34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

# Most Edges in Markov Random Fields for White Matter Hyperintensity Segmentation are Worthless

Christopher G. Schwarz[1], Evan Fletcher[2], Baljeet Singh[2],
Amy Liu[2], Noel Smith[2], Charles DeCarli[2], and Owen Carmichael[2]

*Abstract*— **The time and space complexities of Markov random field (MRF) algorithms for image segmentation increase with the number of edges that represent statistical dependencies between adjacent pixels. This has made MRFs too computationally complex for cutting-edge applications such as joint segmentation of longitudinal sequences of many high-resolution magnetic resonance images (MRIs). Here, we show that simply removing edges from full MRFs can reduce the computational complexity of MRF parameter estimation and inference with no notable decrease in segmentation performance. In particular, we show that for segmentation of white matter hyperintensities in 88 brain MRI scans of elderly individuals, as many as 66% of MRF edges can be removed without substantially degrading segmentation accuracy. We then show that removing edges from MRFs makes MRF parameter estimation and inference computationally tractable enough to enable modeling statistical dependencies within and across a larger number of brain MRI scans in a longitudinal series; this improves segmentation performance compared to separate segmentations of each individual scan in the series.**

## I. INTRODUCTION

Markov Random Fields (MRFs) provide a probabilistic graphical model framework for solving image processing tasks such as denoising, inpainting, and segmentation. By modeling each pixel as a node in a graph, and dependencies between neighboring pixels as edges between nodes, MRFs are able to represent complex statistical relationships between image pixels in a mathematically principled way. Numerous approaches have been presented for the two key computational problems that must be solved to use MRFs in practice: *parameter estimation*, using labeled ground-truth images to estimate the parameters of probabilistic models of inter-pixel dependencies; and *inference*, assigning labels to the pixels that are in accord with the estimated inter-pixel dependencies [1].

Unfortunately, the time and space complexities of all current approaches to parameter estimation and inference increase at least linearly with the number of edges that are included to account for statistical dependencies between adjacent pixels (Table 1). Typically, if each pixel corresponds to a node in the graph, a lattice of edges is induced that connects each node to its $k$ nearest neighbors. Thus, for high resolution images with a large number of pixels, the resulting MRF can include so many nodes and edges that parameter estimation and inference become intractable. This is becoming an especially important problem in neuroimaging, where state-of-the-art studies are collecting longitudinal series of 5 to 10 volumetric MRI scans of the same individual over time [2]. Each scan may contain a 3D array of 256 x 256 x 256 pixels, and besides edges that model dependencies between adjacent pixels within an individual scan, it is desirable for the MRF to encourage biologically-plausible dynamics in segmentation labels over time by including MRF edges that connect a node from one scan to a neighborhood of corresponding nodes in a scan that is adjacent in time. Due to their large number of edges, performing parameter estimation and inference in these large graphs is beyond the scope of even the most state-of-the-art MRF algorithms unless substantial heuristic approximations are employed.

In this paper we argue for making existing MRF algorithms tractable for such large-scale applications by removing edges from the graphs on which they operate. We focus on applications such as brain MRI segmentation for which all images to be segmented are warped to a template space such that each node and edge corresponds to an analogous anatomical location across scans. Before parameter estimation, we use training data to determine which edges to remove from the graph. We then run parameter estimation and inference on the resulting, reduced graph. We test the accuracy of the reduced graphs for MRI-based segmentation of white matter hyperintensities (WMHs), a brain imaging finding important to Alzheimer's disease, multiple sclerosis, depression, and other brain disorders. These experiments suggest that a majority of the edges in an MRF can safely be removed without compromising WMH segmentation performance. Finally, we show that such reduced graphs give rise to increased WMH segmentation accuracy by enabling unified segmentation of large graphs that represent a longitudinal series of three or more high-resolution MRIs along with spatial and temporal label dependencies.

## II. RELATED WORK

Several prior algorithms simplify the structure of graphical models. Hierarchical models connect neighborhoods of pixels not to each other but to a "supernode" that approximates the entire neighborhood. Neighborhoods of supernodes are connected to supernodes at a higher level, and so on. This technique forms tree-structured graphs which allow efficient parameter estimation and inference, but their outputs are often "blocky" due to the fact that dependencies between

This work was supported by NIH grants NIH grants AG10220, AG10129, AG 030514, AG031252, AG021028, and AG024904

[1]C.G. Schwarz is with the Computer Science Department, University of California, Davis, CA 95616 cgschwarz@ucdavis.edu

[2]E. Fletcher, B. Singh, A. Liu, N. Smith, O. Carmichael and C. DeCarli are with the Neurology Department, University of California, Davis, CA 95616

TABLE I

COMPLEXITY OF MRF ALGORITHMS

| Method | Time Complexity | Space Complexity |
|---|---|---|
| Direct Inference | $O(\|L\|^{\|V\|}\|E\|)$ | $O(\|L\|^{\|V\|})$ |
| Simulated Annealing | $O(n\|E\|)$ | $O(\|V\|)$ |
| ICM | $O(n\|E\|\|L\|)$ | $O(\|V\|\|L\|)$ |
| Belief Propagation | $O(n\|E\|\|L\|)$ | $O(\|E\|\|L\|)$ |
| Junction Tree Algorithm | $O(\|L\|^k)$ | $O(\|L\|^k)$ |
| Graph Cuts | $O(\|E\|\|V\|^2)$ | $O(\|V\|)$ |
| IPF | $O(n_{\text{IPF}} n_{\text{BP}} \|E\|\|V\|)$ | $O(\|E\|)$ |
| Pseudolikelihood | $O(n\|T\|\|E\|)$ | $O(\|T\|\|E\|)$ |

Time and space complexity of major algorithms for MRF inference and parameter estimation, assuming a naive implementation for graphs with $|V|$ nodes and $|E|$ edges, each of which correspond to a compatibility function with a single unique free parameter. Other terms are: $n$: the number of iterations, $|T|$: the size of the training set, and $|L|$: the number of possible pixel labels. Note that $|E|$ occurs in the time complexities of all but the Junction Tree Algorithm, for which the tree width $k$ depends indirectly upon $|E|$ as well. The $|E|$ term also appears in many of the space complexities.

neighboring pixels are represented by variable-length paths through the tree [3], [4].

MRF parameter estimation can be performed with an $L_1$ regularizer that encourages zero-valued parameters that exert no influence on inter-pixel label dependencies and thus represent removable edges [5]. However, these methods require solving an expensive regularized parameter estimation problem on a full graph to determine which edges to remove; our starting point is imaging data so large that solving such a parameter estimation problem is computationally intractable.

Another set of methods iteratively removes edges from decomposable models such as Bayesian networks, for which an edge can be removed without modifying any other model parameters [6], [7]. Lattice-structured graphs that are natural for modeling imaging data are generally not decomposable. It is possible to convert lattice models into junction trees, which are decomposable and have several reduction methods designed for them [8], but converting a lattice-structured graph to a junction tree requires adding triangulating edges between all square-shaped configurations, rendering this approach intractable. Our approach is to apply iterative edge removal techniques to large-scale lattice-structured MRFs for which edge removal has not been investigated in any depth.

## III. METHODS

We begin with 3D lattice MRFs with a heterogeneous Potts compatibility function at each edge, Pseudolikelihood maximization as a parameter estimation objective function, and a simplex-based optimizer.

### A. MRFs for image segmentation in a template space

MRFs model the joint probability of fields of random variables. In image processing, each image pixel $i$ typically has a corresponding node $v_i \in V$, and the label $f_i \in L$ assigned to $v_i$ is one such random variable. Each neighboring pair of nodes is connected with an edge $e \in E$ representing a statistical dependency between adjacent pixel labels. For each edge $e$ a *compatibility function* $\Psi$ assigns a probability to each possible assignment of labels to the nodes it connects. For each node $v_i$ the *observation function* $\Phi$ assigns a probability to all labels in $L$ given the image intensities at that location, $o_i$. The MRF models the field of pixel labels with the energy function:

$$U(f) = \sum_{i \in V} \Phi(i,f) + \sum_{e \in E} \Psi(e,f) \quad (1)$$

Here, $f$ is an assignment of labels from $L$ to $V$, and $F$ is the space of all possible labelings. Inference is the process of finding an $f \in F$ that minimizes this energy, and parameter estimation is the process of determining a set of parameters governing $\Phi$ and $\Psi$ that conform to $o_i$ and $f$ provided by labeled training data [1].

Many brain image segmentation tasks, including our WMH segmentation application, are performed in a *template space*. All images are nonlinearly warped to a common *template image* as described previously [9] so that each pixel corresponds to the same anatomical location across subjects. This approach allows us to provide a detailed model of label dependencies that vary from location to location to reflect the spatially-variable properties of distinct anatomical regions. In particular, our compatibility function $\Psi(e,f)$ is a spatially heterogeneous Potts model: each $e \in E$ is assigned its own free parameter $\Theta_e$ representing the amount of energy added to $\Psi$ in the event that the nodes connected by $e$ take on differing labels.

Our observation function is $\Phi(i,f) = \sum_i [O(o_i,f_i) \times \text{Fr}(i,f_i)]$ in which $\text{Fr}(i,f_i)$ is the *label prior:* the frequency of label $f_i$ occurring at location $i$ in the training data. $O(o_i,f_i)$ gives the probability of label $f_i$ being associated with image intensity $o_i$ at pixel $i$. We model $O(o_i,f_i)$ using one log-normal distribution per tissue label, as described previously [9]. These observation and compatibility functions are used throughout all experiments with full and reduced MRFs described in Section IV. For parameter estimation, we used a simplex-based optimizer to maximize graph pseudolikelihood [10], and for inference we used Belief Propagation [11].

### B. Prior-driven edge removal

Under the above formulation, graph reduction is the process of removing as many $e \in E$ from the graph as possible while maintaining the strongest possible connection between minimizing $U(f)$ and maximizing the accuracy of the resulting pixel labels. Our approach is based on the intuitive principle that pixel neighborhoods with little inter-subject or inter-pixel label variability require little or no modeling of inter-pixel label dependencies by $\Psi$. For example, a pixel $i$ in an anatomical region where WMHs rarely occur will usually be assigned a non-WMH label based on the observation function $\Phi$ alone, and the labels of surrounding pixels are so often also non-WMH along with it that they provide little additional useful information about $f_i$. Thus, edges between pixel $i$ and its neighbors can be removed with little impact.

We use label frequency information to quantify the importance of edges in the graph in an approach we call *prior-driven* edge removal. We assign each edge $e \in E$ an *edge prior* $p_e$. If $e$ connects nodes $i$ and $j$, we define $p_e$ as follows:

$$p_e = \min(\max_{l \in L} \text{Fr}(i,l), \max_{l \in L} \text{Fr}(j,l)) \quad (2)$$

in which $\mathrm{Fr}(i,l)$ denotes the frequency at which label $l$ occurs at location $i$, in the labeled training data. The value of $p_e$ becomes greater as the label distribution at both of the nodes becomes more concentrated about a single label; thus, edges with a high $p_e$ are likely to be relatively less relevant because the labels of the pixels that they connect are largely determined by their frequency irrespective of imaging data or neighboring labels.

We remove edges based on $p_e$ using two approaches. In *backward selection*, we begin from a full graph and iteratively remove a designated number of randomly-selected edges from among those with a high $p_e$. In the *forward selection* approach, we begin from a graph with no edges and iteratively add a designated number randomly selected from among those with low $p_e$. We choose edges randomly based on their $p_e$ value, rather than adding or removing individually based on a sorting by $p_e$, because nearby edges often have highly similar $p_e$ values. The random element thus decreases the spatial locality of inserted or removed edges, and thus encourages more global changes to the graph in a smaller number of insertions or removals.

### C. Alternative edge removal criteria

We compared prior-driven edge removal against theoretically-driven, computationally-expensive criteria that evaluated the impact of edge insertions and removals on MRF parameter estimation and inference diagnostics. First, we considered a forward selection approach in which the next edge to be added is the one that provided the greatest increase to the training data pseudolikelihood [10], which can be thought of as approximating an exponential of the $U$ function evaluated over all training examples. We then considered a backward selection approach designed to first train the full graph and remove edges that minimally modified the behavior of that graph in terms of the distribution of $U(f)$ values over all possible label sets $F$. We used the Kullback-Leibler (K-L) divergence [12] to quantify differences between full and reduced graphs in this sense. We evaluated the viability of these more expensive approaches on small graphs in Sec. IV-B to show that they lack substantial advantages over our prior-driven method.

### D. Retraining Approaches

Because lattice-structured MRFs are not decomposable, removal or addition of a single edge could theoretically change the optimal values for compatibility function parameters throughout the graph. This means that retraining is theoretically required: a new run of parameter estimation after every such graph modification. However, in real-world graphs, removing or adding an edge in one corner of a large graph is expected to have little effect on compatibility parameters of distant edges, especially because parameter estimation algorithms applied to large-scale graphs effectively only optimize parameter values with respect to an extended local neighborhood. Therefore, in Section IV-B we experimented with omitting re-training for small graphs,

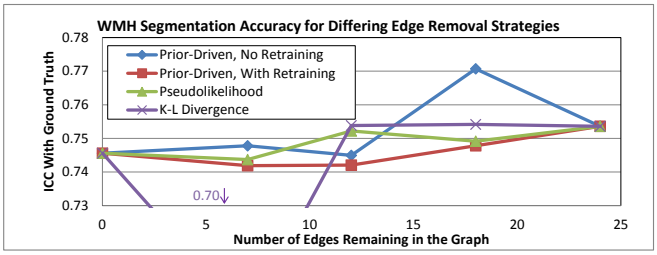


Fig. 1. Results of WMH Segmentation on a small $4 \times 4$ subgraph of MR Images at varying levels of graph reduction with each technique. Note that our proposed prior-driven reduction method, without retraining, performs comparably and often better than the more costly methods that are intractable for larger graphs.

and showed that doing so does not substantially alter the inference performance of the reduced graphs.

## IV. EXPERIMENTS

### A. Data

We evaluated the utility of graph reduction on 958 fluid-attenuated inversion recovery (FLAIR) MRI scans of elderly individuals aged 70-90 in the University of California, Davis Alzheimer's Disease Center (ADC) Longitudinal Cohort covering a range from normal cognition to dementia. Subject recruitment, image acquisition, ground-truth semi-manual WMH segmentation, and warping of these images to a common anatomical template has been described previously [13].

### B. Segmentation of small sub-images

In these experiments we compare our prior-driven edge removal and edge removal based on the more costly pseudolikelihood and K-L divergence criteria (Sec. III-C). We selected the same $4 \times 4$ pixel sub-image from each of the images described in Sec. IV-C and ran WMH segmentation on 88 of them, using the other 870 sub-images as training data. For each method, and for each reduced graph resulting from iterative edge removal or addition, we performed parameter estimation on the 870 training images and inference on the 88 remaining ones, and calculated the intraclass correlation coefficient (ICC) between the volume of pixels the automated method labeled as WMH, and the volume of WMH-labeled pixels provided by ground-truth semi-manual FLAIR segmentation. Higher ICC values denote stronger agreement between estimates and ground truth (Fig. 1). The expensive pseudolikelihood and K-L divergence methods did not perform substantially better than the prior-driven reduction method in graphs of any size. In addition, re-training as described in Sec. III-D did not lead to substantially higher performance either. We concluded that prior driven edge removal without retraining performs comparably to more costly and theoretically more accurate methods and therefore applied this method to the full images.

### C. Segmentation of full images

We performed the forward and backward variants of our prior-driven edge removal and calculated the ICC between ground-truth WMH volumes and those estimated automatically using the reduced graphs (Fig. 2). The ICC drops

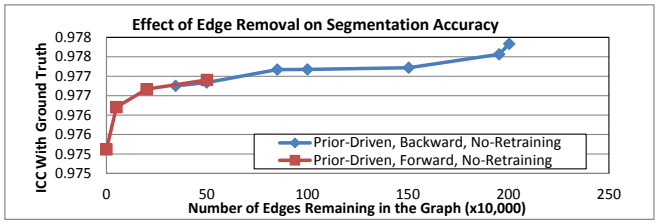


Fig. 2. Results of applying various levels of our proposed prior-driven graph reduction in both the backward-selection and forward-selection variants, without retraining, to a WMH segmentation task. Note that as many as 58% of the edges can be removed without substantially damaging segmentation performance, and that the two directional variants perform comparably for similar levels of reduction.

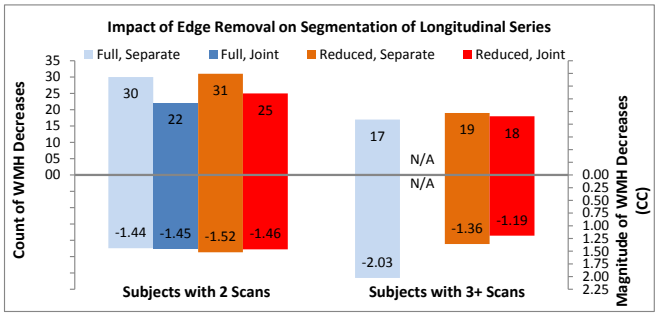


Fig. 3. Occurrences (*above axis*) and magnitudes (*below axis*) of significant decreases in segmented WMH volume, a biologically implausible event, between sequential pairs of segmented scans. Joint inference with full graphs is intractable for 3+ scans at a time, and so was omitted. Note that joint segmentation reduces these occurrences and their magnitudes versus the traditional separate method, and that segmentation performance of reduced graphs is only slightly reduced compared to full graphs.

slightly on removal of the first few edges, but then remains remarkably stable until as many as 58% of edges have been removed. This suggests that more than half of MRF edges can safely be removed without substantially damaging WMH segmentation performance. We also note that the performance of forward- and backward-selection methods are convergent around this point, suggesting that there may be no strong reason to prefer one or the other approach.

### *D. Longitudinal Segmentation of multiple MR Images*

To determine whether graph reduction enables new approaches to segmentation of longitudinal MRI series, we performed graph reduction using the backward-selection, no-retraining variant of our proposed method and used a graph with about 58% of edges removed for longitudinal segmentation. To jointly segment a series of $k$ images for each subject, we created $k$ replicas of the reduced graph and introduced new edges that connected corresponding nodes across adjacent time points. First, we performed joint segmentation with full and reduced graphs on the scans of 179 subjects with exactly two scans. Next, we performed joint segmentation with the reduced graph on each of the 40 subjects with scans at three or more time points. Joint inference on a full graph is intractable for these longer series, and so was not performed. For comparison to a more traditional approach, we also segmented each scan in a series separately from the rest of the series.

To analyze these results, we examined occurrences of an implausible result: WMH volumes decreasing over time. For each subject, we calculated change in segmented WMH volume between subsequent pairs of time points. We then counted those pairs with significant decreases in WMH volume ($> 0.43$ CC), and calculated the average magnitude of these decreases. We present these results in Fig. 3.

In these experiments, joint segmentation led to fewer occurrences of such implausible results versus corresponding separate approaches. Joint segmentation also reduced the magnitudes of these decreases, when they did occur. Consistent with Sec. IV-C, segmentations using reduced graphs were only slightly inferior to their non-reduced counterparts, when available. The lowest average change magnitude overall was achieved by joint segmentation of the multi-time point dataset, which could not practically be performed without the reduced graph.

## V. DISCUSSION

In this work we proposed reducing the computational cost of MRFs for image segmentation by removing edges from the graphs. We showed that for a WMH segmentation task, removing the majority of edges leads to a negligible drop in segmentation accuracy, and we showed that removing edges in this way makes joint segmentation of longer MRI series possible. Such joint segmentation of longitudinal series led to greater biological plausibility in WMH change.

Future work should investigate why more costly, theoretically-driven edge removal metrics showed no substantial benefit over randomly driven edge removal. Additional work will also focus on extending edge removal to other MRF tasks such as 3-tissue brain segmentation and MRI smoothing.

## REFERENCES


[1] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, 3rd ed. London: Springer-Verlag, 2009.

[2] S. G. Mueller *et al.*, "The Alzheimer's Disease Neuroimaging Initiative," *Neuroimaging Clinics of North America*, vol. 15, no. 4, pp. 869–877, 2005.

[3] S. Yu *et al.*, "A hierarchical Markov Random Field model for figure-ground segregation," *Third International Workshop on Energy Minimization Methods in IEEE CVPR*, 2001.

[4] J. K. Johnson and A. S. Willsky, "A Recursive Model-Reduction Method for Approximate Inference in Gaussian Markov Random Fields," *IEEE TSP*, vol. 17, no. 1, pp. 70–83, Jan. 2008.

[5] M. W. Schmidt *et al.*, "Structure learning in random fields for heart motion abnormality detection," *IEEE CVPR*, pp. 1–8, 2008.

[6] U. Kjaerulff, "Reduction of computational complexity in Bayesian networks through removal of weak dependences," *Proceedings of Uncertainty in AI*, pp. 374–382, 1994.

[7] R. A. van Englen, "Approximating Bayesian Belief Networks by Arc Removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 8, pp. 916–920, 1997.

[8] D. Shahaf *et al.*, "Learning Thin Junction Trees via Graph Cuts," *Artificial Intelligence and Statistics (AISTATS)*, vol. 5, 2009.

[9] C. Schwarz *et al.*, "Fully-automated White Matter Hyperintensity detection with anatomical prior knowledge and without FLAIR," *Inf. Proc. in Medical Imaging*, vol. 21, pp. 239–251, Jan. 2009.

[10] J. Besag, "Statistical Analysis of Non-Lattice Data," *The Statistician*, vol. 24, no. 3, p. 179, Sep. 1975.

[11] J. Pearl, "Reverend Bayes on inference engines: A distributed hierarchical approach," in *Proceedings of the AAAI National Conference on AI*, 1982, pp. 133–136.

[12] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[13] O. Carmichael *et al.*, "MRI predictors of cognitive change in a diverse and carefully characterized elderly population," *Neurobiology of Aging*, vol. 33, no. 1, pp. 83 – 95.e2, 2012.