33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Impact of Markov Random Field Optimizer on MRI-based Tissue Segmentation in the Aging Brain

Christopher G. Schwarz*, Alex Tsui*, Evan Fletcher†, Baljeet Singh†, Charles DeCarli †, and Owen Carmichael*†
*Computer Science Department, University of California, Davis, CA 95616. Email: cgschwarz@ucdavis.edu
†Neurology Department, University of California, Davis, CA 95616

***Abstract*—Automatically segmenting brain magnetic resonance images into grey matter, white matter, and cerebrospinal fluid compartments is a fundamentally important neuroimaging problem whose difficulty is heightened in the presence of aging and neurodegenerative disease. Current methods overlap greatly in terms of identifiable algorithmic components, and the impact of specific components on performance is generally unclear in important real-world scenarios involving serial scanning, multiple scanners, and neurodegenerative disease. Therefore we evaluated the impact that one such component, the Markov Random Field (MRF) optimizer that encourages spatially-smooth tissue labelings, has on brain tissue segmentation performance. Two challenging elderly data sets were used to test segmentation consistency across scanners and biological plausibility of tissue change estimates; and a simulated young brain data set was used to test accuracy against ground truth. Belief propagation (BP) and graph cuts (GC), used as the MRF optimizer component of a standardized segmentation system, provide high segmentation performance on aggregate that is competitive with end-to-end systems provided by SPM and FSL (FAST) as well as the more traditional MRF optimizer iterated conditional modes (ICM). However, the relative performance of each method varied strongly by performance criterion and differed between young and old brains. The findings emphasize the unique difficulties involved in segmenting the aging brain, and suggest that optimal algorithm components may depend in part on performance criteria.**

I. Introduction

Fully-automated methods for classifying each pixel in brain Magnetic Resonance Imagery (MRI) into one of three tissue compartments– grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF)– are playing increasingly important roles in characterizing brain changes that accompany brain development, aging, and neurodegenerative diseases. For this reason, a large array of brain tissue segmentation methods have been proposed and validated on real-world data sets. Unfortunately, such methods are generally presented as monolithic end-to-end solutions rather than collections of computational modules, despite the fact that most algorithms have in common a few easily-identifiable components. Prominent components include statistical models that relate image intensities to segmentation labels, formalisms for encouraging certain spatial configurations of tissue labels, models of partial volume effects, and numerical routines that solve for optimal labelings. Although an individual component may be implemented similarly across several state-of-the-art methods, the impact of that component on segmentation performance is generally not explored. Such component-level performance evaluation would enable the construction of high-performing segmentation systems from high-performing components, especially now that a growing number of image processing software platforms (Insight Toolkit, LONI Pipeline, *e*tc.) are based on plug-and-play assembly of overall systems.

Additionally, the rise to prominence of large-scale neuroimaging studies [1] has heightened the need for algorithms that provide biologically-plausible tissue segmentations of large groups of healthy, aging, and diseased brains scanned on multiple scanners at multiple points in time. In aging, brain tissue measurements collected on the same or differing scanners in rapid succession (over the course of days or weeks) should be highly consistent; GM and WM volumes should either remain stable or decline over longer time course (months or years); and the white matter hyperintensities (WMHs) commonly associated with aging should be properly accounted for. To date, detailed validation of brain segmentation methods in this setting has been lacking.

The purpose of this paper is to take a first step toward component-level performance evaluation of brain segmentation methods in multi-site longitudinal studies of aging. We focus on the optimization component of Markov Random Fields (MRFs), comparing three types: Belief Propagation (BP), Graph Cuts (GC), and Iterated Conditional Modes (ICM). Each of these optimizers, which encourage plausible labelings of pixels, was incorporated into an established segmenter [2] to assess the impact of this component on performance. We assessed performance on two challenging elderly scan sets in terms of tissue volume agreement on rapidly-repeated serial scans on differing MRI scanners, and in terms of biological plausibility of tissue changes over time. WMHs were detected on all scans independently and omitted from GM and WM volumes. To highlight the unique challenges presented by the aging brain, we also applied the MRF methods to simulated scans of a young, healthy brain [3], and compared performance against a popular existing package ([4]) that uses this same young brain to initialize its segmentations. We then indexed performance against that of FAST [5], a widely-used end-to-

We thank Karthika Ramanathan for her assistance. This work was supported by NIH grants AG10220, AG10129, AG030514, AG031252, AG021028, and AG024904, The Dana Foundation, and the California Department of Public Health Alzheimer's Disease Program Contracts 06-55311 and 06-55312.

end system.

II. Related Work

Related work largely falls into one of two categories: performance comparisons among multiple off-the-shelf, end-to-end brain segmentation systems; and assessment of individual components of pixel labeling systems, especially the MRF optimizer, for other application domains. For the former, a wealth of papers have focused on FAST and SPM. In two studies that used simulated and real data to assess tissue segmentation accuracy within and between segmenters, SPM was found to be more accurate than FAST. [6], [7]. Excellent agreement between the two methods [8], and similar test-retest reliability between the two [9], has been reported, although one report suggests that FAST may be superior for measuring longitudinal brain volume changes [10]. Meanwhile, performance comparisons among MRF optimizers have been mixed, with two studies suggesting that GC and BP perform comparably on stereo disparity map estimation [11], [12], but another suggesting that among GC, BP, ICM, and other competing methods, GC performs better on photographic image stitching, denoising, and segmentation applications [13]. The one component-level MRI brain segmentation evaluation focused on the model relating image intensity to tissue labels, finding that on simulated data, no one model clearly does best among the ones tested [14].

III. Methods

To provide an objective comparison between MRF optimizers, we implement them within a common framework of an Expectation-Maximization (EM) algorithm that iterates between estimating the statistical distributions of image intensity for each tissue class, based on current voxel-level tissue labels, and estimating tissue labels based on the current tissue intensity models [2].

Our implementation takes as input a T1-weighted brain MRI and a set of initial voxel-level tissue labels. The initial labels are provided by a fully-automated process that warps the input image to a standardized template space via a high-dimensional B-spline transformation [15], [16]. The known tissue labels of voxels in this space, generated by a bootstrapping procedure, are then transformed back to the space of the input image as the initial tissue labeling.

Given the input image and initial segmentation, the main algorithm first computes the mean and variance parameters for Gaussian models of the image intensities within each tissue class. These intensity models are then used to estimate the probability that each voxel belongs to each of the three tissue classes. An MRF optimizer then solves for a voxel labeling that respects the tissue probabilities while encouraging the labelings to be spatially smooth. The voxel labels are then used to calculate new tissue class intensity model parameters, and so on. The algorithm converges when the Kullback−Leibler divergence between tissue intensity models across iterations falls below a threshold.

While the original algorithm [2] used ICM for MRF optimization, we implemented BP, GC, and ICM each for this purpose. Our implementation also diverges from the original algorithm by adding the warping-based label initialization, and by using a gradient filter to prevent the MRF from encouraging spatial smoothness of the tissue labelings across edges.

1) Belief Propagation: Belief propagation (BP) [17] was proposed as a fast way to perform exact label inference on tree-structured graphs, and although it is known to provide sub-optimal label estimates in more general graphs, it performs well in practice on real-world problems [18]. In BP, voxels pass numerical messages to neighboring voxels about their beliefs that the neighbors should be assigned to each of the possible tissue labels. Voxels maintain their beliefs in each of the possible tissue labels based on input from their neighbors together with the tissue intensity models. The messages are passed iteratively until the voxel beliefs converge.

2) Graph Cuts: A well-studied computer science task called maximum-flow/minimum-cut is defined by a graph with two designated positions connected by multiple paths of nodes and edges, where the goal is determining a maximizing configuration of "flow" across edges between them. It was applied to problems of binary image segmentation in 1989, allowing the use of existing algorithms to find an exact global minimum [19]. In 2001, Boykov *et. al.* extended it to n-ary label sets by repeating binary cuts over pairs of possible labels, and proposed a faster, approximate minimization method [20]. Today, GC and BP are both frequently used in current literature.

3) Iterated Conditional Modes: Iterated conditional modes (ICM) [21] is an iterative, greedy algorithm in which each voxel label is set to the most probable according the local evidence from tissue intensity models and the current tissue labels of neighboring nodes only, ignoring longer-range dependencies required for a globally optimal solution. This process is applied repeatedly until convergence. While later methods have surpassed ICM, it is still employed in some current literature for its simplicity and availability.

A. FAST

FAST is a widely used end-to-end system commonly used in MR tissue segmentation tasks [5]. Like our method above, FAST utilizes EM and Gaussian tissue intensity models. Unlike our implementations, its MRF model utilizes ICM for optimization and is homogenous in space. FAST also differs by using histogram-based initializers.

B. SPM

SPM is a popular analysis software package for brain imaging data. In this work we utilize the tissue segmentation method included in SPM version 5 [4] (here referred to as simply SPM). Like the other presented methods, it employs EM and Gaussian intensity models, but unlike the others, a MRF model is not incorporated. SPM uses an initialization method similar to our implementations where the input image

is warped to a known template space, but utilizes a low-dimensional alignment for this rather than a high-dimensional B-spline warp.

IV. Experiments

1) Longitudinal tissue change: We assessed the viability of the methods for estimation of elderly brain tissue change in 57 cognitively-normal elderly individuals and 60 clinically diagnosed with AD (Alzheimer's Disease) [1]. Each individual received MRI scans at baseline and at followup visits 6 months and 12 months later. For each method, subject, and tissue type we fit a linear regression line to the plot of tissue volume against time, and used the line slope to measure rate of tissue change. Line slopes indicating increases in GM or WM, or decreases in CSF, are biologically implausible in this population; we used the percentage of subjects whose rates of tissue change were implausible as a measure of segmentation method validity. To quantify the magnitude of these implausibility errors, we calculated the median and maximum of the magnitudes of such implausible change rates (Table I). ICM generally provided the smallest number of implausible estimates, but BP provided the smallest magnitude of such implausible change. Varying the σ parameter of the gradient filter over a range from 1 to 2, and varying the slope of the gradient-modulating sigmoid function from 2.7 to 5, did not appreciably alter the findings. All methods provided relatively higher numbers of implausible change estimates in WM, whose rate of change in both healthy aging and AD is close to 0. Outside of number of implausible WM change estimates, neither SPM nor FAST ranked highest on any performance criterion.

2) Consistency across scanners: We scanned a group of individuals multiple times over a short period of time, on multiple scanners, and assessed variability in estimated tissue volumes across scanners. A set of 8 cognitively-normal elderly individuals received one T1-weighted and one FLAIR MRI scan on a pair of 1.5T MRI scanners (see [22] for acquisition details); an additional 5 individuals received another T1-weighted and FLAIR scan on a Siemens Trio 3T MRI scanner. For each pair of scanners, segmentation method, and tissue type, we calculated the intraclass correlation coefficient (ICC) in estimated tissue volume between scanners (Table I). ICC values closer to 1 indicate stronger agreement. Among MRF methods, GC and BP generally provided the highest inter-scanner agreement, except for 1.5T agreement in CSF. Inter-scanner agreements in GM, and agreements between 1.5T and 3T in WM, were relatively lower for FAST, but its agreements in CSF volumes were either superior to or comparable to the remaining methods. SPM failed to adequately segment several images, and was the highest-ranking method on none of the performance criteria.

3) Simulated data with ground truth: Each of the five methods were used to segment the BrainWeb [3] template image, which is provided with ground-truth tissue probability maps. Using the BrainWeb simulator [23] we added five levels of Gaussian noise to the template image: 0%, 3%, 5% 7% and 9%. For each noise level, tissue type, and method, a percentage error for the estimated tissue volume was calculated by comparing it to the ground truth tissue volume (Table II). Among MRF methods, BP and ICM provided the highest WM and CSF accuracy respectively. GC and BP provided the highest accuracies among GM, with BP performing slightly better on the higher-noise images. SPM gave higher accuracy GM and CSF estimates than any of these methods on all images; SPM and FAST achieved comparably high performance, especially on WM and GM, and BP gave slightly better WM estimates than SPM on the lower-noise images.

	Inter-Scanner Agreement			**Implausible Longitudinal Change**		
Method	1.5T A vs. 1.5T B	1.5T A vs. 3T	1.5T B vs. 3T	Percent Increasing / Decreasing	Mean Increase / Decrease	Max Increase / Decrease
	White Matter					
Belief Prop.	0.097	0.541	0.116	*58.9%*	***9.5***	***26.5***
Graph Cuts	*0.102*	***0.634***	***0.145***	78.5%	20.7	120.9
ICM	0.101	0.37	0.119	75.7%	27.9	113.7
SPM	-0.144	0.201	-0.337	**36.8%**	11.9	49.9
FAST	**0.271**	0.117	-0.065	55.1%	22.8	200.6
	Gray Matter					
Belief Prop.	0.575	***0.738***	***0.712***	34.6%	***8.2***	***33.3***
Graph Cuts	***0.584***	0.508	0.397	24.3%	15.6	72.9
ICM	0.163	0.694	0.157	***9.3%***	28.4	68.7
SPM	-0.029	0.433	-0.036	29.2%	14.7	54.0
FAST	0.239	0.344	0.084	18.7%	28.6	200.8
	CSF					
Belief Prop.	0.524	*0.451*	0.627	15.9%	***-4.2***	***-31.3***
Graph Cuts	0.52	0.416	*0.653*	10.3%	-7.3	-39.1
ICM	***0.756***	0.338	0.42	***8.4%***	-11.9	-39.8
SPM	0.06	0.071	0.225	20.8%	-11.3	-57.7
FAST	0.752	**0.578**	**0.885**	15.0%	-10.2	-59.6

TABLE I
SUMMARY METHOD PERFORMANCE ON INTER-SCANNER AGREEMENT AND IMPLAUSIBLE LONGITUDINAL CHANGE EXPERIMENTS. FOR INTER-SCANNER AGREEMENT, LISTED VALUES ARE INTRACLASS CORRELATION COEFFICIENTS (ICCS) BETWEEN PAIRS OF MRI SCANNERS, INCLUDING TWO 1.5T SCANNERS (1.5T A AND 1.5T B) AND ONE 3T SCANNER. FOR LONGITUDINAL CHANGE, LISTED VALUES ARE MEAN AND MAXIMUM RATES OF INCREASE AND DECREASE IN UNITS OF CC PER YEAR. IN EACH COLUMN, THE OVERALL METHOD AND THE MRF OPTIMIZER PROVIDING THE LARGEST ICC, FEWEST IMPLAUSIBLE CHANGE RATES, OR SMALLEST MAGNITUDE OF IMPLAUSIBLE CHANGE ARE SHOWN IN BOLD AND ITALICS RESPECTIVELY.

V. Discussion

The key finding of this study is that in aggregate, GC and BP provided elderly brain tissue segmentations that were competitive with, or superior to, the more traditional ICM and state-of-the-art systems SPM and FAST. However, each method had its own strength: BP provided milder errors in depiction of longitudinal change, BP and GC provided comparably high inter-scanner agreement, ICM provided fewer implausible change estimates, SPM excelled in the young brain, and FAST had some advantages, especially in inter-scanner CSF agreement. The two implications of these findings are, first, that in the context of aging, there is room to improve upon the brain tissue segmentation performance provided by off-the-shelf systems; and second, that optimal algorithmic choices for brain segmentation may depend on the relative importance of various performance criteria.

Method	0%	3%	5%	7%	9%
	White Matter				
Belief Prop.	***2.1%***	***0.6%***	*1.4%*	*3.3%*	*4.3%*
Graph Cuts	8.1%	4.0%	6.6%	9.0%	13.1%
ICM	13.4%	14.4%	16.4%	17.3%	19.4%
SPM	18.3%	3.2%	4.5%	**2.5%**	**0.5%**
FAST	43.8%	4.3%	**0.7%**	2.9%	5.4%
	Gray Matter				
Belief Prop.	10.6%	10.7%	*11.2%*	*12.3%*	*13.3%*
Graph Cuts	*8.7%*	*8.7%*	13.8%	16.5%	21.2%
ICM	14.7%	15.1%	16.8%	17.7%	19.9%
SPM	**8.5%**	**2.2%**	**1.1%**	**1.4%**	**0.4%**
FAST	27.2%	3.5%	5.3%	4.8%	4.9%
	CSF				
Belief Prop.	55.8%	52.7%	49.4%	48.4%	48.9%
Graph Cuts	64.5%	54.6%	45.0%	47.8%	53.0%
ICM	***31.3%***	*30.3%*	*31.0%*	*31.7%*	*33.5%*
SPM	33.9%	**26.5%**	**19.0%**	**13.3%**	**11.5%**
FAST	34.6%	33.8%	27.2%	20.2%	14.2%

TABLE II
PERCENT ERROR IN TISSUE VOLUMES FOR THE BRAINWEB IMAGES, SIMULATED WITH NOISE VARYING FROM 0 TO 9%. IN EACH COLUMN, THE OVERALL METHOD AND MRF OPTIMIZER PROVIDING THE LOWEST PERCENT ERRORS ARE SHOWN IN BOLD AND ITALICS RESPECTIVELY.

As expected, SPM excelled in segmenting the young brain that it already uses in the internals of its segmentation routine; among the MRF optimizers, BP and ICM provided the highest accuracy on WM and CSF, respectively, while BP and GC performed best on GM. The differing pattern of results between the young and old brain emphasizes the importance of developing and validating brain segmentation methods that are optimized for aging and aging-associated neurological disease.

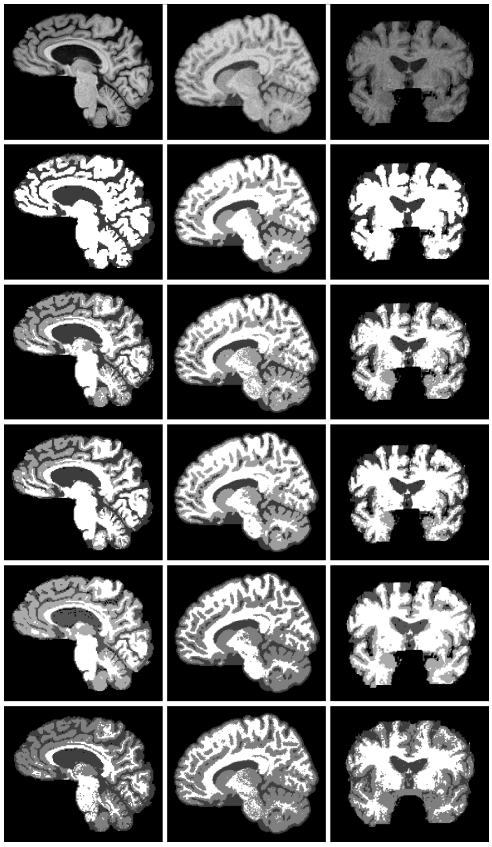

Fig. 1. Tissue segmentation outputs. Seen from top to bottom are the original T1 image and the segmentation by ICM, BP, GC, FAST, and SPM, respectively. From left to right, the data corresponds to a particular subject from the longitudinal tissue change experiment, the BrainWeb simulator with 5% Gaussian noise, and a particular subject from the multiple scanners experiment.

REFERENCES

[1] S. G. Mueller *et al.*, "Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (ADNI)," *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, vol. 1, no. 1, pp. 55–66, 2005.

[2] J. C. Rajapakse *et al.*, "Statistical approach to segmentation of single-channel cerebral MR images," *IEEE TMI*, vol. 16, no. 2, pp. 176–186, 1997.

[3] D. L. Collins *et al.*, "Design and construction of a realistic digital brain phantom," *IEEE TMI*, vol. 17, no. 3, pp. 463–468, Jun. 1998.

[4] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005.

[5] Y. Zhang *et al.*, "Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE TMI*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[6] O. Tsang *et al.*, "Comparison of tissue segmentation algorithms in neuroimage analysis software tools." *Conf. Proc. IEEE EMBS*, vol. 2008, pp. 3924–8, Jan. 2008.

[7] F. Klauschen *et al.*, "Evaluation of automated brain MR image segmentation and volumetry methods." *Human Brain Mapping*, vol. 30, no. 4, pp. 1310–27, Apr. 2009.

[8] H. Lee and I. Prohovnik, "Cross-validation of brain segmentation by SPM5 and SIENAX." *Psychiatry research*, vol. 164, no. 2, pp. 172–7, Nov. 2008.

[9] R. de Boer *et al.*, "Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods." *NeuroImage*, vol. 51, no. 3, pp. 1047–56, Jul. 2010.

[10] J. de Bresser *et al.*, "A comparison of MR based segmentation methods for measuring brain atrophy progression." *NeuroImage*, vol. 54, no. 2, pp. 760–8, Jan. 2011.

[11] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," *IEEE ICCV*, vol. 2, pp. 900–906, 2003.

[12] S. Morales *et al.*, *Graph-Cut versus Belief-Propagation Stereo on Real-World Images*. Springer, 2009, pp. 732–740.

[13] R. Szeliski *et al.*, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors." *IEEE PAMI*, vol. 30, no. 6, pp. 1068–80, Jun. 2008.

[14] M. B. Cuadra *et al.*, "Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images." *IEEE TMI*, vol. 24, no. 12, pp. 1548–65, Dec. 2005.

[15] M. Otte, "Elastic registration of fMRI data using bezier-spline transformations," *IEEE TMI*, vol. 20, no. 3, pp. 193–206, Mar. 2001.

[16] D. Rueckert *et al.*, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE TMI*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[17] J. Pearl, "Reverend Bayes on inference engines: A distributed hierarchical approach," in *AAAI*, 1982, pp. 133–136.

[18] K. P. Murphy *et al.*, "Loopy belief propagation for approximate inference: An empirical study," *UAI*, pp. 467–475, 1999.

[19] D. Greig *et al.*, "Exact maximum a posteriori estimation for binary images," *J. R. Stat. Soc. Ser. B*, vol. 51, no. 2, pp. 271–279, 1989.

[20] Y. Boykov *et al.*, "Fast approximate energy minimization via graph cuts," *IEEE PAMI*, vol. 23, no. 11, pp. 1222–1239, 2001.

[21] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. R. Stat. Soc. Ser. B*, vol. 48, no. 3, pp. 259–302, 1986.

[22] O. Carmichael *et al.*, "MRI predictors of cognitive change in a diverse and carefully characterized elderly population," *Neurobiol Aging*, 2010.

[23] C. A. Cocosco *et al.*, "Brainweb: Online interface to a 3D MRI simulated brain database," *NeuroImage*, vol. 5, p. 425, 1997.