Temporal Hierarchical Adaptive Texture CRF for Automatic Detection of Gadolinium-Enhancing Multiple Sclerosis Lesions in Brain MRI

Zahra Karimaghaloo*, Hassan Rivaz, *Member, IEEE*, Douglas L. Arnold, D. Louis Collins, and Tal Arbel, *Member, IEEE*

***Abstract*—We propose a conditional random field (CRF) based classifier for segmentation of small enhanced pathologies. Specifically, we develop a temporal hierarchical adaptive texture CRF (THAT-CRF) and apply it to the challenging problem of gad enhancing lesion segmentation in brain MRI of patients with multiple sclerosis. In this context, the presence of many nonlesion enhancements (such as blood vessels) renders the problem more difficult. In addition to voxel-wise features, the framework exploits multiple higher order textures to discriminate the true lesional enhancements from the pool of other enhancements. Since lesional enhancements show more variation over time as compared to the nonlesional ones, we incorporate temporal texture analysis in order to study the textures of enhanced candidates over time. The parameters of the THAT-CRF model are learned based on 2380 scans from a multi-center clinical trial. The effect of different components of the model is extensively evaluated on 120 scans from a separate multi-center clinical trial. The incorporation of the temporal textures results in a general decrease of the false discovery rate. Specifically, THAT-CRF achieves overall sensitivity of 95% along with false discovery rate of 20% and average false positive count of 0.5 lesions per scan. The sensitivity of the temporal method to the trained time interval is further investigated on five different intervals of 69 patients. Moreover, superior performance is achieved by the reviewed labelings of our model compared to the fully manual labeling when applied to the context of separating different treatment arms in a real clinical trial.**

***Index Terms*—Automatic segmentation, magnetic resonance imaging (MRI), multiple sclerosis (MS), probabilistic graphical models.**

Manuscript received October 15, 2014; revised December 09, 2014; accepted December 10, 2014. Date of publication December 18, 2014; date of current version May 29, 2015. This work was supported by a Canadian National Science and Engineering Research Council Strategic Grant (STPGP 350547-07) and a Canadian National Science and Engineering Research Council collaborative Research and Development Grant (CRDPJ 411455-10). *Asterisk indicates corresponding author.*

*Z. Karimaghaloo is with the Centre for Intelligent Machines, McGill University, Montreal, QC, H3A 2A7 Canada (e-mail: naghaloo@cim.mcgill.ca).

H. Rivaz is with the Electrical and Computer Engineering Department and PERFORM Centre, Concordia University, Montreal, QC, H3G 1M8 Canada (e-mail: hrivaz@ece.concordia.ca).

D. L. Arnold is with NeuroRx Research, Montreal, QC, H2X 3P9 Canada (e-mail: doug@neurorx.com).

D. L. Collins is with the Montreal Neurological Institute, McGill University, Montreal, QC, H3A 2B4 Canada (e-mail: louis.collins@mcgill.ca).

T. Arbel is with the Centre for Intelligent Machines, McGill University, Montreal, QC, H3A 2A7 Canada (e-mail: arbel@cim.mcgill.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2014.2382561

I. Introduction

MULTIPLE SCLEROSIS (MS) is a chronic inflammatory disease of the central nervous system that is common among young adults and is characterized by demyelinating lesions varying widely over time and spatial position. At the moment, there is no cure for MS. Magnetic resonance imaging (MRI) is widely used to diagnose and monitor the activity of this disease. Newly formed lesions are associated with enhancement after the administration of contrast agents containing gadolinium (hence called gadolinium-enhanced lesions or *gad lesions* in short). Disease activity in MS is generally quantified on MRI based on the frequency of these lesions. In fact, quantifying the frequency of gad lesions has led to new insights into the natural history of MS and, perhaps more importantly, has provided an objective measure of the disease activity for new anti-inflammatory MS therapies. As a result, gad lesion frequency is routinely used in clinical trials to provide biological evidence of drug efficacy [1]. These trials usually involve the assessment of thousands of scans at different timepoints from hundreds of patients, and so the effort required to detect the lesions is not trivial. However, at the moment, these lesions are often fully manually segmented by several raters. A task that is both time consuming and prone to intra and inter reader variability which hinders further robust statistical analysis of the results. Hence, robust automatic detection and segmentation of these lesions is highly desirable.

Automatic segmentation of pathologies is generally more difficult than segmentation of healthy structures due, in part, to the shortage of available shape, size, and location priors, and to the difficulty in modeling intensities and texture patterns over a population because of their large variability. In the context of gad lesion segmentation, the problem is further complicated because of their general small sizes and the huge variability in their appearance and location within the white matter. Some gad lesions are large and easy to detect while others are as small as three voxels. Some lack enough contrast to be identified on the contrast image alone without comparison to the pre-contrast image. Some are in the deep white matter while others are very close to the cortex. Furthermore, the presence of many nonlesional enhancements in the contrast image associated with normal structures such as blood vessels or noise in MRI makes the problem more challenging. Examples of lesional and nonlesional enhancements are shown in Fig. 1. Here, only the green rectangles correspond to true gad lesions.

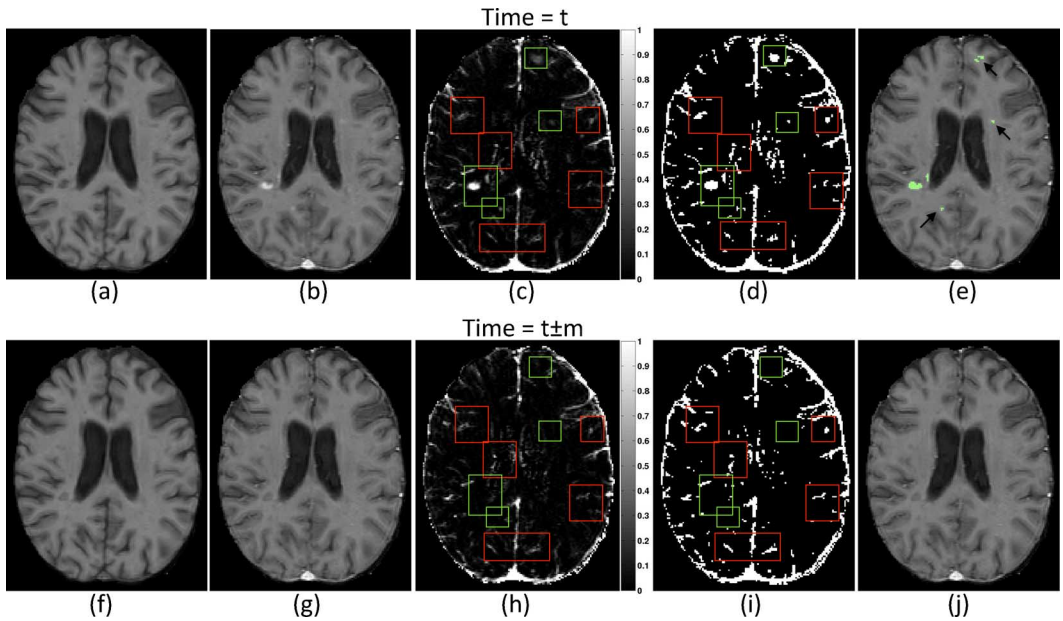

Fig. 1. Examples of brain MRI of patients with gad enhancing MS lesions. Top and bottom rows show two timepoints of the same patient m months apart (m is 6 here). First and second columns show axial slices of T1w images pre-contrast (T1p) and post-contrast (T1c). Third column shows the normalized enhancement map for all voxels computed as $(T1c - T1p)/T1p$. Brighter regions correspond to the stronger enhancements. Fourth column shows the enhancement map thresholded to only show voxels with required amount of enhancements (i.e., respectively more than 20% in our work) for gad lesions. Green and red rectangles correspond to lesional and nonlesional enhancements. Last column shows gad lesions as per the "ground truth" segmentation shown in green (there is no gad lesion in the second row image). In (e), arrows point to small and low contrast lesions that are hardly visible considering only the contrast image in (b).

Intensity alone is therefore insufficient in order to correctly distinguish the gad lesions from the pool of other enhancements [2], [3], therefore the inclusion of larger scale information from the surrounding neighborhood at different scales should improve the results. In addition, as MS is a chronic disease, there often exist multiple scans of the same patient over time in a clinical trial to monitor the disease activity over time and to evaluate the efficacy of a new drug. Specifically, there are several clinical studies investigating the enhancement duration of gad lesions [4], [5], which indicate that typically enhancements last less than six months. It is important to note that the enhancement indicates new lesion activity. If the interval between the two scan is around six months, the enhancement pattern of almost all lesions (even the small ones) changes, as old activities cease or new ones begin. In fact, two of the gad lesions shown in Fig. 1(e) have only three voxels. Comparison between Fig. 1(d) and (i) shows that these lesions have, indeed, changed over time. Therefore, leveraging the temporal data can provide an additional source of information increasing the discrimination power of the classifier.

In this work, we show how to combine this *spatio-temporal* information at different scales within a probabilistic graphical model. A probabilistic graphical model represents a structured probability distribution over a set of random variables following a graphical structure with its associated parameterization. Markov random field (MRF) and its discriminative variant, conditional random fields (CRF), are two widely used examples of probabilistic graphical models [6]–[13]. Accordingly, we present a temporal hierarchical adaptive texture CRF (THAT-CRF) where spatio-temporal patch-based features are incorporated to express more complex patterns. As considering patch textures are computationally prohibitive for the *entire MR volume* at *different scales*, we use these descriptors within a hierarchical approach, where candidate lesions are first specified by a voxel-wise CRF. The extensive texture analysis is then only performed on these selected candidates in the second level.

An overview of our proposed model is shown in Fig. 2. We first perform a CRF-based classifier to detect candidate lesions (step I). Here, unlike most of the CRF-based approaches where interactions are modeled only up to pairs, we incorporate higher order cliques of size three which can capture more complex interactions in the image. Once the candidate lesions are detected, stationary and temporal higher order features are calculated for the patches that contain the candidates (step II). Stationary features model the texture at the current timepoint, while temporal features model the textural profile across two timepoints. In modeling these features, we extensively explore the effect of robust descriptors (independently and combined) such as local intensity histogram description (spin image) [14], rotationally invariant feature transform (RIFT) [14], and local binary pattern (LBP) [15]. These descriptors encode intensity patterns and gradient orientation around a reference point and are invariant to rotation and local intensity distortion. A higher order CRF model is designed by combining the higher order textures and voxel-wise interactions and is applied to the candidate lesions to remove falsely detected regions (step III). This level is also adaptive in that the size and shape of candidate lesions are continuously refined. This is due to the more comprehensive patch-based information used in this level, in addition to the voxel-wise features used in the first level.

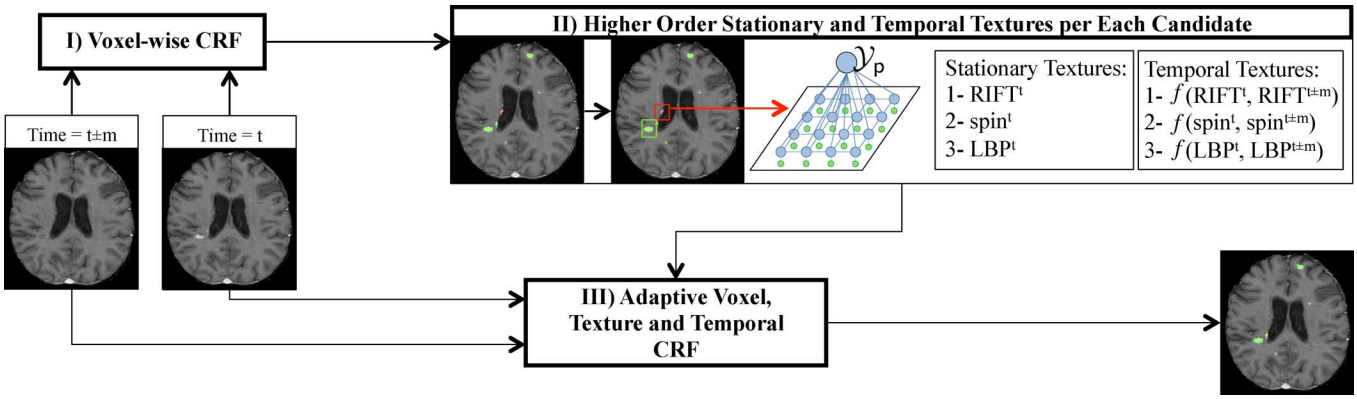

Fig. 2. Proposed THAT-CRF classifier. Lesion candidates are inferred as the result of the voxel-wise CRF (step I). Stationary and temporal higher order features are computed for the patches containing the detected candidats (step II). These higher order features are then combined with the voxel-wise interactions to remove the remaining false detections and also to further refine the boundaries of the correct detections (step III).

The proposed model is trained on 2380 relapsing remitting (RRMS) scans from a clinical trial and the effect of different proposed spatio-temporal textures is extensively evaluated on 120 RRMS scans from a *separate* clinical trial to evaluate the model robustness against different trials. Our analysis show that the incorporation of the spatio-temporal features at different scales results in a general decrease of the false discovery rate while maintaining a high sensitivity. The sensitivity of the temporal model to the time interval trained on is further investigated by testing on five different intervals of 69 patients. We also evaluated the performance of our model in separating different treatment arms (patients on the drug and those on placebos) as this is one of the primary goals of the clinical trials evaluating the efficacy of new drugs. Better separation is achieved by the reviewed labelings of our model (corrected only for the false detections) compared to the fully manual labeling. Therefore, the automatic framework is more sensitive at locating the gad lesions.

After describing previous work, we present our approach and then detail the experiments and results in the following sections.

A. Previous Work

In computer vision and medical imaging, many segmentation algorithms have been proposed for the context of segmenting a central object (e.g., a building or a healthy structure such as hippocampus) in an image or in a region of interest from a cluttered background [6]–[9], [16]. In these contexts, often rich features can be extracted from the objects to be segmented, based on color, intensity, or texture patterns, that render the object distinctive from the surrounding background. Moreover, location, size, and shape models can be learned and exploited in order to further improve the segmentation results. In the context of pathology segmentation, where the pathology of interest is large and there is only one in the image (e.g., brain tumors), techniques have exploited prior knowledge and texture information to delineate the pathology, particularly if one can leverage texture homogeneity within sub-regions [10], [11], [13], [12]. However, due to the aforementioned challenges associated to small pathologies, none of these methods are directly applicable to our problem. The most similar to our approach is the work of [12] where they include different hypothesis based on all image cues to train a single CRF framework. However, their framework highly relies on texture homogeneity within sub-regions and the performance is only shown on relatively large breast lesions. Hence, the efficacy of their approach in detecting small lesions is not clear.

The form of the higher order cliques defined in our model is very similar to the work of [9] and [7] in computer vision for the problems of object (such as a building, tree, or car) detection and scene understanding in natural images. However, their model focuses on the problem of multi-class segmentation and the type of the higher order cliques they exploit requires complex learning and inference algorithms. The higher order term proposed in this work is specifically tailored for binary classification problems and it can be easily decomposed to pairwise interactions. As a result, conventional learning and inference methods are readily applicable.

Most of the existing methods for gad lesion segmentation described in the literature are either not fully automatic [17], or depend on nonconventional MRI acquisition sequences [18], or require prior segmentation of T2 lesions in order to remove the falsely detected regions [19], [20]. In our previous work [2], [3] a single timepoint CRF-based classifier for detection of small pathology was developed and showed superior performance over several state-of-the-art classifications such as support vector machine (SVM) and traditional MRF when applied to the problem of gad lesion classification. However, only voxel-based features were used. We proposed a single timepoint hierarchical CRF based approach in [21] where simple patch-based statistics such as mean and standard deviation of a given neighborhood were used. However, false positive detections still remained. Preliminary work [22] showed some promise in leveraging temporal information into the model and we now further develop and explore this premise in the following section.

II. Methodology

A. Background

Let X and Y respectively represent a set of input variables (e.g., image intensities) and a set of output variables (e.g., labels in the task of image segmentation). A CRF models the conditional posterior probability distribution $p(Y|X)$ over a

predefined graph whose nodes and edges represent the voxels and their inter-dependencies. The conditional probability of the label Y given an image X can be written as a product of factors (also called potentials)

$$p(Y|X) = \frac{1}{Z(X)} \prod_{c \in C} \psi_c(\mathbf{y}_c, X, \lambda_{\mathbf{y}_c}) \tag{1}$$

where ψ_c is a factor function mapping its input variables to a non-negative real value. $\mathbf{y}_c$ is the output variables associated to the nodes in clique c and $\lambda_{\mathbf{y}_c}$ is the parameter associated with a configuration $\mathbf{y}_c$. Z is the partition function and C is the set of all possible cliques in the graph. It is generally assumed that each of the factors is a member of the exponential families $\psi_c(\mathbf{y}_c, X, \lambda_{\mathbf{y}_c}) = \exp(-\lambda_{\mathbf{y}_c} E(\mathbf{y}_c, X))$, where $E(\mathbf{y}_c, X) > 0$ represents the energy of a configuration $\mathbf{y}_c$. Hence (1) can be written as

$$P(Y|X) = \frac{1}{Z(X)} \exp(-\sum_c \lambda_c E_c) \tag{2}$$

where $\lambda_{\mathbf{y}_c}$ and $E(\mathbf{y}_c, X)$ are shortly written as λ_c and E_c.

In the following section we elaborate on different steps of the THAT-CRF model outlined in Fig. 2.

B. Voxel-Wise CRF

Following the general CRF model [(2)], we can formulate the problem of gad lesion classification by defining several cliques that capture different characteristics of lesions. At this level, we model voxel-wise interactions up to triplet cliques to obtain a set of candidate regions that are further explored at the next level through the incorporation of more complex terms. Let X^t and $\mathrm{X}^{t\pm m}$, respectively, denote the images of two timepoints of the same patient coregistered using [23]. 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 $X^{t\pm m}$ (i.e., $m = 6$ in this study). We incorporate the voxel-wise temporal information by using the voxel intensities of both X^t and $\mathrm{X}^{t\pm m}$ for all cliques. Hence, the probability of the configuration Y^t at time t given X^t and $X^{t\pm m}$ is

$$\begin{aligned} &p^v(Y^t|X^t, X^{t\pm m}, \boldsymbol{\lambda}^v) \\ &= \frac{1}{Z} \exp(-\sum_c \lambda^v_{\mathbf{y}_c} E^v_c(\mathbf{y}^t_c, X^t, X^{t\pm m})) \\ &= \frac{1}{Z} \exp\Big[-\Big(\sum_i \lambda^v_\phi \phi(y^t_i|\mathbf{x}^t_i, \mathbf{x}^{t\pm m}_i) \\ &\quad + \sum_{i,j\in N_i} \lambda^v_\varphi \varphi(y^t_i, y^t_j|\mathbf{x}^t_i, \mathbf{x}^t_j, \mathbf{x}^{t\pm m}_i, \mathbf{x}^{t\pm m}_j) \\ &\quad + \sum_{i,(j,k)\in N_i} \lambda^v_\psi \psi(y^t_i, y^t_j, y^t_k|\mathbf{x}^t_i, \mathbf{x}^t_j, \mathbf{x}^t_k, \mathbf{x}^{t\pm m}_i, \mathbf{x}^{t\pm m}_j, \mathbf{x}^{t\pm m}_k) \Big) \Big] \end{aligned} \tag{3}$$

where ϕ, φ, and ψ represent the voxel-wise potentials for the unary, pairwise, and triplet cliques, respectively. Superscript v denotes voxel-wise analysis. N_i represents the first-order neighborhood of voxel i (i.e., four in-plane and two out-of-plane neighbors). The voxel-level weights $\boldsymbol{\lambda}^v$, modulate the effect of each term in the final decision and are learned at training. It should be noted that a separate weight is assigned to different configurations of cliques. In a binary classification, each clique has 2^m different configurations where m is the size of the clique. For instance, for the unary clique $m = 1$ and for the pairwise and triplet cliques $m = 2$ and $m = 3$, respectively. As such, there are 2^1, 2^2, and 2^3 λ's associated to the unary, pairwise and triplet cliques. However, to prevent clutter, the dependency of λ to the labeling configuration is not explicitly shown in (3). The clique energies can be modeled as

$$\begin{aligned} \phi &= -\log p(y^t_i|\mathbf{x}^t_i, \mathbf{x}^{t\pm m}_i) \\ \varphi &= -\log p(y^t_i, y^t_j|\mathbf{x}^t_i, \mathbf{x}^t_j, \mathbf{x}^{t\pm m}_i, \mathbf{x}^{t\pm m}_j) \\ \psi &= -\log p(y^t_i, y^t_j, y^t_k|\mathbf{x}^t_i, \mathbf{x}^t_j, \mathbf{x}^t_k, \mathbf{x}^{t\pm m}_i, \mathbf{x}^{t\pm m}_j, \mathbf{x}^{t\pm m}_k). \end{aligned} \tag{4}$$

Essentially any classifier can be used to model the probabilities in (4). In this work, we choose to use random forest [24]. A random forest is a discriminative classifier that consists of an ensemble of decision tree classifiers, where the final classification is determined by summing the votes cast by each individual tree. Due to random selection of subset of training data and features, contrary to traditional decision trees, random forest is less prone to overfitting. Also it is computationally efficient both at the training and test and provides probabilistic outputs.

C. Candidate Region Detection

After the voxel-wise inference is completed, each voxel is assigned a probability of being gad lesion. These probabilities are then thresholded to obtain a binary result at each voxel (0 or 1). The goal at this stage is to capture all of the lesions (i.e., high sensitivity) at the expense of additional FPs. Therefore, the threshold is selected such that the highest sensitivity is achieved on the training data. Then candidate lesions are determined as neighboring set of voxels (defined by 26-connectedness in three dimensions) labeled as 1. A patch p whose size is proportional to the size of the detected region is considered around each candidate. The rational behind this is as follows: the selected patch should be large enough to include sufficient surrounding tissue to capture the contextual information, at the same time, it should not be too large to suppress the lesion related textures. Specifically, if a shows the size of a given side of the detected region's bounding box, the size of the corresponding side of the patch, a_p is set as

$$a_\mathsf{p} = \begin{cases} a+4 & a < 4 \text{ pixels} \\ 2a & \text{o.w.} \end{cases}. \tag{5}$$

In fact, we examined a few different sizes and the aforementioned patch size consistently yielded the best performance on the training data. Fig. 3(b) illustrates the patch selection for a given detected region. The region inside this patch is forwarded to the next level.

D. Higher Order Stationary and Temporal Textures

At this level, patches are examined more closely by considering higher order features besides voxel-wise interactions. Three types of higher order features are extracted for each patch: 1) spin image, 2) RIFT, and 3) LBP. These features are chosen

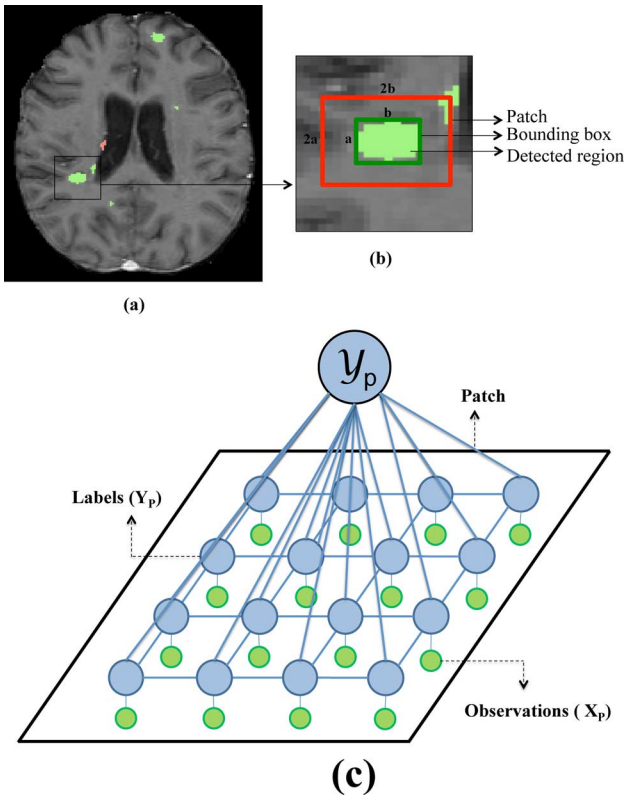

Fig. 3. (a) Post-contrast T1w image with the detected candidates of the first level. (b) Zoomed view of the detected region. The selected patch is shown in red. Size of the patch is twice that of the bounding box (in green). (c) Schematic view of our higher order clique defined over the voxels inside the patch. A new variable, $\mathcal{Y}_p$, is defined over all the variables inside the patch and its value is propagated to them.

due to their popularity and superior performance in computer vision applications [14], [25]. They lead to histograms encoding the appearance pattern inside each patch, based on the intensity content and both magnitude and orientation of gradients.

Spin image is a 2-D histogram encoding spatial and intensity information in a neighborhood of a particular reference point [14]. The neighborhood is a circular region divided into concentric rings (making the result rotation invariant) centred on the reference. The two dimensions of the histogram are distance from the center, and the normalized intensity [Fig. 4(a)]. To generate the spin image, the intensity values inside the patch are first normalized to the [0 1] range, resulting in invariance to spatial intensity changes (e.g., as a result of bias fields in MRI). Every pixel in the patch contributes to all histogram bins according to its location and normalized intensity using a Parzen window weighting.

RIFT is a 2-D histogram descriptor that encodes spatial information and gradient orientations weighted by the gradient magnitude [14]. Its main difference with respect to SIFT is that it considers the gradient orientations relative to the radial direction, and therefore is rotation invariant [Fig. 4(b)]. Each pixel inside the patch contributes to all histogram bins using a Parzen window weight. The Parzen window makes RIFT and spin image less sensitive to small deformations and intensity distortions.

LBP encodes local patterns at circular neighborhoods of a point. A label is assigned to every point in the image by thresholding its neighborhood points with its value and considering the result as a binary number [Fig. 4(c)]. The descriptor is the histogram of the labels of all the points inside the patch [15]. Specifically, we used the rotational invariant version of LBP that assigns a unique identifier to all possible rotations of the same pattern [15].

It should be noted that the three proposed textures are rotation invariant which is a desired property in our context since lesions can have different orientations. The spin image is complementary to RIFT as it encodes the *intensity* pattern of the given patch while RIFT is based on the *orientation of the gradients* inside the patch. Contrary to the spin image and RIFT which encode information relative to a reference point, the LBP encodes the pattern around *each* voxel inside the patch by assigning a code to it based on its eight neighbors. As such, it captures more local patterns compared to other two making it more sensitive to the pattern of small patches. Furthermore, all three features are scale invariant and also invariant to affine intensity distortions. The former is desired since lesions can have different sizes. The latter makes the results robust to intensity distortions happening due to bias field in homogeneities.

Fig. 5 shows examples of the above textures derived for a lesional and a nonlesional enhancement. Specifically, there are two types of higher order features: 1) *stationary textures* that correspond to the textures extracted at the given patch at time t, 2) *temporal textures* that correspond to the differences between the textures derived at time t and $t \pm m$. Accordingly, two separate classifiers based on the stationary and temporal features are learned in order to discriminate between the lesional and nonlesional enhancements.

Prior to defining these classifiers, we first define a new variable, $\mathcal{Y}_p$, that is connected to all voxels inside the patch under study through pairwise edges [Fig. 3(c)]. $\mathcal{Y}_p$ is 1 if the region contains a lesion and 0 otherwise. We now elaborate on the aforementioned texture classifiers.

1) Stationary Texture Classifier: The goal of this classifier is to model $p(\mathcal{Y}_p|H_p^t)$, where H_p^t is the higher order textural pattern derived from the patch at time t. In other words, we wish to determine whether H_p^t is more similar to the texture of a lesional enhancement or to that of a nonlesional enhancement. Our analysis showed that in order to distinguish between the stationary textures, it is more efficient to use a histogram-based distance metric, such as the earth movers distance (EMD) [26] as opposed to comparing their attributes one by one (as it is done by methods such as random forest). For this reason, we have adapted a kernel-based classifier such as the relevance vector machine (RVM) where the kernel matrix is computed based on the EMD distances between the stationary textures. RVM [27] is a Bayesian discriminant classifier and its main difference to SVM is that it provides a probabilistic output and usually results in a more sparse solution than SVM. This means it tends to generalize better and also needs less computations [27]. RVM models the probability distribution of the labels as

$$p(\mathcal{Y}_p|H_p^t) = \frac{1}{1+\exp(-f(H_p^t, \mathbf{w}_{\mathrm{rvm}}))} = \sigma(H_p^t, \mathbf{w}_{\mathrm{rvm}})) \quad (6)$$

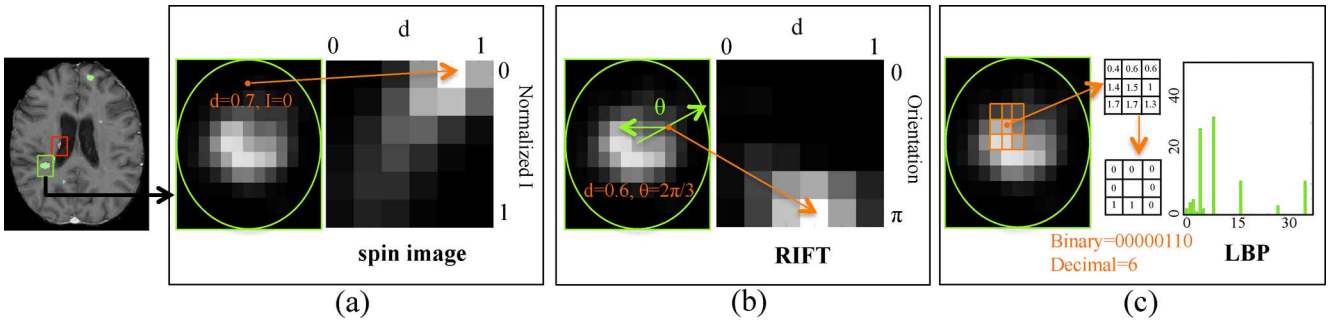

Fig. 4. Images (a), (b), and (c) illustrate computation of the spin image, RIFT and LBP textures for a lesional enhancement. In all three cases, the image on the left is the zoomed view of the region under the analysis and the image on the right is its associated texture. Contribution of a point inside the patch (marked with an orange circle) to each texture is shown.

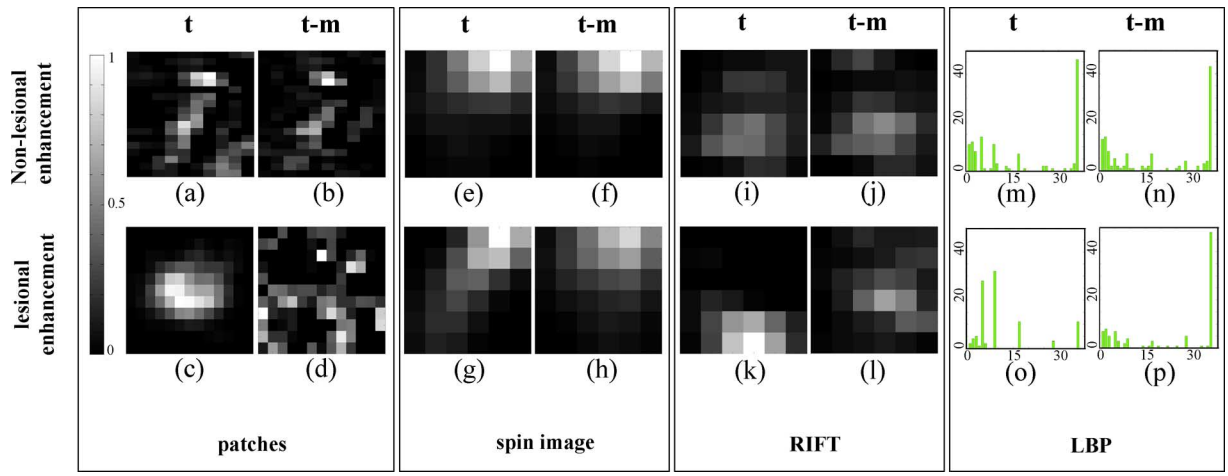

Fig. 5. Spin image, RIFT, and LBP features for a lesional and a nonlesional enhancement. All features are shown for both t and t-m (m = 6).

where the decision boundary is determined by the function f as follows:

$$f(H_{\mathsf{p}}^t, \mathbf{w}_{\text{rvm}}) = \sum_{j=1}^{N} w_j K_j(H_{\mathsf{p}}^t). \tag{7}$$

N is the total number of stationary textures used for training and $\mathbf{w}_{\text{rvm}} = [w_1, \ldots, w_N]^t$ is the set of weights. $\boldsymbol{K} = [K_1(.), \ldots, K_N(.)]^t$ is the kernel matrix comprised of N basis functions. More implementation details are given in Section III-E. Three types of stationary textures are used and hence three different stationary classifiers are learned: $p_{\text{tex}}(\mathcal{Y}_{\mathsf{p}}|H_{\mathsf{p}}^t)$ where $tex \in \{\text{spin image}, \text{RIFT}, \text{LBP}\}$.

2) Temporal Texture Classifier: Here, we wish to compare the texture of the selected candidate at time t and $t \pm m$. The idea is that lesional enhancements change over time whereas nonlesional enhancements do not change as much. Therefore, we expect their associated textures to follow the same trend as well. An example is shown in Fig. 5. Here, the spin image textures derived at t and $t-m$ for a nonlesional enhancement [i.e., (e) and (f)] are more similar than that of a lesional enhancement [i.e., (g) and (h)]. A similar trend is seen for the RIFT and LBP features as well.

Therefore, the goal is to evaluate $p(\mathcal{Y}_{\mathsf{p}}|f(H_{\mathsf{p}}^t, H_{\mathsf{p}}^{t\pm m}))$ where H_{p}^t and $H_{\mathsf{p}}^{t\pm m}$ are the textures derived from the same location at t and $t \pm m$ and f is a function that finds the distance between H_{p}^t and $H_{\mathsf{p}}^{t\pm m}$. It should be noted that here the feature vector is the *distance* between the two temporal textures and not the textures themselves (more details are given in Section III-E). Hence, we can directly compare the attributes of the feature vector. Therefore, similar to the voxel-wise potentials, a random forest classifier is used to evaluate this probability. As before, for each texture type, a random forest models is learned resulting in three different classifiers: $p_{\text{tex}}(\mathcal{Y}_{\mathsf{p}}|f(H_{\mathsf{p}}^t, H_{\mathsf{p}}^{t\pm m}))$ where $tex \in \{\text{spin image}, \text{RIFT}, \text{LBP}\}$.

E. IV- Adaptive Voxel, Texture, and Temporal CRF

For the voxels inside the patch, we now consider a second CRF model that includes both voxel-wise interactions and higher order stationary and temporal features

$$\begin{aligned} &p^{\mathsf{p}}(Y_{\mathsf{p}}^t|X_{\mathsf{p}}^t, X_{\mathsf{p}}^{t\pm m}, \boldsymbol{\lambda}^{\mathsf{p}}) \\ &= \frac{1}{Z}\exp\Big[-\Big(\sum_{c\in C_{\mathsf{p}}} \lambda_{\mathbf{y}_c}^{\mathsf{p}} E_c^v(\mathbf{y}_c^t, X^t, X^{t\pm m}) \\ &\quad + \sum_{i=1}^{N_{\mathsf{p}}}\Big(\sum_{\text{tex}} \lambda_{\Omega,tex}^{\mathsf{p}} \Omega_{\text{tex}}(y_i^t, \mathcal{Y}_{\mathsf{p}}|H_{\mathsf{p}}^t)\Big) \\ &\quad + \sum_{i=1}^{N_{\mathsf{p}}}\Big(\sum_{\text{tex}} \lambda_{\Gamma,tex}^{\mathsf{p}} \Gamma_{\text{tex}}(y_i^t, \mathcal{Y}_{\mathsf{p}}|f(H_{\mathsf{p}}^t, H_{\mathsf{p}}^{t\pm m}))\Big)\Big)\Big] \end{aligned} \tag{8}$$

where N_{p} denotes the total number of voxels inside the patch and p denotes patch level analysis. X_{p}^t and Y_{p}^t indicate the ob-

servations and labels inside the patch at t and $X_{\mathsf{p}}^{t\pm m}$ indicates the observation for the patch at $t \pm m$. C_{p} denotes unary, pairwise and triplet cliques inside the patch and E_c^v is their corresponding energy terms. These terms are similar to those used at the voxel-wise CRF [(3)]. The only difference is the usage of separate parameters ($\lambda_{\mathbf{y}_c}^{\mathsf{p}}$) that are learned together with the parameters of the higher order terms ($\lambda_{\Omega,tex}^{\mathsf{p}}$ and $\lambda_{\Gamma,tex}^{\mathsf{p}}$). Ω_{tex} and Γ_{tex} represent the higher order cliques reflecting the stationary and temporal texture classifiers. Essentially, like any pairwise energy, they encourage voxels inside the patch to have similar labels as $\mathcal{Y}_{\mathsf{p}}$ by assigning a lower energy to label agreements and a higher one to label disagreements. To that end, we define Ω_{tex} as

$$\Omega_{\text{tex}}(y_i^t, \mathcal{Y}_{\mathsf{p}}|H_{\mathsf{p}}^t) = \begin{cases} \mathcal{F}_{\text{tex}}^{\Omega}(\mathcal{Y}_{\mathsf{p}}) & y_i^t = \mathcal{Y}_{\mathsf{p}} \\ \mathcal{G}_{\text{tex}}^{\Omega}(\mathcal{Y}_{\mathsf{p}}) & \text{o.w.} \end{cases} \tag{9}$$

where

$$\begin{aligned} \mathcal{F}_{\text{tex}}^{\Omega}(\mathcal{Y}_{\mathsf{p}}) &= -\log p_{\text{tex}}(\mathcal{Y}_{\mathsf{p}}|H_{\mathsf{p}}^t) \\ \mathcal{G}_{\text{tex}}^{\Omega}(\mathcal{Y}_{\mathsf{p}}) &= -\log(1 - p_{\text{tex}}(\mathcal{Y}_{\mathsf{p}}|H_{\mathsf{p}}^t)) \\ &\forall\, tex \in \{\text{spin image, RIFT, LBP}\}. \end{aligned} \tag{10}$$

So, for example, if $p_{\text{tex}}(\mathcal{Y}_{\mathsf{p}}|H_{\mathsf{p}}^t)$ is high, then voxels are encouraged to have similar label to $\mathcal{Y}_{\mathsf{p}}$ because $\mathcal{F}_{\text{tex}}^{\Omega}$ has a lower value than $\mathcal{G}_{\text{tex}}^{\Omega}$.

Γ_{tex} is defined similar to (9) where

$$\begin{aligned} \mathcal{F}_{\text{tex}}^{\Gamma}(\mathcal{Y}_{\mathsf{p}}) &= -\log p_{\text{tex}}(\mathcal{Y}_{\mathsf{p}}|f(H_{\mathsf{p}}^t, H_{\mathsf{p}}^{t\pm m})) \\ \mathcal{G}_{\text{tex}}^{\Gamma}(\mathcal{Y}_{\mathsf{p}}) &= -\log(1 - p_{\text{tex}}(\mathcal{Y}_{\mathsf{p}}|f(H_{\mathsf{p}}^t, H_{\mathsf{p}}^{t\pm m})) \\ &\forall\, tex \in \{\text{spin image, RIFT, LBP}\}. \end{aligned} \tag{11}$$

It should be noted that in the CRF model used in [22] the higher order potentials are added in the form of unary terms [i.e., $\Omega_{\text{tex}}(y_i^t|H_{\mathsf{p}}^t)$ and $\Gamma_{\text{tex}}(y_i^t|f(H_{\mathsf{p}}^t, H_{\mathsf{p}}^{t\pm m}))$] whereas in this work we use a more complete graphical model by introducing pairwise edges between the image labels and higher order variables. More modulating parameters are associated to pairwise potentials as opposed to unary terms. Hence, the resulting model better represents the interactions between the variables.

Finding the parameters of (11) results in a new set of voxel-level parameters (different from those learned when considering the voxel-level cliques alone). As a result, the boundaries of the detected regions may change. Intuitively, if the higher order features show the presence of a gad lesion in the patch under study, voxel-level parameters become more relax and let more of the boundary voxels to be included.

F. Parameter Learning and Inference

1) Learning: We now consider the problem of learning the parameters of our model given a set of labeled training instances. We learned the parameters of the voxel-level CRF and patch level CRF independently. The standard maximum log-likelihood approach to find the optimum parameters at each level from K training cases is

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} \sum_{k=1}^{K} \log p(Y_k|X_k, \boldsymbol{\lambda}) \tag{12}$$

where $\boldsymbol{\lambda}$ is the set of model parameters and the log-likelihood of the kth example is (the subscript k is removed to avoid clutter)

$$\sum_{c\in C} -\lambda_{\mathbf{y}_c} E_c(\mathbf{y}_c|X) - \log(Z(\boldsymbol{\lambda})). \tag{13}$$

The partition function $Z(\boldsymbol{\lambda})$ is modeled as

$$Z(\boldsymbol{\lambda}) = \sum_{\mathbf{y}'\in\mathbb{Y}} \exp(-\sum_{c\in C} \boldsymbol{\lambda}_{\mathbf{y}_c} E_c(\mathbf{y}'_c|X)) \tag{14}$$

where $\mathbb{Y}$ is the set of all possible image configurations with cardinality of L^N (L: the number of classes, N: the total number of image voxels). Hence, computation of the partition function requires summation over an exponential number of terms making exact parameter learning intractable.

As a result, different approximation methods have been proposed to address this problem. In this work, we investigate pseudo log-likelihood parameter learning to obtain optimal parameters of the THAT-CRF model. Pseudo log-likelihood is a widely used approach to approximate the log-likelihood by factorizing it over the individual nodes of the graph. Hence (12) is written as

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} \sum_{k=1}^{K} \sum_{i=1}^{N} \log p_i(y_{i_k}|X_k, \boldsymbol{\lambda}) \tag{15}$$

where

$$\log p_i(y_{i_k}|X_k, \boldsymbol{\lambda}) = -\lambda_{\mathbf{y}_{c_i}} E_{c_i}(\mathbf{y}_{c_i}|X) - \log Z_i(\boldsymbol{\lambda}) \tag{16}$$

and

$$Z_i(\boldsymbol{\lambda}) = \sum_{\mathbf{y}'_i\in L} \exp(-\sum_{c_i\in C} \lambda_{\mathbf{y}_{c_i}} E_{c_i}(\mathbf{y}'_{c_i}|X)). \tag{17}$$

Here, p_i indicates the pseudo likelihood term and Z_i is the local partition function. Intuitively, instead of computing the partition function for an exponential number of label configurations, at each time we only allow the label of one voxel to change. This reduces the exponential complexity to a linear one.

2) Inference: Considering the CRF model at each level and its learned parameters, we now seek the most probable labeling that maximizes the conditional probability of (3) and (8). Here, we use iterated conditional mode (ICM). This yields probabilistic outputs and also makes our learning and inference consistent since both ICM and pseudo-likelihood consider local marginals to approximate the original intractable problem.

III. Experiments and Results

A. Data

Two different multi-center clinical data sets were used for training and testing. They contain subjects with relapsing-remitting MS (RRMS) with varying numbers of Gad-lesions located in different areas of the brain white matter. The training data set, *data set A*, is comprised of 2380 scans from 1190 subjects from 247 different centers. Each subject has two timepoints taken at six months intervals. Each timepoint has various multi-channel MRI data: pre-contrast (T1p) and post-contrast (T1c) T1-weighted images, T2-weighted (T2), proton density

weighted (PD), and fluid attenuated inversion recovery (FLR). As is typical of data used clinically in most centers, all scans were acquired axially with in-slice resolution of 1 mm and slice thickness of 3 mm. The test data set, *data set B*, comprised of two subsets, B_1 and B_2. Data set B_1 contains 120 pairs of scans at time t and $t \pm 6$ (all time intervals are in months) from 24 centers. Data set B_2 contains 69 patients from 27 centers all having the baseline scan and four other scans at months one, three, four, five, and six.

The same type of MRI sequences were available as for data set A. Different data sets from different clinical trials were used for training and testing in order to demonstrate the robustness of our model.

B. Preprocessing

Prior to analysis, several preprocessing steps are applied on MR data to remove nonbrain portions of the image, correct for nonuniformity effects and bring MR images into a common spatial and intensity space. All inter-subject modalities and timepoints are coregistered using a normalized cross correlation-based technique [23]. After coregistration, a deformable model-based skull-stripping algorithm called Brain Extraction Tool (BET) [28] is used to generate an initial brain tissue mask use for skull stripping. Finally, for each image volume intensity nonuniformity (NU) correction is done using N3 [29]. Intensity normalization is also performed within and across different subjects by a histogram matching technique to map each patient into a global intensity space. This allows us to learn classifiers based on the intensity across many subjects. This normalization was done using the method of [30], using intensity deciles as control timepoints.

C. Ground Truth

Gad-lesions in both data set A and B were available as a ground truth reference for training and evaluating the performance of our classifier. Specifically, these manual labels were determined using a protocol where two *trained experts* separately labeled the gad lesions using a software that displays coronal, transverse and sagittal slices all together. The silver standard ground truth was then generated by consensus agreement among them. If the two could not reach an agreement, the case was then reviewed by a third *highly trained expert* who made the final decision. These precisely controlled labels have been used for several real clinical trials of MS treatments. All MRI modalities (T1p, T1c, T2, PD, FLR) were consulted during the manual segmentation, as were timepoints from the same patient other than the one being considered (including, prior and subsequent scans).

D. Training Setup

In order to verify the robustness of our model to different training sets, three models were learned by selecting two thirds of the whole training data at random and repeating this three times. This allowed us to analyze the sensitivity of our classification results to different training sets. For learning each of these three models, the available training data itself was divided into four subsets: $s1$, $s2$, $s3$, and $s4$ to avoid over learning. First, the parameters of the potential functions of the voxel-level CRF [i.e., the random forest classifiers used to evaluate the probabilities in (4)] were learned with $s1$. Learning the voxel-wise classifier CRF^v was then completed by learning the $\boldsymbol{\lambda}^v$ modulating parameters using $s2$. CRF^v was then run on $s3$ to generate candidate lesions. Both stationary and temporal higher order texture features were inferred for the candidates. Given these features, the RVM and random forest classifiers used to evaluate probabilities in (10) and (11) were learned. Finally, $s4$ was used to learn the modulating parameters of the second level, $\boldsymbol{\lambda}^{\mathrm{p}}$. Learning different parts of the model with separate data decreased the chance of over fitting.

E. Implementation Details

The voxel-wise feature vector for the unary, pairwise and triplet interactions comprised of the voxel's intensity in all five MRI modalities[1] at t and $t \pm m$, the x, y, z location of the voxel and the value of three tissue priors: the white matter, partial volume, and T2 MS lesion priors. The white matter prior is estimated by registering the icbm152 (MNI) average brain atlas [31] to the patient images. The partial volume prior models mixtures of Cerebrospinal fluid and gray matter priors from the icbm152. The T2 MS lesion prior was available from [32] where manual T2 MS lesion segmentations for a series of clinical trials (totaling 3714 RRMS scans) were all nonlinearly registered to icbm152 space to provide a probabilistic lesion atlas. All registrations were performed using the method of [23].

All of the random forest classifiers used in (4) and (11) contain 100 decision trees. The number of variables to sample are half of the total available training data and the number of used features are half of the available features.

The stationary texture feature vector for the multi-modal RVM classifier in (10) comprised of the textures of the candidate location in all five MRI modalities, i.e., $H^t_{\mathrm{p}} = [h_{t1p}, h_{t1c}, h_{t2}, h_{flr}, h_{pd}]$ [Fig. 6(a)]. For simplicity, we denote H^t_{p} as H from now on. Gaussian kernels are used for the basis function in (7). Hence for a given texture H, the jth basis is written as

$$K_j(H) = \exp\left(-\frac{Dist(H_j, H)}{2\sigma^2}\right) \tag{18}$$

where H_j is the jth stationary texture feature vector from the training set. The distance between any two feature vectors is defined as the sum of the distances between their individual components

$$Dist(H, H_j) = \sum_{d=1}^{5} \mathrm{EMD}(h_d, h_{j,d}) \tag{19}$$

where d is one of the five MRI sequences: $d \in \{\mathrm{T1p}, \mathrm{T1c}, \mathrm{T2}, \mathrm{FLR}, \mathrm{PD}\}$. We use the EMD [26] to evaluate the distance between two textures in each MR modality. EMD finds the minimum cost required to transform one histogram into another. The cost is the histogram mass moved times a weight associated with the distance between the two bins. We use EMD-L1 [33] in this work where the weight between two bins is their L1 distance. EMD avoids

[1]It should be noted that the intensities of different MR modalities are all used together resulting in multi-modal classifiers.

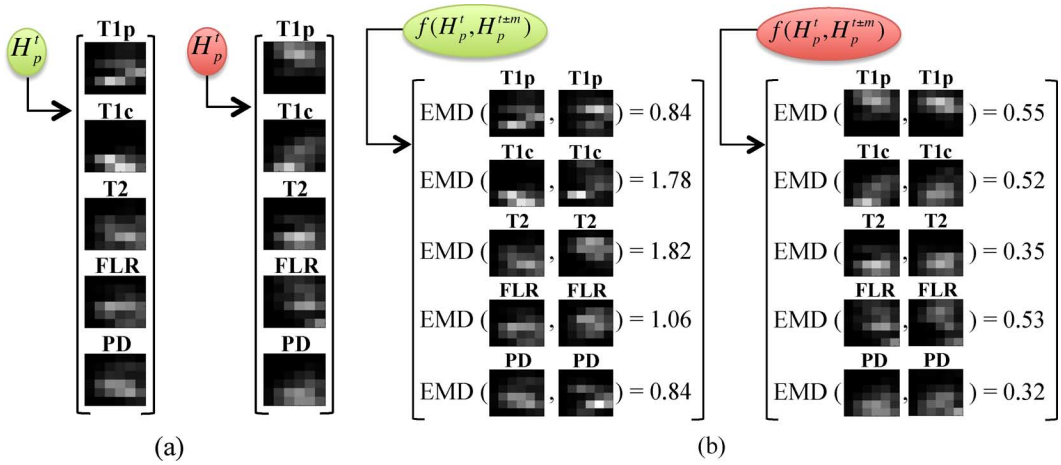

Fig. 6. Higher order feature vectors for a lesional (green) and a nonlesional (red) enhancement in all five MRI sequences computed for the RIFT texture. (a) and (b) Stationary and temporal features respectively. Note that in (b), the distance between features at different timepoint are significantly larger for lesional enhancements as compared to nonlesional enhancements. This is due to the temporal change of MS lesion activity in that time interval (m). $H_{\mathsf{p}}^{t} = [h_{t1p}, h_{t1c}, h_{t2}, h_{flr}, h_{pd}]$. (a) Stationary higher order feature vector. (b) Temporal higher order feature vector.

quantization and binning problems associated with histograms, and has been shown [33] to outperform other histogram comparison techniques.

The temporal texture feature vector for the multi-modal random forest classifier in (11) comprised of the distance between the individual components of H_{p}^{t} and $H_{\mathsf{p}}^{t\pm m}$

$$f(H_{\mathsf{p}}^{t}, H_{\mathsf{p}}^{t\pm m}) = \{\mathrm{EMD}(h_d^t, h_d^{t\pm m})\} \quad \forall d \tag{20}$$

where $d \in \{\mathrm{T1p},\ \mathrm{T1c},\ \mathrm{T2},\ \mathrm{FLR},\ \mathrm{PD}\}$. This feature vector is shown for a lesional and a nonlesional enhancement in Fig. 6(b).

Due to the larger spatial resolution of our data set in the transverse direction, all higher order textures are computed on 2-D axial slices. The final value of $p_{\mathrm{tex}}(\mathcal{Y}_{\mathsf{p}}|H_{\mathsf{p}}^{t})$ and $p_{\mathrm{tex}}(\mathcal{Y}_{\mathsf{p}}|f(H_{\mathsf{p}}^{t}, H_{\mathsf{p}}^{t\pm m}))$ for a detected region is obtained by a weighted sum of the probabilities in the individual slices it spans over. The weights are the ratio of the area in that slice to the over all volume of the detected area.

F. Experimental Validation

Fig. 7 shows three qualitative examples of gad lesion classification results of the proposed THAT-CRF method. There exists one gad lesion in each shown example that is successfully captured by the THAT-CRF model. Figs. 8 and 9 present two different examples illustrating the adaptive aspect of the THAT-CRF model in refining the boundaries of the detected regions. It should be noted that if at the second level (similar to commonly used hierarchical methods) we had only accepted or rejected the detected regions based on their textures, we would have obtained the labeling shown in Fig. 8(c) as the final results. Since this labeling consists of only two voxels, it would have been removed according to a clinical protocol, which stipulates that MS lesions must have at least three voxels. Hence, the lesion would have been missed. However, by considering the higher order textures together with the voxel-level cliques, we permit more of the boundary voxels to be included if the textures show the presence of a lesion.

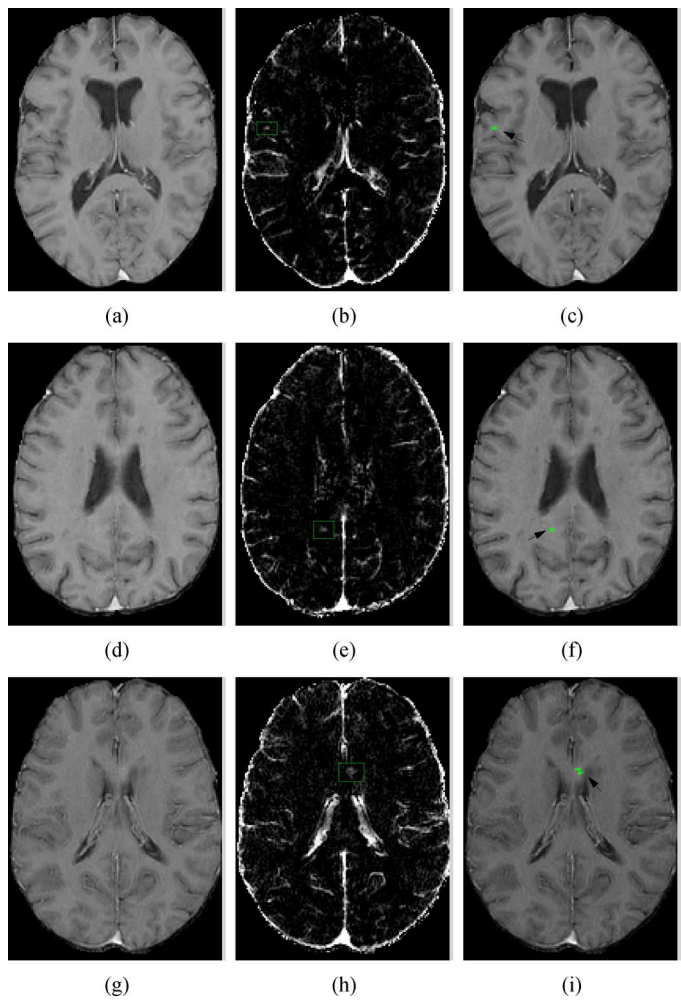

Fig. 7. Qualitative results of the performance of our proposed classifier. Each row shows an example image from a different patient. First column shows the post-contrast T1w images. Second column shows the enhancement map with the enhanced regions corresponding to the lesion detected in the last column marked with green rectangles. Third column shows the classification result of the THAT-CRF where arrows point to the gad lesions shown with green labels.

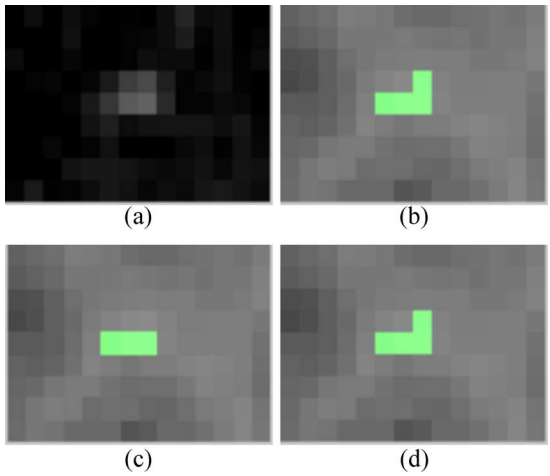

Fig. 8. Illustrating the adaptive aspect of the model. Images are the zoomed view of the example shown in the first row of Fig. 7. Output of the classifier at different stage of the THAT-CRF model is shown in green. CRF^{v} (unary) is when only the unary term is used. CRF^{v} (unary, pairwise, triplet) is the final result of the voxel-level analysis with inclusion of unary, pairwise and triplet cliques. Finally, THAT-CRF shows the final output of the model. (a) Enhancement. (b) CRF^{v} (unary). (c) CRF^{v} (unary, pairwise, triplet). (d) THAT-CRF.

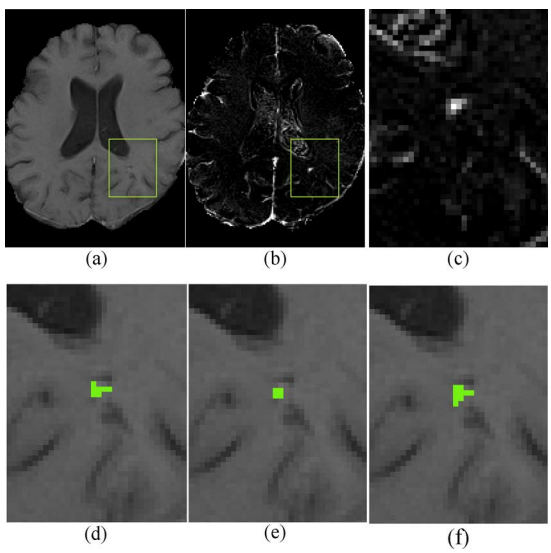

Fig. 9. Illustrating the adaptive aspect of the model. Output of the classifier at different stage of the THAT-CRF model is shown in green. CRF^{v} (unary) is when only the unary term is used. CRF^{v} (unary, pairwise, triplet) is the final result of the voxel-level analysis with inclusion of unary, pairwise and triplet cliques. Finally, THAT-CRF shows the final output of the model. (a) Post-contrast T1. (b) Enhancement. (c) Zoomed view. (d) CRF^{v} (unary). (e) CRF^{v} (unary, pairwise, triplet). (f) THAT-CRF.

We perform different experiments to measure the effect of different aspects of our model. All performances are evaluated by comparing to our ground truth reference. Voxel-wise metrics such as Kappa or Dice coefficient have been previously used to measure the performance of lesion segmentation algorithms [34]–[36]. However, these metrics have a strong bias toward larger lesions as small lesions contribute little to an overall volumetric measure [32], [37]. Moreover, due to the pathological nature of the lesions, their exact boundaries are often ambiguous and even an expert has been shown to label the boundaries differently if asked to perform the task twice. In addition, the over all gad lesion $counts$ is the outcome measure typically used in clinical trails. As such the primary focus of our work is to detect the individual lesions. As a result, voxel-wise metrics are not directly applicable to our problem, so we use lesion-wise metrics (such as counts) to evaluate our classifier.

In order to measure the performance of the output of a classifier at any step, gad lesions are determined as a neighboring set of voxels defined by 26-connectedness in three dimensions.

Prior to computing the metrics, regions with size 1 or 2 voxels are removed to comply with the aforementioned protocol that lesions should have at least size 3 (the same criteria was used for the manual labeling). Since our primary focus is to $detect$ all the gad lesions, we consider a detected region as a true positive (TP) if it has at least one voxel overlapping with our ground truth, otherwise it is counted as a false positive (FP). Any labeling in the "ground truth" that is not detected by our method is defined as a false negative (FN). Overall sensitivity and false discovery rate (FDR) are evaluated as: $TP/(TP+FN)$ and $FP/(TP+FP)$. By varying the acceptance threshold on the final probability at each voxel, we present our results in the form of a ROC-like curve by plotting the sensitivity versus FDR. It should be noted that our FDR definition is different from what is commonly used in the conventional ROC curves, where it is defines as the proportion of the false discoverys over all negative counts (i.e., $FP/(TN+FP)$ where TN is the true negative counts). As computing TN counts is not feasible in this context, the definition is modified here.

We compare our proposed temporal model, THAT-CRF, against 1) its single timepoint variant i.e., hierarchical adaptive texture CRF (HAT-CRF) and 2) its semi temporal variant (semi THAT-CRF). HAT-CRF is a single timepoint classifier that does not consider any temporal data when classifying the scan at time t. That means the voxel-wise feature vector consists of the intensity information only at time t, and the lesion-level higher order features consists of only the stationary features, i.e., no Γ term in (8). We previously proposed a simpler version of the HAT-CRF model in [21]. However, the HAT-CRF model used in this work for comparison is more complete as it uses up to triplet clique potentials at the voxel-level (as opposed to pairwise interactions used in [21]) and it includes robust higher order textures (as opposed to simple patch-based statistics used in [21]). Moreover, the parameters of the HAT-CRF model are learned using a substantially larger training data set than the one used in [21]. In the semi THAT-CRF, only *new* enhancing voxels at time t compared to timepoint $t \pm m$ are considered, and therefore its mathematical expression is the same as that of the HAT-CRF model. The only difference is that instead of starting with all of the enhancements at time t [as in Fig. 1(d)], only new enhancements compared to $t \pm m$ are considered. In other words, all voxels that are enhanced in *both* Fig. 1(d) and (i) are removed in the semi-temporal analysis. Comparison to the semi THAT-CRF is performed to highlight the necessity of the temporal patch based comparison as apposed to easier alternative of removing common enhancements between the two timepoints. The very local voxel-wise comparison performed by the semi THAT-CRF model makes it more sensitive to slight temporal misregistrations while the patch based comparisons carried in the THAT-CRF model

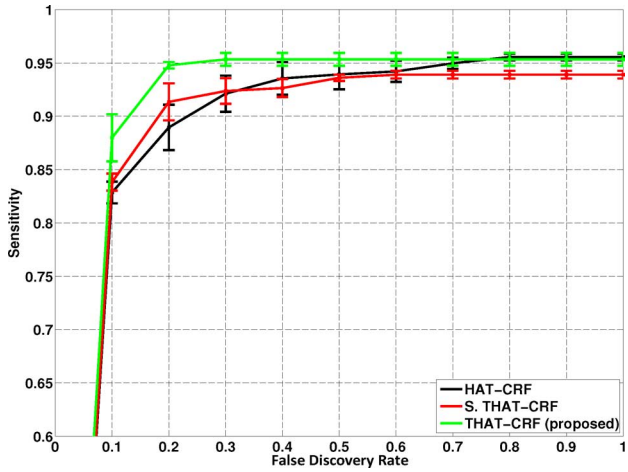

Fig. 10. Sensitivity versus false discovery rate for the proposed THAT-CRF, semi THAT-CRF (S. THAT-CRF), and HAT-CRF. Individual points on the curve were generated by varying the acceptance threshold on the final classification results at each voxel.

TABLE I
Mean sensitivity ± standard deviation at Two False Discovery Rates for the THAT-CRF Along With the Semi THAT-CRF (S. THAT-CRF) and HAT-CRF Classifiers

	HAT-CRF	S. THAT-CRF	THAT-CRF
Sens@ FDR = 0.2	0.889±0.021	0.913±0.017	**0.948±0.003**
Sens@ FDR = 0.1	0.828±0.010	0.838±0.008	**0.880±0.022**

provide robustness to slight mis alignments between the two timepoints. As such the proposed THAT-CRF model yields higher sensitivities.

Several experiments are conducted for comparing the aforementioned models:

1) Evaluating the Effect of Temporal Information: Fig. 10 shows a plot of sensitivity versus false discovery rate for the final output of the the proposed temporal model (in green), the semi-temporal model (in red), and the single timepoint classifier (in black) using dataset B_1. Different points on the curve were generated by varying the target sensitivity for the final result. Points on each curve show the mean classification sensitivity along with its standard deviation as a function of the false discovery rate for the three learned models. Curves have been extrapolated (as a straight line at maximum achievable sensitivity for each classifier) to cover the whole range of false discovery rates even if in practice these false discovery rates are not achievable.[2] Sensitivity at the final operating points of interest are shown in Table I. The results show that the semi THAT-CRF slightly improves the results over the single-time timepoint classifier. The THAT-CRF technique further improves the performance by incorporating patch-based stationary and temporal higher order features. It should be noted that even though only one voxel overlap with the ground truth is required for a TP detection, on average more than 95% of the lesion area is captured in all three classifiers.

[2]This is mainly performed to facilitate computing the mean curve for reporting the mean performance of the three trained models. Also, it facilities comparing different curves present in each plot.

TABLE II
Mean Sensitivity Sensitivity for the a Fixed Threshold of 0.88

	HAT-CRF	S. THAT-CRF	THAT-CRF
Sens	0.877	0.894	0.948
FDR	0.155	0.176	0.196

While Table I compares the performance of the three models for a fixed sensitivity (which translates into a different acceptance threshold for the three methods), it is also informative to compare their performance for a fixed acceptance threshold. The optimal threshold can be found by choosing the desired operating point along the curves. For instance, in the context of pathology detection, it is usually desirable to achieve high sensitivities ($> 90\%$ or 95%) while keeping the false discovery rate as low as possible. The threshold yielding this result for the three models is 0.88. Results are summarized in Table II. A fixed threshold would result in various sensitivities and false discovery rates for the three models. However, still the THAT-CRF model yields the highest sensitivity, in expenses of a slightly higher false discovery rate .

2) Evaluating the Effect of Different Potential Functions in CRF: Experiments were performed to evaluate the effect of each term in the HAT-CRF [Fig. 11(a)], semi THAT-CRF [Fig. 11(b)], and THAT-CRF [Fig. 11(c)] classifiers using dataset B_1. Results are shown in the form of plots of the sensitivity as a function of the false discovery rate. Similar to the previous section, each curve shows the mean performance over the three learned models. For readability, the standard deviations are not shown.

Several conclusions can be drawn from the results. 1) In the voxel-level, incorporating cliques of up to size three improves the sensitivity of the results (which is the main goal of the voxel-level) compared to cliques of up to size two[3] [see the dashed lines in Fig. 11(a)–(c)]. 2) Exploiting stationary texture features in the second level improves the results of the voxel-level [compare all the solid lines versus the dashed lines in Fig. 11(a) and (b)]. 3) Embedding the combination of all three stationary texture features (RIFT, spin image and LBP) leads to the best performance [see Fig. 11(a) and (b)]. 4) The proposed THAT-CRF technique gives the best results compared to both the HAT-CRF and semi THAT-CRF [all three stationary texture features are used in models in Fig. 11(c)]. 5) Similar to the case of stationary features, using the combination of temporal RIFT, spin image and LBP texture features results in the best performance. It should be noted that the highest achievable sensitivity of the final result in all three models is determined based on the working point of their corresponding voxel-level CRF; this is 0.95 in our experiments and point of saturation of the curves.

Curves with the best performance in Fig. 11(a) to (c) (marked in black) were shown together in a single plot in Fig. 10 for the ease of comparison.

[3]It should be noted that in computing the statistics of the voxel-level CRF, thresholds resulting in acceptance of all of the voxels in a given MRI volume were excluded because based on our TP and FP definitions they will be counted as one TP resulting in overall sensitivity of one and FDR of zero. For this reason the maximum achievable sensitivity for the voxel-level curves is not one.

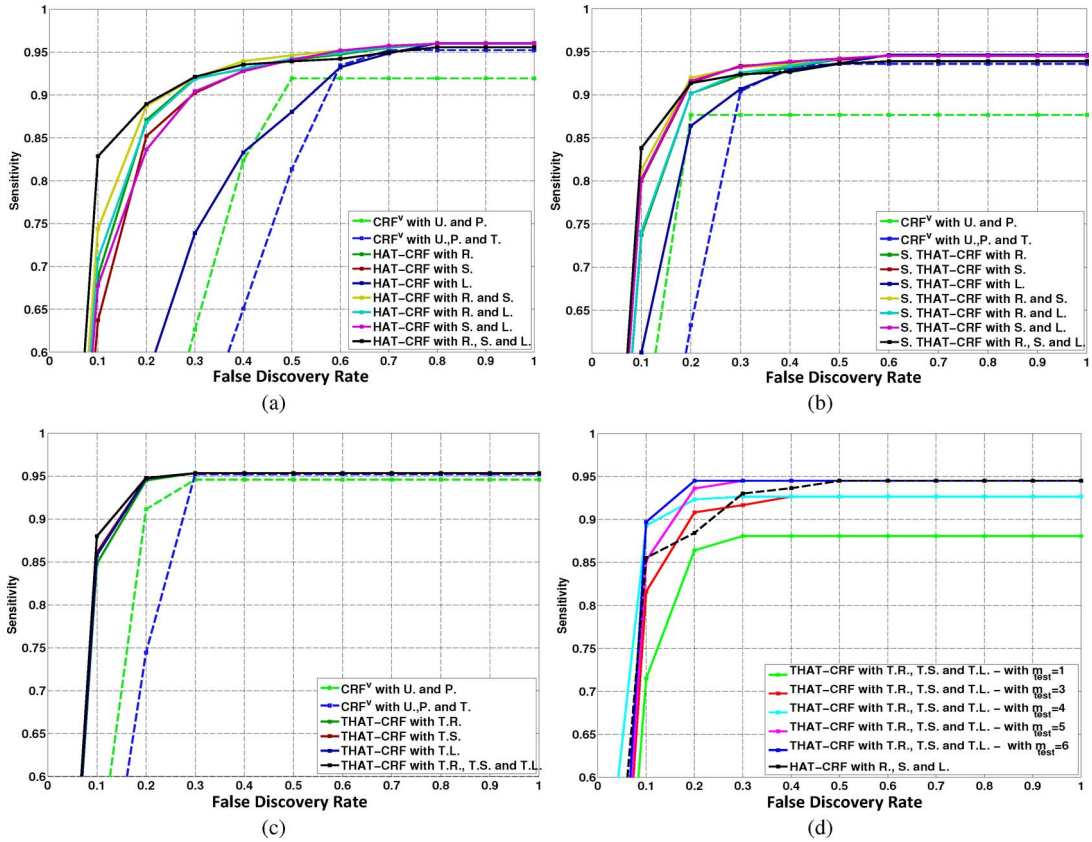

Fig. 11. Plot of sensitivity versus false discovery rates for (a) HAT-CRF, (b) semi THAT-CRF (S. THAT-CRF), and (c) THAT-CRF models. (a) and (b) Compare performance of the voxel wise CRF (CRF^v) with unary, pairwise and triplet cliques (U., P., and T.) as well as the final classification results with combinations of RIFT, spin image and LBP features (R., S., and R.). In (c), aside from the voxel-wise comparisons, the performance of THAT-CRF with temporal textures namely temporal RIFT, spin image and LBP features (T.R., T.S., and T.L.) is shown. All the curves in (a) to (c) show the mean performance of the three learned models on the test subset B_1. Individual points on each curve show the mean value over the three learned models and are generated by varying the acceptance threshold on the final probability at each voxel. (d) Plot of sensitivity versus false discovery rates of the THAT-CRF model for different testing time intervals. Model is trained with $m = 6$ and is tested with varying m_{test} intervals ($m_{\mathrm{test}} \in \{1, 3, 4, 5, 6\}$). Performance of the HAT-CRF model is also shown for comparison.

TABLE III
PERFORMANCE OF THE THAT-CRF MODEL FOR DIFFERENT LESION SIZES AT TARGET SENSITIVITY OF 95%. TP = TRUE POSITIVE COUNTS, SENS = SENSITIVITY, FP = FALSE POSITIVE COUNTS, FDR = FALSE DISCOVERY RATE, #LES = NUMBER OF LESIONS

	overall	3-5	6-10	11-20	21-50	51-100	101+
#les	235	65	45	38	52	20	15
TP	223	58	42	37	51	20	15
SENS	0.95	0.89	0.93	0.97	0.98	1	1
FP	55	33	14	5	3	0	0
FDR	0.20	0.36	0.25	0.12	0.05	0	0

3) Evaluating the Effect of Lesion Size: Large gad lesions that span over multiple slices are much easier to capture than the ones with only a few voxels in a single slice. In this section we examine the performance of the THAT-CRF algorithm as a function of lesion size. Table III shows sensitivity and false discovery rates for different lesion sizes from very small (3–5 voxels) to very large (101+ voxels) when operating at overall target sensitivities of 0.95. Sensitivities for different lesion sizes range from 89% for very small lesions to 100% for very large lesions.

4) Evaluating the Effect of Timepoint Interval: So far, we limited our studies to pairs of scans acquired six months apart. While the scan interval in *phase I* of clinical trials is typically six months or more, in *phase II*, more frequent scans are common. Therefore, it is of great interest to measure the sensitivity of our model to the acquisition interval between the two scans.

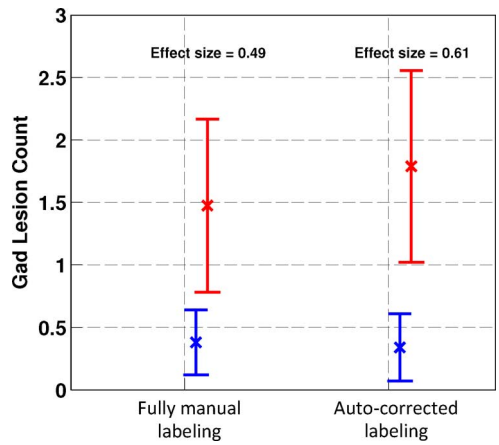

Fig. 12. Separation of two treatment arms [patients on the drug (blue) and those on placebos (red)] based on gad lesion counts for fully manual labeling and automatic labeling corrected only for false detections (auto-corrected labeling). Effect size quantifies the difference between two groups and is evaluated as: (mean of group1 – mean of group 2)/standard deviation. It is observed that auto corrected labeling leads to a better separation of the two groups.

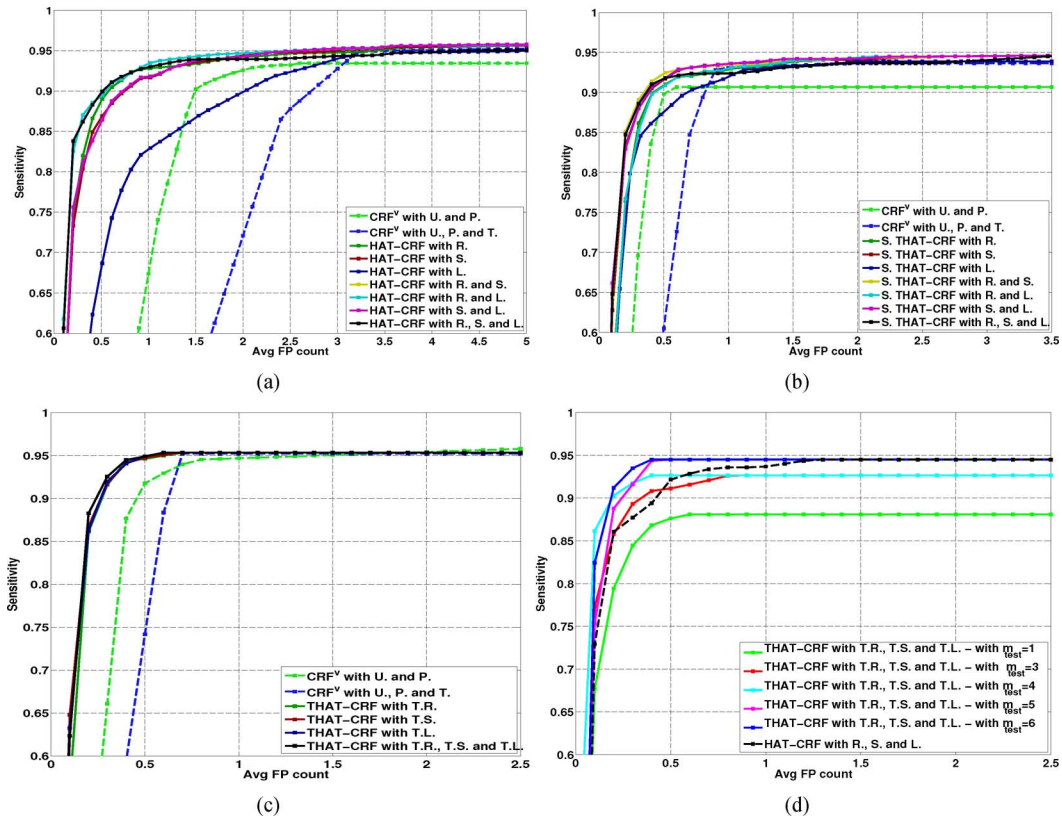

Fig. 13. Plots of sensitivity vs average false positive count. (a)–(c) Single timepoint (HAT-CRF), semi-temporal (S. THAT-CRF), and the proposed temporal classifier (THAT-CRF). (d) Corresponds to applying the temporal model learned for $m_{\text{train}} = 6$ on pairs of scans with different m_{test} intervals.

In other words, we would like to explore the performance of our detection framework at different ranges of scanning intervals, when the model is learned with a fixed acquisition interval (currently set at six months). This will help to determine the intervals for which retraining is required. To this end, we test the model learned for six month intervals on pairs of scans of the same patients with variable scanning intervals attained from dataset $B2$. Specifically, we classify the gad lesions at the baseline where the second timepoint being used is selected at one, three, four, five, and six months intervals. The plots of the sensitivity as a function of the false discovery rate are shown in Fig. 11(d). As expected, the best performance corresponds to $m_{\text{test}} = 6$. However, for $m_{\text{test}} >= 4$ months the THAT-CRF model still works well and in fact outperforms the single timepoint HAT-CRF model.

5) Sensitivity Versus Average False Positive Count: It is informative to plot the sensitivity versus the average false positive count per patient. Fig. 13 shows these curves for all the plots shown in Fig. 11. It is observed that the proposed temporal model yields less than 0.5 false positive count on average per patient at the sensitivity of 95% (the highest achievable sensetivity among all models). This indicates that further review of the automatic labelings by a human expert, if required, can be accomplished very quickly.

6) Evaluating the Discriminatory Power of the Automatic Labeling in Differentiating Treatment Arms: One of the important goals of clinical trials is to assess the efficacy of a new drug. This is commonly performed in phase II trials for MS by measuring changes in the gad lesion count in response to therapy for groups of subjects on the drug under investigation and a comparator treatment, either placebo or active. Accordingly, we performed an experiment to assess the effectiveness of the automatic labeling in separating these two treatment arms in a clinical trial. Following clinical convention [38], we studied the MRI of 69 patients at months 3, 4, 5, and 6. Only scans at month 6 were analyzed by the THAT-CRF using the baseline scan as a reference. For the other timepoints the HAT-CRF model was used. Fully manual labeling of the scans performed by trained experts using the protocol described in Section III-C were available. The results of the automatic labeling of all scans were provided to the trained experts asking them to correct only false positives. We call this new set the *auto-corrected* labels. The mean gad lesion count for both the placebo and treated groups was computed using the fully manual labels and again with the auto-corrected labels. Fig. 12 summarizes the results. The effect size measure is used to quantify the separation obtained by each method. Higher effect sizes indicate better separations. It appears that the two treatment arms are more effectively separated by using the auto-corrected labels than the fully manual ones. This is because the automatic method has captured some lesions that were missed in the fully manual labeling (two examples of such lesions are shown in the last two rows of Fig. 7)[4]. It should be noted that manually finding missing lesions (false negatives) is very challenging and much more time-consuming

[4]The fully automatic labeling without manual correction yields the effect size of 0.33.

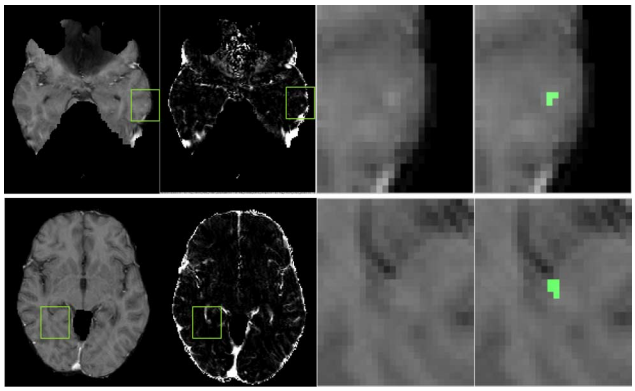

Fig. 14. Examples of two false negatives. Each row shows an example image from a different patient. First column shows the post-contrast T1w images. Second column shows the enhancement map. Enhanced regions corresponding to the missed lesions are marked with green rectangles. Last two columns show the zoomed view of the false negatives without and with manual labeling.

than correcting false detections, and thus not feasible in practice. Due to the high sensitivity of the proposed technique, we only investigated the effect of a human rater removing false detections (which are very few) in this experiment. Note that the low false negative rate of the automatic method does not affect the discriminatory power of the algorithm in separating the treatment arms.

We also measured the total processing time that is saved by applying the automatic technique. The required time by the readers to perform fully manual labelling of gad-enhancing lesions for each MRI volume ranges from 10 minutes to several hours depending on the number of lesions in the volume. However, it takes only 53.75 and 58.10 s on average (on a 2.66 GHz CPU) for the HAT-CRF and THAT-CRF, respectively. Moreover, our analysis showed that correcting the results of the automatic labeling takes on average less than half of the time required for the fully manual labeling. It should be noted that this was the first time this type of experiment was conducted. It is expected that by repeating this experiment a few more times, readers would gain expertise in correcting the automatic labeling, which would result in further reductions in the total required time.

7) Some Examples of the False Negatives: As it is observed from Fig. 11(c), all curves saturate at sensitivity of 95%. This is because there are a few lesions that are hardly detectable and hence are missed at the voxel-level analysis. Two examples of these false negatives are shown in Fig. 14. These false negatives are mainly caused due to the lack of accurate priors for these locations. The example in the first row depicts a lesion very close to the gray matter. The second example shows a lesion at the base of the ventricular horn. In both cases the white matter prior yields low values resulting in the poor performance of the classification algorithm.

IV. Conclusion

In this paper, we proposed a THAT-CRF classifier to detect and segment Multiple Sclerosis gad-enhancing lesions in brain MRI. The small size (mostly 3–5 voxels) and various shapes of these lesions make the problem more challenging than the typical segmentation problems. We showed how to integrate higher order descriptors and spatio-temporal features at different scales in order to discriminate between the true enhanced lesions and the pool of all other nonlesional enhancements. The model is learned on a very large multi-center clinical trial consisting of 2380 scans and the effect of different components of the model is extensively evaluated. Results show that the proposed model achieves a very high sensitivity (95%) along with a low false discovery rate (20%) and very few average false positive counts (0.5) per patient. The sensitivity of the temporal method to the chosen time interval is further investigated on five different intervals of 69 patients. Finally, when applied to the context of separating different treatment arms (the final goal in many of the clinical trials), superior performance is achieved by the reviewed labelings of our model (corrected only for the false detections) compared to the fully manual labeling. In conclusion, the high sensitivity along with the few false positive counts of our THAT-CRF model offer a fast and accurate solution to be employed in real clinical trials where final review of the automatic results is routinely performed.

References

[1] M. P. Sormani *et al.*, "Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: A meta-analytic approach," *Ann. Neurol.*, vol. 65, no. 3, pp. 268–275, 2009.

[2] Z. Karimaghaloo, M. Shah, S. J. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Detection of gad-enhancing lesions in multiple sclerosis using conditional random fields," in *MICCAI*, 2010, pp. 41–48.

[3] Z. Karimaghaloo *et al.*, "Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1181–1194, Jun. 2012.

[4] F. Cotton, H. L. Weiner, F. A. Jolesz, and C. R. Guttmann, "MRI contrast uptake in new lesions in relapsing-remitting MS followed at weekly intervals," *Neurology*, vol. 60, no. 4, pp. 640–646, 2003.

[5] D. S. Meier, H. L. Weiner, and C. R. Guttmann, "Time-series modeling of multiple sclerosis disease activity: A promising window on disease progression and repair potential?," *Neurotherapeutics*, vol. 4, no. 3, pp. 485–498, 2007.

[6] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr, "Associative hierarchical CRFs for object class image segmentation," in *Proc. 12th Int. Conf. IEEE Comput. Vis.*, 2009, pp. 739–746.

[7] P. S. L'ubor Ladicky, K. Alahari, C. Russell, and P. H. Torr, "What, where and how many? Combining object detectors and CRFs," in *Proc. 11th Eur. Conf. Comput Vis.*, 2010, pp. 424–437.

[8] S. Kumar and M. Hebert, "Discriminative random fields," *Int. J. Comput. Vis.*, vol. 68, no. 2, pp. 179–201, 2006.

[9] X. Boix *et al.*, "Harmony potentials," *Int. J. Comput. Vis.*, vol. 96, no. 1, pp. 83–102, 2012.

[10] M. W. Schmidt, K. P. Murphy, G. Fung, and R. Rosales, "Structure learning in random fields for heart motion abnormality detection," in *Proc. CVPR*, 2008, vol. 1, p. 1.

[11] C. H. Lee *et al.*, "Segmenting brain tumors with conditional random fields and support vector machines," in *Computer Vision for Biomedical Image Applications*. New York: Springer, 2005, pp. 469–478.

[12] Z. Hao *et al.*, "Combining CRF and multi-hypothesis detection for accurate lesion segmentation in breast sonograms," in *MICCAI*, 2012, pp. 504–511.

[13] S. Bauer, L. P. Nolte, and M. Reyes, "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in *MICCAI*, 2011, pp. 354–361.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1265–1278, Aug. 2005.

[15] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.* , vol. 24, no. 7, pp. 971–987, Jul. 2002.

[16] P. Coupé *et al.*, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, no. 2, pp. 940–954, 2011.
[17] Y. Miki *et al.*, "Computer-assisted quantitation of enhancing lesions in multiple sclerosis: Correlation with clinical classification," *Am. J. Neuroradiol.*, vol. 18, no. 4, pp. 705–710, 1997.
[18] B. J. Bedell and P. A. Narayana, "Automatic segmentation of gadolinium-enhanced multiple sclerosis lesions," *Magn. Reson. Med.*, vol. 39, no. 6, pp. 935–940, 1998.
[19] R. He and P. A. Narayana, "Automatic delineation of GD enhancements on magnetic resonance images in multiple sclerosis," *Med. Phys.*, vol. 29, p. 1536, 2002.
[20] S. Datta *et al.*, "Segmentation of gadolinium-enhanced lesions on MRI in multiple sclerosis," *J. Magn. Reson. Imag.*, vol. 25, no. 5, pp. 932–937, 2007.
[21] Z. Karimaghaloo, D. L. Arnold, D. L. Collins, and T. Arbel, "Hierarchical conditional random fields for detection of gad-enhancing lesions in multiple sclerosis," in *MICCAI*, 2012, pp. 379–386.
[22] Z. Karimaghaloo, H. Rivaz, D. L. Arnold, D. L. Collins, and T. Arbel, "Adaptive voxel, texture and temporal conditional random fields for detection of gad-enhancing multiple sclerosis lesions in brain MRI," in *MICCAI*, 2013, pp. 543–550.
[23] D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans, "Automatic 3-d model-based neuroanatomical segmentation," *Human Brain Mapp.*, vol. 3, no. 3, pp. 190–208, 1995.
[24] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[25] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, 2007.
[26] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.
[27] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
[28] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapp.*, vol. 17, no. 3, pp. 143–155, 2002.
[29] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, 1998.
[30] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 143–150, Feb. 2000.
[31] V. Fonov *et al.*, "Unbiased average age-appropriate atlases for pediatric studies," *NeuroImage*, vol. 54, no. 1, pp. 313–327, 2011.
[32] C. Elliott, D. Collins, D. Arnold, and T. Arbel, "Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 1490–11503, Aug. 2013.
[33] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 840–853, May 2007.
[34] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Trans. Med. Imag.*, vol. 20, no. 8, pp. 677–688, Aug. 2001.
[35] A. Akselrod-Ballin *et al.*, "An integrated segmentation and classification approach applied to multiple sclerosis analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis.Pattern Recognit.*, 2006, vol. 1, pp. 1122–1129.
[36] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, "Probabilistic segmentation of white matter lesions in MR imaging," *NeuroImage*, vol. 21, no. 3, pp. 1037–1044, 2004.
[37] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Med. Image Anal.*, vol. 17, no. 1, pp. 1–18, 2013.
[38] S. L. Hauser *et al.*, "B-cell depletion with rituximab in relapsing-remitting multiple sclerosis," *N. Eng. J. Med.*, vol. 358, no. 7, pp. 676–688, 2008.