserting voxels into a data structure optimized for nearest neighbor searching [5]. Given the memory constraints of a lower-end desktop computer and the overhead of this data structure we randomly sub-sampled 15% of voxels from the training data. Following training, the sample images were histogram-matched to the first patient in the training set before classifying the voxels to their nearest neighbors in feature space with a majority rule. The third method we investigated is a k-means cluster algorithm with the number of clusters, k = 7. The fourth method uses Iterative Self Organizing Data Analysis (ISODATA) [6] where initial number of clusters, K=9; maximum number of iterations, I=100; maximum number of pairs of clusters lumped in one iteration P=3; cluster size, $Theta_N$=1000; splitting parameter, $Theta_S$=0.05; lumping parameter, $Theta_C$=0.03. ISODATA and k-means methods were applied to images normalized to values between 0 and 1. After voxel classification with a Mahalanobis distance metric, in the case of the unsupervised methods the cluster masks were then divided into spatially coherent regions using connected component labeling (CCL). Regions of >200 voxels with a mean DWI signal intensity (SI) greater than two standard deviations above the mean DWI SI of all the tissue voxels and an ADC more than half a standard deviation under the mean ADC were then marked as potential lesions. Performance of the four techniques was evaluated based on sensitivity (TP/(TP+FN)), specificity (TN/(TN+FP)), accuracy ((TP + TN) / (TP+TN+FP+FN)), Dice's similarity measure (DSM=(2TP)/(2TP + FP +FN)) and run time. Runtime for each algorithm is measured on a high-end desktop machine (3.00 GHz Intel Xeon CPU, 16GB RAM). Differences between training and testing group were compared (two-sided Wilcoxon-test for continuous variables and Pearson chi-square test for categorical variables). All results are mean±standard deviation or median [interquartile range]. One-way ANOVA was performed with post-hoc Student Newman-Keuls test was performed comparing performance of DSM across the four methods. Subset analysis was performed between DWI lesion volumes ≥ 5 cm^3 (N=76) and lesions < 5 cm^3 (N=45). P<0.05 were considered to indicate a significant difference in all analyses.

Results: Patient demographics are shown in Table 1. Results of the four segmentation methods are shown in Table 2. An example of the results of the algorithms is shown in Figure 1. A significant difference was found between the four techniques (P<0.001). ISODATA had significantly lower DSM compared to the other techniques except for simple thresholding. K-means and k-NN performed equivalently. Performance was significantly higher for lesions more than 5 cm^3 (P<0.001) for all techniques: thresholding (0.20±0.18 vs. 0.01±0.01), k-means (0.34±0.34 vs 0.05±0.16), ISODATA (0.23±0.32 vs 0.02±0.12), k-NN (0.32±0.20 vs 0.07±0.10). To be expected, simple thresholding had the fastest execution time. k-NN took the longest.

Table 1. Demographics

Characteristic	Training data (N = 38)	Testing data (N = 121)	p-value
Age (years)	66±18	69±15	0.49
Male gender	14	65	0.07
NIH Stroke Scale	8 [3-17]	6 [3-15]	0.32
DWI lesion volume (cm^3)	31.3±59.0	31.3±59.0	0.92
Stroke subtype			0.01*
Cardioembolic	19	51	
Large vessel	5	33	
Small vessel	3	3	
Undetermined	6	31	
Other determined	5	3	

* p < 0.05

Table 2. Algorithm performance.

Algorithm	Sensitivity	Specificity	Accuracy	DSM	Time/case (s)
ADC threshold	0.60±0.22	97.4%±1.7	97.6%±1.7	0.13±0.17	0.64 ± 0.08
k-means	0.21±0.31	99.9%±0.3	99.7%± 0.6	0.23±0.32	6.92±1.62
ISODATA	0.15±0.27	99.9%±0.3	99.6%±0.6	0.16±0.28	5.69±1.61
k-NN	0.21±0.18	99.9%±0.06	99.6%±0.7	0.22±0.21	121.6±8.7

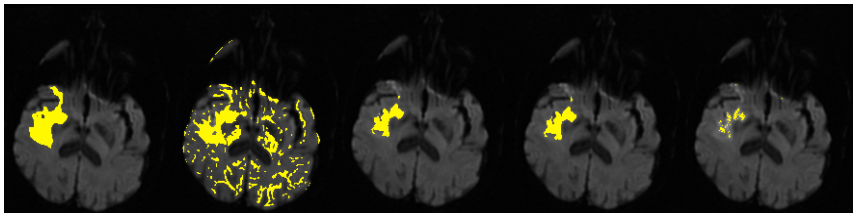

Figure 1. Segmentations of a representative lesion. From left to right: manual outline, ADC threshold, k-means, ISODATA, k-NN.

Discussion: We have shown that real-time automatic delineation of DWI lesions is possible. K-means appears to provide the best performance of all the methods examined in terms of computational time and accuracy. Particularly thresholding, but also k-NN, might have suffered in terms of specificity and benefited in sensitivity as a result of not having the spatial coherence of CCL and region selection. The techniques are highly specific, driven by the large number of normal voxels. The sensitivities are low, resulting in underestimation of lesions. Increased sensitivity can be obtained but at the cost of specificity, as seen by the ADC threshold technique. There is a clear correlation between lesion detection and lesion volume. Poorest DSMs were found with small lesion volumes.

Cases were randomly divided into training and sample sets without knowledge of the stroke subtype, which is typically the situation when patients are seen acutely. We found a significant difference in the distribution of stroke subtypes between training and testing data sets. A training set more representative of the test cases may improve results for k-NN. Its run time was high as a result of having to load a large training set for each case. In future research it may be worthwhile to include algorithms that maximize spatial coherence, such as graph-cuts- or wavelet-based approaches as they should be less sensitive to noise and artifacts than CCL.

References: 1. Schellinger P, et al, Neurology, July 13, 2010; 75:177-185 2. Saver JL, Stroke 2006; 37:263–67 3. Otsu N, IEEE Trans. Sys., Man., Cyber. 1979; 9: 62–66. 4. Miyabe M, et al. J Cereb Blood Flow Metab. 1996; 16:881-89 5. Arya S, 5th Ann. ACM-SIAM Symposium on Discrete Algorithms, 1994, 573-582. 6. Ball GH, Storming Media, 1965.