Adaptive Voxel, Texture and Temporal Conditional Random Fields for Detection of Gad-Enhancing Multiple Sclerosis Lesions in Brain MRI

Zahra Karimaghaloo[1], Hassan Rivaz[2], Douglas L. Arnold[3]
D. Louis Collins[2], and Tal Arbel[1,⋆]

[1] Centre for Intelligent Machines, McGill University, Canada
[2] Montreal Neurological Institute, McGill University, Canada
[3] NeuroRx Research, Montreal, Canada

Abstract. The detection of Gad-enhancing lesions in brain MRI of Multiple Sclerosis (MS) patients is of great clinical interest since they are important markers of disease activity. However, many of the enhancing voxels are associated with normal structures (i.e. blood vessels) or noise in the MRI, making the detection of Gad-enhancing lesions a challenging task. Furthermore, these lesions are typically small and in close proximity to vessels. In this paper, we present a probabilistic Adaptive Multi-level Conditional Random Field (AMCRF) framework, capable of leveraging spatial and temporal information, for detection of MS Gad-enhancing lesions. In the first level, a voxel based CRF with cliques of up to size three, is used to identify candidate lesions. In the second level, higher order potentials are incorporated leveraging robust textural features which are invariant to rotation and local intensity distortions. Furthermore, we show how to exploit temporal and longitudinal images, should they be available, into the AMCRF model. The proposed algorithm is tested on 120 multimodal clinical datasets acquired from Relapsing-Remitting MS patients during multi-center clinical trials. Results show a sensitivity of 93%, a positive predictive value of 70% and average False Positive (FP) counts of 0.77. Moreover, the temporal AMCRF results show the same sensitivity as the AMCRF model while decreasing the FP counts by 22%.

1 Introduction

Multiple Sclerosis (MS) is one of the most common neurological disease in young adults. Conventional Magnetic Resonance Imaging (MRI) techniques, such as T2-weighted (T2) and Gadolinium-enhanced T1-weighted (T1) sequence are sensitive in detecting its white matter (WM) plaques known as lesions. Specifically, due to their ability to reflect areas of blood-brain barrier disruption and acute inflammations, Gad-enhancing lesions[1] lesions serve as a measure of disease activity. At present, the number of Gad lesions is a widely used MRI outcome

⋆ This work was supported by an NSERC CRD grant (CRDPJ 411455-10).
[1] We refer to Gad-enhancing lesions simply as Gad lesions hereafter.

K. Mori et al. (Eds.): MICCAI 2013, Part III, LNCS 8151, pp. 543–550, 2013.

parameter in MS clinical trials. Gad lesions are generally segmented manually, a laborious task which is subject to intra- and inter-rater variability and very expensive for clinical trials that involve enormous amounts of data from multiple centers. It is desirable to have an automatic segmentation method that is robust to data variability due to different scanners and protocols. Moreover, it is necessary for any automatic technique to have high sensitivity and low False Positive (FP) rate to be clinically relevant. Unfortunately, there exists huge variability in the size (as small as 3 voxels), texture, intensity and location of Gad lesions making the detection task very challenging. Furthermore, the presence of numerous non-lesional enhancements (e.g. blood vessels, MRI noise and partial volume effects) renders maintaining low FP rate a challenging task. Most of the existing methods for Gad segmentation described in the literature are either not fully automatic [1,2], or depend on non-conventional MRI acquisition sequences [2,3], or require prior segmentation of T2 lesions in order to remove the FPs [3,4]. In [5,6,7], conditional random field models are proposed for addressing this problem, which were shown to outperform standard MRF, SVM and linear regression models. However, these models incorporate mainly local, voxel-level features and FPs still remain. As Gad lesions are typically very small and noisy, higher order features could be integrated in order to express more complex patterns. However, computing such features for all enhancing voxels is computationally prohibitive. Also, since MS is a longitudinal disease, clinical trials often consist of multiple scans of each patient over time which can provide additional information to the manual raters by observing persistence of enhancements in scans acquired at least six months apart (Gad lesions are generally enhance for less than six months). No automated methods have explored how additional temporal information (if available) can be leveraged in Gad lesion segmentation for further removal of the possible FPs.

In this work, we present an Adaptive Multi Level Conditional Random Field (AMCRF) classifier which incorporates both local voxel level and robust higher order textural patterns into the model. Specifically, at the voxel level, a local CRF (with cliques of up to size 3) is developed to infer binary labels at each voxel (i.e. lesions/non-lesion). At this level, the classifier is tuned to be highly sensitive at the expense of additional FP detections. At the second level, voxels with the same label are grouped together to form lesion candidates. Each candidate is further analyzed by considering new textural patterns, derived from a larger neighborhood, along with its voxel-wise observations to differentiate true and false lesion detections. To this end, SPIN image and RIFT features [8] are explored, two texture descriptors that are invariant to rotation and local intensity distortions. In addition to removing false lesions, the AMCRF also refines the boundaries of lesions at the second level and as such, is adaptive. We also show effective ways to exploit the temporal information from a past or future time point (should it be available) into our AMCRF model to further improve our results. The temporal AMCRF outperforms other methods with a sensitivity of 93%, a positive predictive value of 75% and average False Positive (FP) counts of 0.60.

2 Method

2.1 Adaptive Multi-level CRF

Our single timepoint AMCRF model infers the posterior distribution of the labels in two levels: the first voxel-based level, and the second level which incorporates higher order texture information. The first level is similar to the model presented in [7], except that in addition to the unary and pairwise interactions, here triplet cliques are considered as well. We first describe these two levels at a single timepoint, and then present algorithms for exploiting temporal information, should it be available.

Let $\mathbf{x}_i \in \Re^d$ denote the observation vector (e.g. intensity values) at voxel i in the image and $y_i \in \{1, 0\}$ be a binary random variable indicating its label (e.g. lesion vs. non-lesion). Given a test image, X, the goal of a probabilistic classifier is to infer the posterior distribution of the labels given the observations, i.e $p(Y|X)$ where $X = \{\mathbf{x}_i\}_1^n$, $Y = \{y_i\}_1^n$, and n is the total number of voxels in the image. At the first level, we introduce a voxel-based CRF with cliques of size up to 3 to formulate p^v, the posterior of labels given the observations:

$$p^v(Y|X, \boldsymbol{\lambda}^v) = \frac{1}{Z}\, exp[-(\sum_{i=1}^{n} \boldsymbol{\lambda}_\phi^v\, \phi(y_i|\mathbf{x}_i) + \sum_{i,j\in N_i} \boldsymbol{\lambda}_\varphi^v\, \varphi(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j) + \sum_{i,j\in N_i} \boldsymbol{\lambda}_\delta^v\, \delta_2(y_i, y_j) + \sum_{i,(j,k)\in N_i} \boldsymbol{\lambda}_\psi^v\, \psi(y_i, y_j, y_k|\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)] \quad (1)$$

where Z is the partition function. ϕ , φ and ψ represent the voxel level potentials for the unary, pairwise and triplet cliques respectively. The smoothing constant $\delta_2(\cdot)$ is for penalizing discrepancies in the labels of neighbouring pairs. It is zero if the two labels are equal and is one otherwise. N_i represents the first order neighborhood of voxel i. The voxel level parameters $\boldsymbol{\lambda}^v$, modulate the effect of each term in the final decision and are learned at the training stage (Sec.3.1). ϕ, φ and ψ are modeled as:

$$\phi(y_i|\mathbf{x}_i) = -\log\ p(y_i|\mathbf{x}_i), \quad (2)$$

$$\varphi(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j) = -\log\ p(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j) \quad (3)$$

$$\psi(y_i, y_j, y_k|\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = -\log\ p(y_i, y_j, y_k|\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \quad (4)$$

Random Forest [RF] [9] can be used to model the probabilities in Eq.(2) to (4). A set of labeled voxels are inferred as the result of the voxel level analysis. At this stage, the goal is to capture all of the lesions at the expense of additional FPs.

So far only voxel-wise interactions are considered. However, the pure intensity at each voxel might be distorted due to the presence of noise or other artifacts such as partial volume. Hence, higher order textural patterns that are robust to local intensity distortions are incorporated in to the model to remove the possible FPs. To that end, at the second level, voxels with the same label are grouped together to form a set of lesion candidates. A bounding box (BB) is

considered around each candidate with a 2 voxel margin from each side, and a new CRF is constructed for the voxels inside the BB modeling voxel-wise interactions together with higher order textural patterns:

$$p(Y_{BB}|X_{BB}, \boldsymbol{\lambda}^l, \boldsymbol{\lambda}^l_\Omega) = p^v(Y_{BB}|X_{BB}, \boldsymbol{\lambda}^l)\ exp(-\sum_i \boldsymbol{\lambda}^l_\Omega\ \Omega(y_i|H(X_{BB}))) \quad (5)$$

where X_{BB} and Y_{BB} indicate the observations and labels inside BB. $p^v(Y_{BB}|X_{BB}, \boldsymbol{\lambda}^l)$ represents a set of voxel-wised cliques similar to Eq.(1) with a new set of modulating parameters $\boldsymbol{\lambda}^l$ which along with $\boldsymbol{\lambda}^l_\Omega$ are learned for the second level. $H(X_{BB})$ is the textural pattern derived from the region inside the BB. In principal, any textural feature can be used to represent $H(X_{BB})$. In this work, inspired by [8], we chose two novel descriptors: *SPIN image* and *RIFT* which are 2D histograms encoding the appearance pattern inside each BB based on its intensity and gradient orientation distributions. $\Omega(y_i|H(X_{BB})) = -log(p(lesion_{BB}|H(X_{BB}))$ which represents the likelihood of detecting a Gad lesion inside BB given $H(X_{BB})$. It should be noted that the SPIN image and RIFT descriptors are computationally expensive, but the proposed hierarchical framework computes them only at the second level where we are left with only a few candidates. During training, we apply the voxel level model to a subset of the training data to obtain a set of lesion candidates. Spin image and RIFT features are computed for each candidate and are saved as a *textural pattern dictionary* according to whether it is a true or false detection. At test time, we use a KNN classifier (e.g. K=100) to find the K closest match between the textural patterns of the test candidate and the ones in the dictionary. Specifically, Earth Movers Distance (EMD) [10] can be used to find the distance between textural patterns. The probability of $lesion_{BB} = 1$ is proportional to the number of true detections among the K nearest neighbours.

2.2 Leveraging Temporal Information

In clinical practice, temporal information can be available to help the rater detect Gad lesions. Let X^t and $\mathrm{X}^{t\pm m}$, respectively, denote the image at the current time point and the one acquired m months before or after. In this paper, we focus on the context where the temporal interval, m, is large enough such that, if a Gad lesion is enhanced in X^t, it is most likely not enhanced in $X^{t\pm m}$. In clinical practice, this typically translates to scanning intervals of 6 months or more (i.e. m=6). In order to incorporate this temporal information, at the voxel level of the AMCRF model, we use the voxel intensities of both X^t and $\mathrm{X}^{t\pm m}$ for all cliques. At the second level, in addition to comparing the textural pattern of the detected region at X^t with those in the dictionary, we also compare it with the textural pattern at the same location at $\mathrm{X}^{t\pm m}$. Hence, the second level is modeled as:

$$p(Y_{BB}|X^t_{BB}, X^{t\pm m}_{BB}, \boldsymbol{\lambda}^{l'}, \boldsymbol{\lambda}^{l'}_\Omega, \boldsymbol{\lambda}^{l,'}_\Gamma) = p^v(Y_{BB}|X^t_{BB}, X^{t\pm m}_{BB}, \boldsymbol{\lambda}^{l'})\ exp(-\sum_i \boldsymbol{\lambda}^{l'}_\Omega\ \Omega(y_i|H(X^t_{BB})))$$
$$exp(-\sum_i \boldsymbol{\lambda}^{l,'}_\Gamma\ \Gamma(y_i|G(X^t_{BB}, X^{t\pm m}_{BB}))) \quad (6)$$

where $\Gamma(y_i|G(X^t_{BB}, X^{t\pm m}_{BB})) = -log(p(Lesion_{BB}|G(X^t_{BB}, X^{t\pm m}_{BB}))$ and $G(X^t_{BB}, X^{t\pm m}_{BB})$ is the EMD between textural patterns of X^t and $X^{t\pm m}$ at the same location. A RF classifier is designed to model this term. As before, $\boldsymbol{\lambda}^{l'}$, $\boldsymbol{\lambda}^{l'}_{\Omega}$, and $\boldsymbol{\lambda}^{l,'}_{\Gamma}$ are modulating parameters learned in training. For a non-lesional enhancement at X^t, textural patterns are similar to the ones extracted from the same location at $X^{t\pm m}$ (compare Fig.1(e)-(f) to Fig.1(g)-(h)). This is typical for enhancing structures such as blood vessels, for example. However they look different for a lesional enhancement (compare Fig.1(m)-(n) to Fig.1(o)-(p)).

3 Experiments and Results

3.1 Parameter Learning and Inference

There are two sets of parameters in our model: the RF parameters used in ϕ, φ, ψ and Γ and the modulating parameters. RF parameters are learned separately for each clique. However, due to the complexity of the partition function, exact learning of the modulating parameters is intractable. In this work we used an iterative approach proposed by Taskar *et al.* [11] in order to find the modulating parameters.

In the inference stage, considering the CRF model at each level and its learned parameters, the most probable labeling is found. Graph Cuts are chosen to solve this optimization problem primarily because of their ability to find globally optimal solutions for binary classifications [12].

3.2 Data Pre-processing

The training and test data was acquired from multi center clinical trials with RRMS patients with varying numbers of Gad lesions, each located in different areas of the brain WM. Each acquisition was composed of five sequences: pre- and post-contrast T1, T2, PD and FLAIR. Therefore, our voxel-wise observation vector, $\mathbf{x}$, consists of the intensities of the above five modalities, WM and partial volume tissue priors[2] and spatial locations of each voxel. For the particular data set that we had access to, the "silver standard" manual labels were determined using a protocol where two trained experts label the data separately, followed by consensus agreement. Prior to classification, pre-processing steps including bias-field inhomogeneity correction as well as removal of non-brain regions from the MRI are performed [13]. Furthermore, all intra-subject sequences are registered to a common coordinate space and the intensity histogram of all sequences is normalized [14]. The training data consists of 1760 scans (880 pairs of two time points) from 160 different centers and testing is based on 120 scans (60 pairs of two time points) from a *separate* clinical trial consisting of 24 centers in order

[2] The WM prior is built based on statistical tissue frequencies of 152 manually labelled brains (ICBM 152). The PV atlas was built based on overlapping locations between Grey Matter and Cerebrospinal Fluid (CSF) atlases and WM and CSF atlases.

to examine the robustness of the method to different multi center trials. Before computing the statistics, any detected region with size 1 or 2 is deleted according to clinical protocol that requires Gad lesions to consists of at least 3 voxels. If at least three voxels of a lesion are classified correctly, it is counted as a TP, otherwise it is a False Negative (FN). Any candidate with size greater than two that does not correspond to an enhancing lesion is counted as an FP. Sensitivity ($\frac{TP}{TP+FN}$), Positive Predictive Value ($\frac{TP}{TP+FP}$) and average number of FPs are reported.

3.3 Single Timepoint Results

Figure 1 compares the higher order textural descriptors for a non-lesional (first row) and lesional (second row) enhancement. As it is observed, the proposed textural patterns look very different for false (Fig.1(e)-(f)) and true (Fig.1(m)-(n)) detections and hence when compared to the dictionary of textural patterns (by computing the EMD), they can be distinguished from each other.

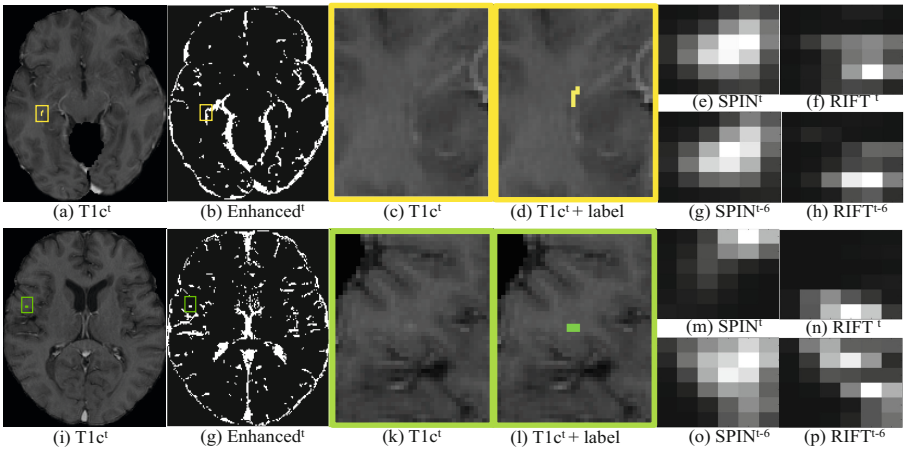


Fig. 1. Comparisons of the higher order textural features for a non-lesional and lesional enhancement. Post contrast T1 images are shown in the first column. Zoomed views are shown without and with labels in the third and forth columns. The second column shows all the high enhancement voxels. The textural features are shown for the detected regions at the current timepoint image and a previous timepoint.

In Table 1(a), we quantitatively compare the performance of the AMCRF model with the HCRF model proposed in [7] and an MRF. Here, a conventional MRF model is considered consisting of a unary clique and a non data dependant smoothing term (i.e. Eq.(1) without φ and ψ). For this experiment, a mask outlining the "new" enhancing voxels at the current time point is made available to all three methods. The results show that the proposed model has the highest sensitivity and PPV rate over all methods. Lower sensitivity in the MRF stems from the lack of observations in the smoothing term resulting in over smoothing

the small lesions. In HCRF, the high level features used in the second level were not robust enough to capture all lesions. We also show the overall performance of the AMCRF model along with its break down based on the size of the detected regions in Table 1(b). The AMCRF model achieves overall sensitivity of 0.93, PPV of 0.70 and average FP count of 0.77. As the size of the detected regions get larger both sensitivity and PPV values increase. Furthermore, Fig.2(a) shows that majority of the false detections are very small (i.e. less than 5 voxels).

Table 1. (a) Quantitative comparison of the performance of the AMCRF, HCRF and MRF models. (b) The performance of the AMCRF based on the voxel size.

(a)

	AMCRF	HCRF [7]	MRF
Sens	**0.93**	0.86	0.78
FPs	**0.77**	0.76	0.80
PPV	**0.70**	0.68	0.66

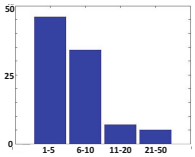
(b)

	overall	1-5	6-10	11-20	21-50	51-100	101^{+}
#Les	**231**	64	44	35	53	20	15
Sens	**0.93**	0.89	0.93	0.94	1	1	1
PPV	**0.70**	0.37	0.55	0.82	0.91	1	1

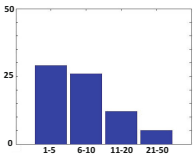
3.4 Temporal Model Results

Fig.1 shows qualitative results depicting the improvement caused by the higher order textural patterns when longitudinal data is available. The higher order textural features of a false detection at time t are very similar to those of the same location at time $t-6$ (compare Fig.1(e)-(f) to Fig.1(g)-(h)). On the other hand, the higher order textural features of a true detection at time t are very different from those of the same location at time $t-6$ (compare Fig.1(m)-(n) to Fig.1(o)-(p)). The classifier designed to capture these differences (by computing the EMD between the two textures) can appropriately distinguish true and false detections.

Fig.2(c) shows the quantitative comparisons of the AMCRF model with the temporal AMCRF. As it is observed, incorporation of the temporal data has increased the PPV value by 5% without changing the sensitivity. Also, the average number of false detections has been reduced by 22%. The histogram of the voxel size of the FPs is also shown in Fig.2(b). Once again we see that majority of the FP counts are small.



(a)



(b)

	AMCRF	temporal-AMCRF
Sens	0.93	0.93
FPs	0.77	0.60
PPV	0.70	0.75

(c)

Fig. 2. (a) and (b) show the histograms of the voxel wised size of the total FP detections for the AMCRF and temporal AMCRF models respectively. (c) is the quantitative comparison of the performance of the AMCRF and temporal AMCRF.

4 Discussion

In this paper, we propose a new a Adaptive Multi-level CRF (AMCRF) model to detect Gad lesions in brain MRI that embeds contextual information at multiple levels. At the first level, a local voxel-based CRF is used to identify candidate lesions. In the second level, a CRF model is designed to further examine the lesion candidates. We also proposed exploiting temporal data into our model. The temporal AMCRF outperforms other methods with a sensitivity of 93%, a positive predictive value of 75% and average False Positive (FP) counts of 0.60.

References

1. Miki, Y., et al.: Computer-assisted quantitation of enhancing lesions in multiple sclerosis: correlation with clinical classification. Am. J. Neur 18, 705–710 (1997)
2. Bedell, B., Narayana, P.: Automatic segmentation of Gadolinium-enhanced multiple sclerosis lesions. Magn. Reson. Med. 39, 935–940 (1998)
3. He, R., Narayana, P.: Automatic delineation of Gd enhancements on magnetic resonance images in multiple sclerosis. Med. Phys. 29, 1536–1546 (2002)
4. Datta, S., et al.: Segmentation of gadolinium-enhanced lesions on MRI in multiple sclerosis. J. Magn. Reson. Imag. 25, 932–937 (2007)
5. Karimaghaloo, Z., Shah, M., Francis, S.J., Arnold, D.L., Collins, D.L., Arbel, T.: Detection of gad-enhancing lesions in multiple sclerosis using conditional random fields. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part III. LNCS, vol. 6363, pp. 41–48. Springer, Heidelberg (2010)
6. Karimaghaloo, Z., et al.: Automatic detection of Gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields. TMI (2012)
7. Karimaghaloo, Z., Arnold, D.L., Collins, D.L., Arbel, T.: Hierarchical conditional random fields for detection of gad-enhancing lesions in multiple sclerosis. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 379–386. Springer, Heidelberg (2012)
8. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE PAMI 27, 1265–1278 (2005)
9. Leo, B.: Random forests. Machine Learning, 5–32 (2001)
10. Rubner, Y., Tomasi, C., Guibas, L.: The earth mover's distance as a metric for image retrieval. IEEE PAMI, 99–121 (2000)
11. Taskar, B., Chatalbashev, V., Koller, D., Guestrin, C.: Learning structured prediction models: a large margin approach. In: ICML, pp. 896–903 (2005)
12. Boykov, Y., Kolmogorov, V.: An experimental comparison of min- cut/max-flow algorithms for energy minimization in vision. PAMI, 1124–1137 (2004)
13. Smity, S.: Fast robust automated brain extraction. Hum. Brain Mapp., 143–155 (2002)
14. Nyul, L., Udupa, J.: On standardizing the MR image intensity scale. Comp. Assisted Tomography 42, 1072–1081 (1999)