A Multi-scale Multiple Sclerosis Lesion Change Detection in a Multi-sequence MRI

Myra Cheng[1], Alfiia Galimzianova[1(✉)], Žiga Lesjak[2], Žiga Špiclin[2], Christopher B. Lock[1], and Daniel L. Rubin[1]

[1] Stanford University, Stanford, CA 94305, USA
alfiia@stanford.edu
[2] University of Ljubljana, 1000 Ljubljana, Slovenia

Abstract. Multiple sclerosis (MS) is a disease characterized by demyelinating lesions in the brain and spinal cord. Quantification of the amount of change in MS lesions in magnetic resonance imaging (MRI) over time is important for evaluation of drug effectiveness in clinical trials. Manual analysis of such longitudinal datasets is time- and cost prohibitive, and also prone to intra- and inter-rater variability. Accurate automated change detection methods would be highly desirable. We propose a new MS lesion change detection method that integrates a voxel's multi-sequence MR intensity with its immediate neighborhood context and the texture of the extended neighborhood in a machine learning framework. On our dataset of 15 patients, the proposed method had higher performance (median AUC-ROC = 0.97, AUC-PR = 0.43, Wilcoxon's signed rank test, $p < 0.001$) than implemented baseline methods. As such, the proposed method has potential clinical applications as an efficient, low-cost algorithm to capture and quantify local lesion change and growth.

Keywords: Multiple sclerosis · Change detection
Multi-scale image descriptors

1 Introduction

Multiple sclerosis (MS) is a disease of the central nervous system (CNS) that affects over 400,000 people in the U.S. and 2.5 million people worldwide. It is one of the leading causes of non-traumatic disability among young and middle-aged adults [1]. Currently, MS has no cure, although there is an ongoing research in search for improved treatment and management of the disease. The success of such research depends on clinical trials, in which the response to treatment and change in disease status must be quantified in an accurate and consistent manner.

Multi-sequence magnetic resonance imaging (MRI) is the standard imaging exam performed to analyze the white-matter lesions for diagnosis and follow-up evaluation of MS. Quantitative evaluation of the changes in the MS lesions

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 353–360, 2018.
https://doi.org/10.1007/978-3-030-00889-5_40

appearance requires annotation of the corresponding areas in the brain, which when done manually is time-consuming and subjective. To address these challenges, reliable automated methods are needed.

The strategies for automated change detection can be categorized as longitudinal volumetric analysis, deformable image registration, and longitudinal analysis of MR intensity [5]. Longitudinal volumetric analysis relies on segmentation of MS lesions at each imaging timepoint independently and can only provide global measures of the lesion change, such as the count of new lesions and the total lesion volume difference. Deformable image registration relies on deformation fields obtained during non-rigid alignment of the MR images at two timepoints and can quantify local changes through analysis of enlarging and shrinking lesions, while detection of new or disappearing lesions is limited. The longitudinal analysis can address the issues of the previous two strategies through rigid of affine registration being followed by intensity analysis at matching anatomical sites, thus allowing for local quantification of all types of lesion change [4]. The computational core of such analysis is detection of lesion change for each voxel of the brain MR image set from two imaging timepoints. Although single-timepoint MS lesion segmentation approaches have incorporated multiple scales of spatial information for context [6], longitudinal MS lesion analysis has been limited to change detection using voxel intensities independently [4]. Texture descriptors specifically can aid in a compact representation of a local context of the multi-sequence MR images and, moreover, have been successfully applied to MS lesion segmentation [11].

In this work, we propose a change detection method that incorporates the local context of the multi-sequence MR images at three scales. The multi-scale descriptors are extracted from the intensity information of the voxels immediate neighborhood, and the intensity and texture information of a larger surrounding image patch. Our experiments demonstrated that incorporating the contextual information to change detection improved the performance at each scale and the proposed method statistically significantly outperformed the baseline state-of-the-art approach [8].

2 Materials and Methods

2.1 Dataset

We used anonymized imaging data from 15 MS patients with two imaging exams, each with three MRI sequences: T1-weighted (T1), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR), acquired at 1.5T. Pre-processing included resampling to the common spatial resolution of $1 \times 1 \times 3\,\mathrm{mm}^3$, inhomogeneity correction on all sequences using N4 bias correction [10], registration of all sequences at both timepoints to a common space [2], and intracranial volume extraction from the T1 sequences using BET 2 [7]. The reference lesion change labels were acquired as a consensus segmentation of the corresponding regions by two neuroradiologists.

2.2 MS Lesion Change Detection

For each patient, we consider six three-dimensional volumes: T1, T2, and FLAIR at two time points. We denote intensity of a voxel v from the intracranial volume mask for patient i, from the imaging study conducted at time t_j $(j = 1, 2)$ by $M_{iv}^{t_j}$, where $M \in \{T1, T2, FLAIR\}$. For common interpretations of voxel intensities, each volume was normalized by computing the z-scores of the intracranial volume intensities:

$$\widetilde{M}_{iv}^{t_j} = \frac{M_{iv}^{t_j} - \mu_{i,M}^{t_j}}{\sigma_{i,M}^{t_j}}$$

where the mean $\mu_{i,M}^{t_j}$ and standard deviation $\sigma_{i,M}^{t_j}$ are computed as sample statistics across the voxels in the intracranial volume mask for patient i at time t_j. For each patient, dissimilarity maps were extracted by voxel-wise subtraction of the normalized image between two time points for each imaging sequence M as $\Delta M_{iv} = \widetilde{M}_{iv}^{t_2} - \widetilde{M}_{iv}^{t_1}$.

Lesion Change Model and Voxel-Level Descriptors. We model the presence of change in a set of preregistered multi-sequence MR images as a function of descriptors extracted from $\widetilde{M}_{iv}^{t_1}$ and ΔM_{iv} for each imaging sequence M. With the voxel-level lesion change represented by a random variable R, the probabilities for each test patient i at voxel v are modeled as a logistic regression:

$$logit[P\{R_{iv} = 1\}] = \alpha_0 + \Sigma_{x=1}^{6} \alpha_x I_{iv}^x \tag{1}$$

where $I^x \in \{\widetilde{FLAIR}^{t_1}, \Delta FLAIR, \widetilde{T1}^{t_1}, \Delta T1, \widetilde{T2}^{t_1}, \Delta T2\}$ constitute the six input image volumes for patient i.

Incorporating Neighborhood Information. The first scale of the context we incorporate into the model in Eq. (1) is the immediate neighborhood of a voxel in the form of $\Delta FLAIR$ values over a $K \times K$ neighborhood of each considered voxel. With x representing the indices of the neighborhood voxels, and these values were used as additional descriptors to learn additional coefficients β:

$$logit[P\{R_{iv} = 1\}] = \alpha_0 + \Sigma_{x=1}^{6} \alpha_x I_{iv}^x + \Sigma_{x=1}^{8} \beta_x \Delta FLAIR_{iv}^x \tag{2}$$

Incorporating Local Texture Descriptors. To incorporate a wider context, we extracted texture descriptors from a larger $L \times L$ neighborhood of a voxel on a $\Delta FLAIR$ sequence. From voxel intensities within the brain mask, descriptors representing intensity statistics, such as mean, standard deviation, and kurtosis, were generated. Texture-based descriptors were extracted from the gray-level co-occurrence matrix (GLCM) of each patch using Haralick descriptors [3] and the gray-length run-length matrix (GLRLM) descriptors [9]. This resulted in 31-dimensional vectors, which were normalized to their z-scores to obtain T_{iv}^x, $x = 1, \ldots, 31$ for each selected voxel v. The logit model using these 31 descriptors is denoted as:

$$logit[P\{R_{iv} = 1\}] = \gamma_0 + \Sigma_{x=1}^{31} \gamma_x T_{iv}^x \tag{3}$$

Table 1. Median ± median absolute deviation values for patient-level AUC-ROC and AUC-PR generated across 45 independent three-fold cross-validated trials.

Metric	Baseline 1	Baseline 2	Baseline 3	Proposed
AUC-ROC	0.94 ± 0.04	0.94 ± 0.04	0.93 ± 0.04	0.97 ± 0.01
AUC-PR	0.14 ± 0.19	0.17 ± 0.18	0.29 ± 0.13	0.43 ± 0.16

Multi-scale Method. The joint multi-scale model combines the multi-sequence voxel intensities with the immediate neighborhood and the local texture descriptors. This descriptor was used as input to fit the logistic regression function:

$$logit[P\{R_{iv} = 1\}] = \alpha_0 + \Sigma_{x=1}^{6}\alpha_x I_{iv}^x + \Sigma_{x=1}^{8}\beta_x \Delta FLAIR_{iv}^x + \Sigma_{x=1}^{31}\gamma_x T_{iv}^x \quad (4)$$

The logistic regression model was first learned using the balanced dataset over a selection mask, computed at voxel v for patient i as:

$$S_{iv} = \begin{cases} 1 & \Delta FLAIR_{iv} > \sigma_\Delta FLAIR_i \\ 0 & \text{otherwise} \end{cases}.$$

In testing, the learned set of coefficients $\alpha_0, \ldots \alpha_6, \beta_1, \ldots \beta_8, \gamma_1, \ldots \gamma_{31}$ was applied to infer the lesion change probability maps for the whole intracranial volume of a test subject.

3 Experiments and Results

The proposed method (Eq. (4)) was compared to an implementation of the existing state-of-the-art method [8] that considers only the multi-sequence voxel intensities (Eq. (1), referred here as Baseline 1); a method using the intensities as well as the immediate neighborhood context information from a 3×3 neighborhood (Eq. (2), referred to as Baseline 2); and a method using texture-based descriptors (Eq. (3), referred to as Baseline 3).

A three-fold cross-validation scheme was used. To minimize the effect of random variation on training and testing subsets within folds, each three-fold experiment was run three times, resulting in 45 patient-level ROC and PR curves for each method. Table 1 and Fig. 2 depict comparisons of AUC-ROC and AUC-PR for the four methods. The higher performance of the Multi-Scale Method compared to other methods was statistically significant ($p < 0.001$) both in terms of AUC-ROC and AUC-PR.

3.1 Importance of Texture-Based Descriptors

Comparing Baseline 2 and the Multi-Scale Method, texture-based radiomic features clearly contribute to the latters higher performance. Here we investigate the discriminative power of the set of texture descriptors. Figure 3 shows the differences in descriptor values between voxels with lesion change and voxels with

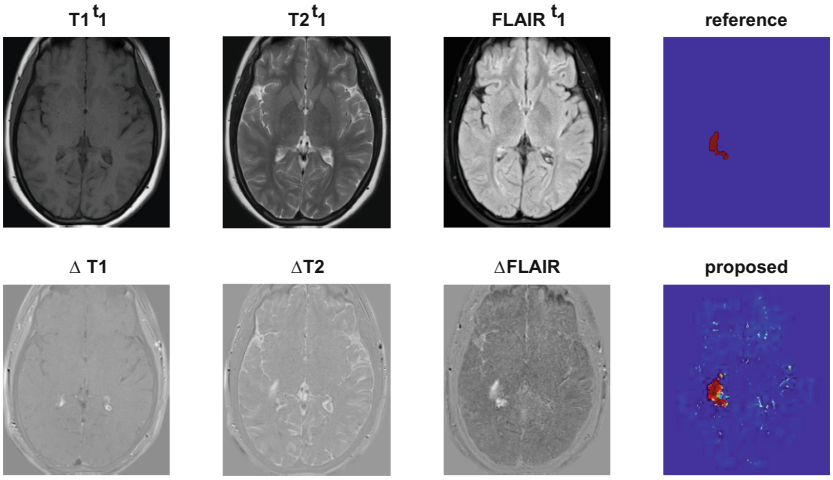

Fig. 1. A sample of the input images, manually annotated reference labels for lesion change, and the lesion change probability map generated by the proposed method.

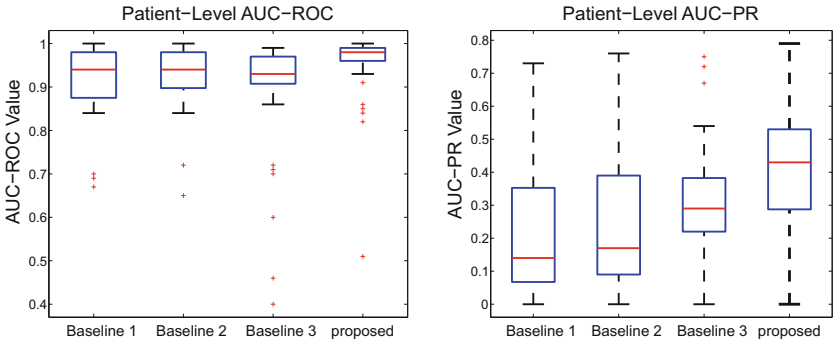

Fig. 2. Distributions of patient-level AUC-ROC (left) and AUC-PR (right) values generated across 45 3-fold cross-validated trials on test patients.

no change. Since the descriptors were normalized for each patients data, the learned coefficients γ_x as in Eq. (4) indicate the significance of the corresponding descriptors in differentiating lesion change detection. Coefficients in Eq. (4) that are consistently large in magnitude suggest that they are especially important to generating the output lesion change probability map.

To study the importance of specific texture-based descriptors, a large set of coefficients were generated using a Monte-Carlo simulation. With repeated random sampling of image data from ten patients, 32 sets of coefficients were generated. A one-sample Wilcoxon signed-rank test was used to evaluate which coefficients consistently differ most from zero (null hypothesis that the median value is zero). The top ten descriptors ($p < 0.01$) were found to be homogeneity,

sum average, sum variance, standard deviation, contrast, dissimilarity, difference variance, difference entropy, short run-length emphasis (RE), and long run high gray-level emphasis (GE).

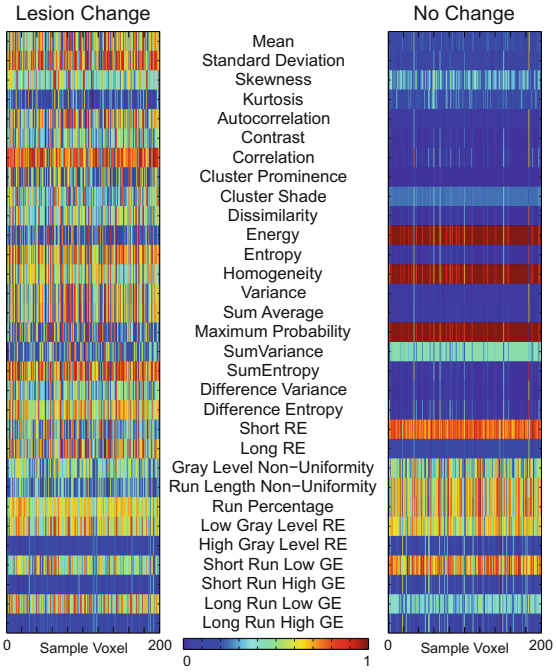

Fig. 3. Normalized descriptor values for sample voxels with lesion change (left) and with no change (right). The voxels were randomly selected from a sample patient, and all voxels are within the intracranial volume mask and voxel selection mask. Note the differences in descriptor values between the two classes, which are indicative of discriminative potential of these radiomic descriptors. In the descriptor index, GE stands for gray-level emphasis and RE stands for run-length emphasis.

4 Discussion

We proposed a multi-scale MS lesion change detection method, which had the highest performance when compared to three baseline methods that employed the multi-sequence MR images at fewer scales. The proposed method differs from others by extracting descriptors from not only the voxel itself, as in Ref. [8], which corresponded to Baseline 1, but also at two additional spatial scales, i.e. at immediate and extended neighborhoods. Since a radiologist annotates image voxels not in isolation but in the context of its surrounding voxels, it is intuitive that multi-scale context information improves lesion change detection.

Comparing Baseline 2 and the Multi-Scale Method, texture descriptors clearly contribute to the latter's higher performance. The Multi-Scale Method

and Base-line 3 had higher AUC-PR than Baseline 1 and 2, suggesting that texture-based descriptors improve the algorithms performance in the task of differentiating lesion change areas from non-lesion change areas. The top ten descriptors reported in the Results section (e.g. homogeneity, sum average) were especially critical in distinguishing between areas with and without lesion change, since their learned weights differ most from a median value of zero across multiple experiments. Extracting only these descriptors resulted in comparable performance (median AUC-ROC = 0.97, AUC-PR = 0.40) to using all 31 texture-based descriptors (median AUC-ROC = 0.97, AUC-PR = 0.43), which suggests that they may be sufficient for less computationally expensive use in a clinical setting.

While radiomic descriptors have made progress in tumor detection and segmentation for other diseases, this work is one of the first to use these descriptors in MS lesion change detection and highlights the importance of specific descriptors for this purpose. As expected, despite the significant contributions of texture descriptors, the intensities of the central voxel and its immediate neighborhood nonetheless contributed significantly to the performance. As per our experiments, the proposed Multi-Scale Method outperformed Baseline 3 due to these additional descriptors. This once again emphasizes the importance of a multi-scale approach. As a reference point, we also evaluated the coefficients directly provided by Ref. [8]. It resulted in a median AUC-ROC of 0.82, which was significantly lower than the results from training on our dataset as presented in Table 1, but the result is high enough to confirm the generalizability of such an approach, which can be attributed to the standard image preprocessing routines applied in both studies.

Quantitative longitudinal MS lesion analysis could provide insight into disease progression that occurs subtler and earlier than clinical markers like deterioration of physical movement. In the current clinical practice, lesion analysis is limited to recording lesion count and location. Manual annotation of the MS lesion change is challenging due to the number of sequences (T1, T2 and FLAIR in our study) that need to be taken into account when performing the labeling of each potential new lesion area. Moreover, the task becomes even more time and cost expensive in a setting of a large-scale clinical trials, where, due to the larger number of imaging timepoints of interests, it is crucial to have a consistent annotation in order to reliably evaluate the effectiveness of the tested treatment. Reliable automated methods such as our approach can be used to expedite and assist the radiologists daunting task of labeling lesion change for many patients and provide objective evaluations in a consistent and efficient manner.

To conclude, we proposed a multi-scale MS lesion change detection method, which incorporates not only information at voxel level, but also information from neighborhood and texture-based descriptors from the larger patch surrounding each voxel. The method statistically significantly improved over the state-of-the-art method. We also showed the importance of texture-based descriptors to effective lesion change detection, which to the best of our knowledge has not been explored in previous works.

Acknowledgments. This material is based upon work supported by Philips Healthcare.

References

1. NIH fact sheets - multiple sclerosis. https://report.nih.gov/nihfactsheets/ViewFactSheet.aspx?csid=103
2. Avants, B.B., Tustison, N.J., Stauffer, M., Song, G., Wu, B., Gee, J.C.: The insight ToolKit image registration framework. Front. Neuroinformatics **8**, 44 (2014). https://doi.org/10.3389/fninf.2014.00044
3. Haralick, R.M.: Statistical and structural approaches to texture. Proc. IEEE **67**(5), 786–804 (1979). https://doi.org/10.1109/PROC.1979.11328
4. Lesjak, Ž., Pernuš, F., Likar, B., Špiclin, Ž.: Validation of white-matter lesion change detection methods on a novel publicly available MRI image database. Neuroinformatics **14**(4), 403–420 (2016). https://doi.org/10.1007/s12021-016-9301-1
5. Lladó, X., et al.: Automated detection of multiple sclerosis lesions in serial brain MRI. Neuroradiology **54**(8), 787–807 (2012). https://doi.org/10.1007/s00234-011-0992-6
6. Mechrez, R., Goldberger, J., Greenspan, H.: Patch-based segmentation with spatial consistency: Application to MS lesions in brain MRI. https://doi.org/10.1155/2016/7952541
7. Smith, S.M.: Fast robust automated brain extraction. Hum. Brain Mapp. **17**(3), 143–155 (2002). https://doi.org/10.1002/hbm.10062
8. Sweeney, E.M., Shinohara, R.T., Shea, C.D., Reich, D.S., Crainiceanu, C.M.: Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. Am. J. Neuroradiol. **34**(1), 68–73 (2012). https://doi.org/10.3174/ajnr.A3172
9. Tang, X.: Texture information in run-length matrices. IEEE Trans. Image Process. **7**(11), 1602–1609 (1998). https://doi.org/10.1109/83.725367
10. Tustison, N.J., et al.: N4itk: improved n3 bias correction. IEEE Trans. Med. Imaging **29**(6), 1310–1320 (2010). https://doi.org/10.1109/TMI.2010.2046908
11. Zhang, Y.: MRI texture analysis in multiple sclerosis. https://doi.org/10.1155/2012/762804