1-1-2016

Manual-protocol inspired technique for improving automated MR image segmentation during label fusion

Nikhil Bhagwat
University of Toronto, Institute of Biomedical Engineering

Jon Pipitone
Centre for Addiction and Mental Health

Julie L. Winterburn
University of Toronto, Institute of Biomedical Engineering

Ting Guo
SickKids Research Institute

Emma G. Duerden
SickKids Research Institute, eduerden@uwo.ca

See next page for additional authors

frontiers in Neuroscience

METHODS
published: 19 July 2016
doi: 10.3389/fnins.2016.00325

Manual-Protocol Inspired Technique for Improving Automated MR Image Segmentation during Label Fusion

Nikhil Bhagwat[1, 2, 3], Jon Pipitone[3], Julie L. Winterburn[1, 2, 3], Ting Guo[4, 5], Emma G. Duerden[4, 5], Aristotle N. Voineskos[3, 6], Martin Lepage[2, 7], Steven P. Miller[4, 5], Jens C. Pruessner[2, 8], M. Mallar Chakravarty[1, 2, 7, 9]* and Alzheimer's Disease Neuroimaging Initiative*

[1] Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada, [2] Cerebral Imaging Centre, Douglas Mental Health University Institute, Verdun, QC, Canada, [3] Kimel Family Translational Imaging-Genetics Research Lab, Research Imaging Centre, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada, [4] Neurosciences and Mental Health, The Hospital for Sick Children Research Institute, Toronto, ON, Canada, [5] Department of Paediatrics, The Hospital for Sick Children and the University of Toronto, Toronto, ON, Canada, [6] Department of Psychiatry, University of Toronto, Toronto, ON, Canada, [7] Department of Psychiatry, McGill University, Montreal, QC, Canada, [8] McGill Centre for Studies in Aging, Montreal, QC, Canada, [9] Biological and Biomedical Engineering, McGill University, Montreal, QC, Canada

OPEN ACCESS

Edited by:
Xi-Nian Zuo,
Chinese Academy of Sciences, China

Reviewed by:
Gang Li,
University of North Carolina at Chapel Hill, USA
Hyunjin Park,
Sungkyunkwan University, South Korea
Andreas Holzinger,
Medical University of Graz, Austria

***Correspondence:**
Nikhil Bhagwat
nikhil.bhagwat@mail.utoronto.ca
M. Mallar Chakravarty
mallar@cobralab.ca

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *21 December 2015*
Accepted: *28 June 2016*
Published: *19 July 2016*

Citation:
Bhagwat N, Pipitone J, Winterburn JL, Guo T, Duerden EG, Voineskos AN, Lepage M, Miller SP, Pruessner JC, Chakravarty MM and Alzheimer's Disease Neuroimaging Initiative (2016) Manual-Protocol Inspired Technique for Improving Automated MR Image Segmentation during Label Fusion. Front. Neurosci. 10:325. doi: 10.3389/fnins.2016.00325

Recent advances in multi-atlas based algorithms address many of the previous limitations in model-based and probabilistic segmentation methods. However, at the label fusion stage, a majority of algorithms focus primarily on optimizing weight-maps associated with the atlas library based on a theoretical objective function that approximates the segmentation error. In contrast, we propose a novel method—Autocorrecting Walks over Localized Markov Random Fields (AWoL-MRF)—that aims at mimicking the sequential process of manual segmentation, which is the gold-standard for virtually all the segmentation methods. AWoL-MRF begins with a set of candidate labels generated by a multi-atlas segmentation pipeline as an initial label distribution and refines low confidence regions based on a localized Markov random field (L-MRF) model using a novel sequential inference process (walks). We show that AWoL-MRF produces state-of-the-art results with superior accuracy and robustness with a small atlas library compared to existing methods. We validate the proposed approach by performing hippocampal segmentations on three independent datasets: (1) Alzheimer's Disease Neuroimaging Database (ADNI); (2) First Episode Psychosis patient cohort; and (3) A cohort of preterm neonates scanned early in life and at term-equivalent age. We assess the improvement in the performance qualitatively as well as quantitatively by comparing AWoL-MRF with majority vote, STAPLE, and Joint Label Fusion methods. AWoL-MRF reaches a maximum accuracy of 0.881 (dataset 1), 0.897 (dataset 2), and 0.807 (dataset 3) based on Dice similarity coefficient metric, offering significant performance improvements with a smaller atlas library (<10) over compared methods. We also evaluate the diagnostic utility of AWoL-MRF by analyzing the volume differences per disease category in the ADNI1: Complete Screening dataset. We have made the source code for AWoL-MRF public at: https://github.com/CobraLab/AWoL-MRF.

Keywords: MR Imaging, segmentation, multi-atlas label fusion, hippocampus, Alzheimer's disease, first-episode-psychosis, premature birth and neonates

INTRODUCTION

The volumetric and morphometric analysis of neuroanatomical structures is increasingly important in many clinical applications. For instance, structural characteristics of the hippocampus have been used as an important biomarker in many neuropsychiatric disorders including Alzheimers disease (AD), schizophrenia, major depression, and bipolar disorder (Harrison, 2004; Frey et al., 2007; Lerch et al., 2008; Kempton et al., 2011; Meda et al., 2013; Weiner, 2013). The gold standard for neuroanatomical segmentation is manual delineation by an expert human rater. However, with the increasing ubiquity of magnetic resonance (MR) imaging technology and neuroimaging studies targeting larger populations, the time and expertise required for manual segmentation of large MR datasets becomes a critical bottleneck in analysis pipelines (Mazziotta et al., 1995, 2001; Pausova et al., 2007). Manual rater performance is dependent on specialized knowledge of the neuroanatomy. A generic manual segmentation protocol leverages this anatomical knowledge and uses it in tandem with voxel intensities to enforce structural boundary conditions during the delineation process. This is, of course, the premise of many automated model-based segmentation approaches.

Multi-atlas based approaches have been shown to improve segmentation accuracy and precision over model-based approaches (Collins et al., 1995; Pruessner et al., 2000; Warfield et al., 2004; Heckemann et al., 2006, 2011; Aljabar et al., 2009; Chakravarty et al., 2009, 2013; Leung et al., 2010; Lötjönen et al., 2010; Sabuncu et al., 2010; Wolz et al., 2010; Wang et al., 2012; Yushkevich et al., 2012). The processing pipelines of these approaches can be divided into multiple stages. First, several atlas images are registered to a target image, i.e., an image to be segmented. Subsequently, the atlas labels are propagated to produce several candidate segmentations of the target image. Finally, a label fusion technique such as voxel-wise voting is used to merge these candidate labels into the final segmentation for the target image. For the remainder of the manuscript we refer this latter stage within a multi-atlas based segmentation pipeline as "label fusion," which is the core interest of this work.

Traditionally, in many image processing and computer vision applications in neuroimaging, the use of Markov Random Field (MRF) has been a popular approach for modeling spatial dependencies and has been used in several model-based segmentation techniques. Existing software packages such as FreeSurfer (Fischl et al., 2002) and FMRIB Software Library (Smith et al., 2004) use MRF for gray matter, white matter, and cerebrospinal fluid classification as well as for segmentation of multiple subcortical structures. For example, FreeSurfer uses an anisotropic non-stationary MRF that encodes the inter-voxel dependencies as a function of location within the brain. Pertaining to multi-atlas label fusion techniques, STAPLE (Simultaneous Truth And Performance Level Estimation; Warfield et al., 2004), uses a probabilistic performance framework consisting of an MRF model and an Expectation-Maximization (EM) inference method to compute the probabilistic estimate of a true segmentation based on an optimal combination of a collection of segmentations. STAPLE has been explored in several studies for improving a variety of segmentation tasks (Commowick and Warfield, 2010; Akhondi-Asl and Warfield, 2012; Commowick et al., 2012; Jorge Cardoso et al., 2013).

Alternatively, a majority of modern multi-atlas approaches treat label fusion as a weight-estimation problem, where the objective is to estimate optimal weights for the candidate segmentation propagated from each atlas. In a trivial case with uniform weights, this label fusion technique boils down to a simple majority vote. In other cases (Aljabar et al., 2009), the weights can be used to exclude atlases that are dissimilar to a target image to minimize the errors from unrepresentative anatomy. In a more general case, weight values are estimated using some similarity metric between the atlas library and the target image. A comprehensive probabilistic generative framework is provided by Sabuncu et al. (2010) that models such an underlying relationship between the atlas and target data, exploited by the methods belonging to this class. More recently, several methods (Coupé et al., 2011; Rousseau et al., 2011; Wang et al., 2012) have extended this label fusion approach by adopting spatially varying weight-maps to capture similarity at a local level. These algorithms usually introduce a bias during label fusion when the weights are assigned independently to each atlas, allowing several atlases to produce similar label errors. These systematic (i.e., consistent across subject cohort) errors can be mitigated by taking pairwise dependencies between atlases into account during weight assignment as proposed in the Joint Label Fusion (JLF) approach (Wang et al., 2012; Yushkevich et al., 2012).

In contrast, the proposed method—Autocorrecting Walks over Localized Markov Random Field (AWoL-MRF)—pursues a different idea for tackling the label fusion problem. We hypothesize that we could achieve superior performance by mimicking the behavior of the manual rater, since virtually all segmentation methods use manual labels to define the gold-standard. Consequently, the label fusion objective developed here comprises capturing the sequential process of manual segmentation rather than optimizing atlas library weights based on similarity measure proxies and/or performing iterative inference to estimate optimal label configurations based on MRFs. Hence the novelty of the approach lies in the methodological procedure as we combine the strong prior anatomical information provided by the multi-atlas framework with the local neighborhood information specific to the given subject.

In the context of segmentation of anatomical structures such as hippocampus, the challenging areas for label assignment are mainly located at the surface regions of the structure. We observe that a manual rater traces these boundary regions by balancing intensity information and anatomical knowledge, while enforcing smoothness requirements and tackling partial volume effects. In practice, this behavior translates into a sequential labeling process that depends on information offered by the local neighborhood around a voxel of interest. For instance, a manual rater would begin by marking a boundary of a structure that they believe to be correct (high-confidence)

based on anatomical knowledge. Next, the rater would identify certain regions that require further refinement (low-confidence). Then, region-by-region (patches), the rater would perform these refinements by moving from high-confidence areas to low in a sequential manner, while taking into account the information offered by neighborhood voxels from orthogonal planes. While not all groups may use this process, this tends to be a dominant order-of-operations for those using the Display tool from the MINC toolkit. This process has been used in many publications by our group (Chakravarty et al., 2008, 2009; Winterburn et al., 2013; Park et al., 2014), and serves as intuition for the development of AWoL-MRF.

The proposed label fusion method attempts to incorporate these observations into an automated procedure and is implemented as part of a segmentation pipeline previously developed by our group (Pipitone et al., 2014). The algorithmic steps of AWoL-MRF can be summarized as follows. First based on a given multi-atlas segmentation method, we initialize the label distribution for a neuroanatomical structure to be segmented. This initial label-vote distribution is leveraged to partition the given target volume in two disjoint subsets comprising regions with high and low confidence label values based on the vote distribution at the voxels. Next we construct a set of local 3-dimensional patches comprising a certain ratio of high and low confidence voxels. The spatial dependencies in these patches are modeled using independent MRFs. Finally, we traverse these patches moving from high to low confidence voxels in a sequential manner and perform the label distribution updates based on a localized (patch-based) MRF model. We implement a novel spanning-tree method to build these ordered sequences of voxels (walks).

We provide a description and extensive validation of our approach in this manuscript, which is organized as follows. First, we describe the AWoL-MRF method and the underlying assumptions in detail. Then, we provide a thorough validation of the method for the whole hippocampus segmentation by conducting multi-fold validation over three independent datasets that span the entire human lifespan. The quantitative accuracy evaluations are performed on three datasets: (1) a subset of the Alzheimer's Disease Neuroimaging Database (ADNI) dataset; (2) a cohort of First Episode Psychosis (FEP) patients; and (3) a cohort of preterm neonates scanned early in life and at term-equivalent age. Additionally we evaluate the diagnostic utility of the method by analyzing the volume differences per disease category in the ADNI1: Complete Screening dataset. We assess the accuracy and robustness of this proposed method (source code: https://github.com/CobraLab/AWoL-MRF) by comparing it with three other approaches. Our group has recently validated the performance of MAGeT-Brain (Pipitone et al., 2014) pipeline against several other automated methods. Here, we make use of MAGeT-Brain to generate candidate labels on which variety of label fusion methods can be implemented. We first compare the performance of AWoL-MRF with the default majority-vote based label fusion used in MAGeT-Brain. In addition, we compare AWoL-MRF with STAPLE (Warfield et al., 2004) and JLF (Wang et al., 2012) label fusion methods.

MATERIALS AND METHODS

Baseline Multi-Atlas Segmentation Method

MAGeT-Brain (https://github.com/CobraLab/MAGeTbrain)—a segmentation pipeline previously developed by our group, is used as a baseline method for comparison (Pipitone et al., 2014). MAGeT-Brain uses multiple manually labeled anatomical atlases and a bootstrapping method to generate a large set of candidate labels (votes) for each voxel for a given target image to be segmented. These labels are generated by first randomly selecting a subset of target images, which is referred as a template library. Then the atlas segmentations are propagated to the template library via transformations estimated by nonlinear image registration. Subsequently, these template library segmentations are propagated to each target image and these candidate labels are fused using a label fusion method. The number of candidate labels is dependent on the number of available atlases and number of templates. In a default MAGeT-Brain configuration, the candidate labels are fused by a majority vote. In previous investigations by our group (Chakravarty et al., 2013; Pipitone et al., 2014), we observed no improvements when we used cross correlation and normalized mutual information based weighted voting (Studholme et al., 1999). For the purposes of this manuscript, candidate labels generated using MAGeT-Brain will be used to serve as the input to AWoL-MRF, STAPLE, and the default majority vote label fusion methods. The use of candidate labels is non-trivial in the case of label fusion with JLF, as this method requires coupled atlas image and label pairs as input. The permutations in MAGeT-Brain pipeline generate candidate labels totaling to *number of atlases* $\times$ *number of templates*. These candidate labels no longer have unique corresponding intensity images associated with them. The use of identical atlas (or template) library images as proxies is likely to deteriorate the performance of JLF, as it models the joint probability of two atlases making a segmentation error based on intensity similarity between a pair of atlases and the target image (Wang et al., 2012). Therefore, no template library is used during JLF evaluation. Note that even though MAGeT-Brain is used as a baseline method for the performance validation in this work, AWoL-MRF is a generic label fusion algorithm that can be used with any multi-atlas segmentation pipeline that produces a set of candidate labels.

Proposed Label Fusion Method: AWoL-MRF

A generic label fusion method involves some sort of voting technique, such as a simple majority or some variant of weighted voting, which combines labels from a set of candidate segmentations derived from a multi-atlas library. These voting techniques normally yield accurate performance at labeling the core regions of an anatomical structure; however, the overall performance is dependent on the structural variability accounted by the atlas library. Especially in cases where only a small number of expert atlases are available, the resultant segmentation of a target image can be split into two distinct regions - areas with (near) unanimous label votes and areas with divided label votes. The proposed method incorporates this observation by

partitioning the given image volume into two subsets based on the label vote distribution (number of votes per label per voxel) obtained from candidate segmentations. Subsequently, these partitions are used to generate a set of patches on which we construct MRF models to impose homogeneity constraints in the given neighborhood spanned by each patch. Finally, the voxels in these localized MRFs are updated in a sequential manner incorporating the intensity values and label information of the neighboring voxels. A detailed description of this procedure is provided below.

Image Partitioning

Let S be a set comprising all voxels in a given 3-dimensional volume. Then an image I comprising gray-scale intensities and the corresponding label volume are defined as:

$$I(S) : \{x \in S\} \rightarrow \mathbb{R} \quad (1)$$

$$L^j(S) : \{x \in S\} \rightarrow \{0,\ 1\} \quad (2)$$

Thus, L^j represents the jth candidate segmentation volume comprising binary label values (background:0 and structure:1) for a given image. Then with J candidate segmentations, we can obtain a label-vote distribution through voxel-wise normalization.

$$V(S) = \frac{\sum_j w^j L^j(S)}{J} \quad (3)$$

Where, w^j is the weight assigned to the jth candidate segmentation. Now, $V(S)$ represents the label probability distribution over all the voxels in the given image. For an individual voxel, it provides the probability of belonging to a particular structure: $V(x_i) = P(L(x_i) = 1) = 1 - P(L(x_i) = 0)$. Now, we split set S into two disjoint subsets S_H (high-confidence region) and S_L (low-confidence region) such that.

$$S_H = \{x \in S \,|\, V(x_i) > L_T^0 \cup V(x_i) > L_T^1\}$$
$$S_L = \{x \in S \,|\, x \notin S_H\} \quad (4)$$

where, L_T^0 and L_T^1 are the voting confidence thresholds for $L = 0$ and $L = 1$, respectively. Note that in the generic majority vote scenario $L_T^0 = L_T^1 = 0.5$ and S_L collapses to an empty set. In order to identify and separate low-confidence regions, these thresholds are set at higher values (>0.5) and can be adjusted based on empirical evidence (see Section Parameter Selection). As mentioned earlier, voting distributions usually form a near consensus (uni-modal) toward a particular label at certain locations, such as the core regions of structures, and therefore these voxels are assigned to the high-confidence subset. In contrast, other areas that have split (flat) label distribution are assigned to low-confidence subset.

Patch Based Graph Generation

From here on, we will refer to voxels as nodes, in keeping with graph-theory convention. The partitioning operation reduces the number of nodes to be re-labeled by a significant amount. However, considering the size of the MR images, selecting a single MRF model consisting of all S_L nodes and their neighbors is a computationally expensive task. Additionally, the unified model usually considers global averages over an entire structure during parameter estimation for choice of prior distributions, such as $P(intensity \mid label)$, which may not be ideal in cases where local signal characteristics show spatial variability. Therefore, we propose a patch-based approach, which further divides the given image in smaller subsets (3-dimensional cubes) comprising S_H as well as S_L nodes. The subsets are created with a criterion imposing a minimum number requirement of S_H nodes in a given patch. This criterion essentially dictates the relative composition of S_H and S_L nodes in the patch—which is referred as the "mixing ratio" parameter in this manuscript. The impact of this heuristic method of patch generation is discussed in Section Parameter Selection. The basic idea behind this approach is to utilize the information offered by the S_H neighbors via pairwise interactions (doubleton clique) along with the local intensity information to update the label-likelihood of S_L voxels. The implemented algorithm to generate these patches is described below.

First, the S_L nodes are sorted based on the number of S_H nodes in their 26-node neighborhood. Next, thresholding on the mixing ratio parameter, top S_L nodes from the sorted list are selected as seeds. Then, the patches are constructed centered at these seeds with pre-defined length (L_{patch}). **Figure 1A** shows the schematic representation of the S_H, S_L partitions based on initial label distribution ($V(S)$), as well as the overlaying patch-based subsets comprising S_H and S_L nodes. Note that depending on parameter choice (mixing-ratio and patch-length), these patches may not be strictly disjoint. In this case, the nodes in overlapping patches are assigned to a single patch based on a simple metric, such as its distance from the seed node. Additionally, these patches may not cover the entire S_L region. These unreachable S_L nodes are labeled according to the baseline majority vote. These two edge cases can be mitigated with sophisticated graph partitioning methods—nevertheless based on our exploratory investigation, such methods prove to be computationally expensive, and yield minimal accuracy improvements.

Localized Markov Random Field Model

As seen from **Figures 1A,B**, the MRF model is built on nodes in a given patch (S_p). The probability distribution associated with the particular field configuration (label values of the voxels in the patch) can be factorized based on the cliques of the underlying graph topology. With first-order connectivity assumption, we get a 3-dimensional grid topology, where each node (excluding patch edges) has six connected neighbors along the Cartesian axes. Consequently, this graph topology yields two types of cliques. The singleton clique (C_1) of S_P is a set of all the voxels contained in that patch. Whereas the doubleton clique (C_2) is a set consisting of all the pairs of neighboring voxels in the given patch. Then, for the MRF model, the total energy (U) of a given label configuration (y) is given by the sum of all clique potentials (V_C) in this MRF model:

$$U(y) = \sum\nolimits_{c \in C} V_c(y) = \sum\nolimits_{i \in C_1} V_{C_1}(y_i) + \sum\nolimits_{i,j \in C_2} V_{C_2}(y_i,\ y_j) \quad (5)$$

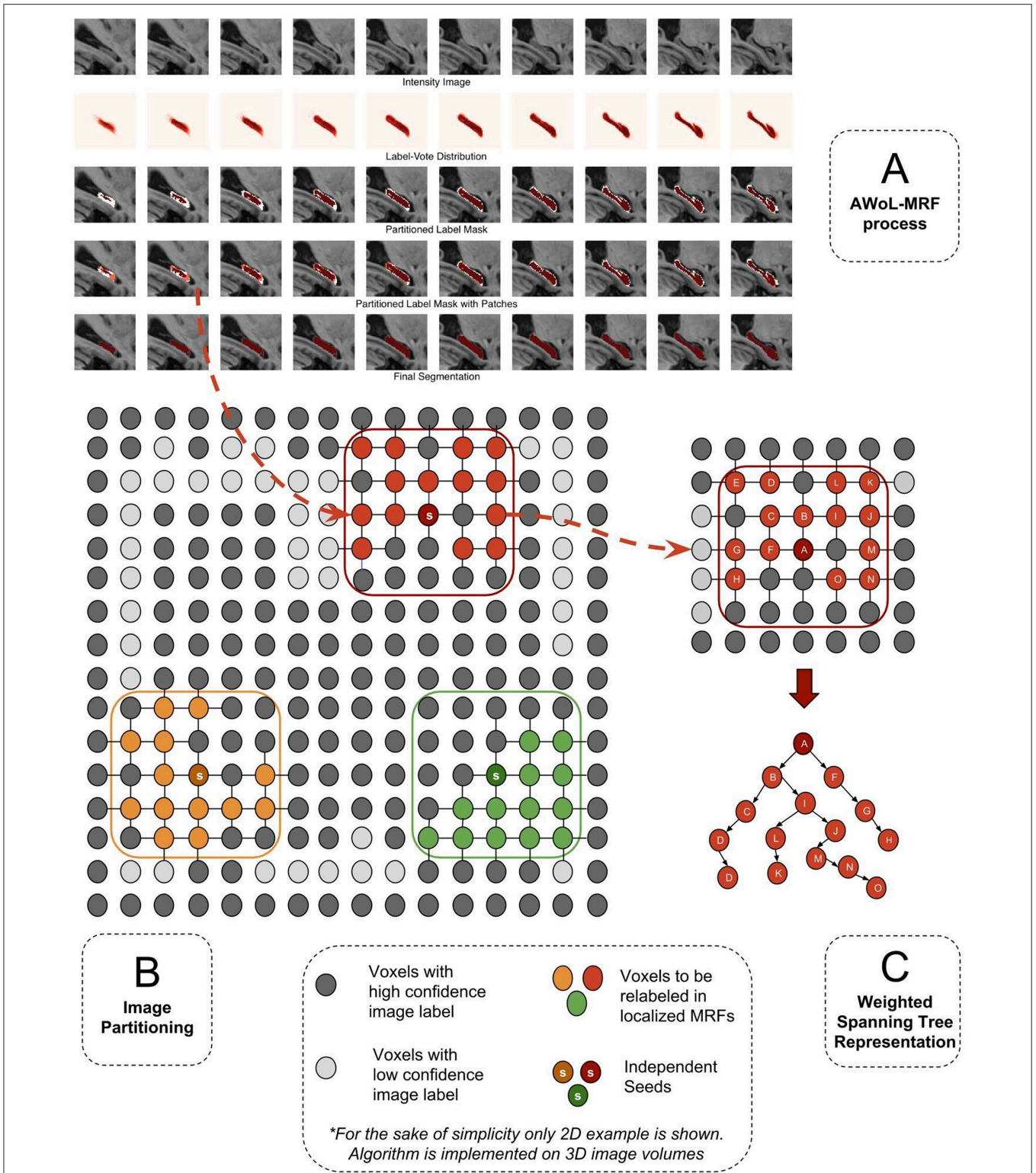

FIGURE 1 | (A) The segmentation of a sample hippocampus in sagittal view during various stages of algorithm. Row 1: The target intensity image to be segmented. Row 2: The voxel-wise label vote distribution map for the target image based on candidate labels. Row 3: Image partitioning comprising two disjoint regions (high confidence: red, low confidence: white). Row 4: Orange Patches (localized MRFs:) comprising low confidence voxels. Row 5: Fused target labels. **(B)** Image partitioning into certain and uncertain regions and generation of patches. **(C)** Transformation of MRF graph into spanning tree representation. The tree is traversed starting from the root (seed) node and successively moving toward the leaf nodes.

where, $y : \{L(x_i) | x_i \in S_p\}$. Now, assuming that voxel gray-scale intensities ($f_i = I(x_i)$) follow a Gaussian distribution given the label value, we get the following relation for the singleton clique potential based on the MRF model.

$$V_{C_1}(y_i) = \log(P(f_i|y_i)) = -\log(\sqrt{2\pi}\sigma_{y_i}) - \frac{(f-\mu_{y_i})^2}{2\sigma_{y_i}^2} \quad (6)$$

The mean and variance of the Gaussian model can be estimated for each patch empirically, utilizing the S_H nodes in the given patch as a training set. This approach proves to be advantageous especially in the context of T1-weighted images of the brain, as intensity distributions tend to fluctuate spatially. The doubleton clique potentials are modeled to favor similar labels at the neighboring nodes and are given by the following relation.

$$V_{C_2}(y_i, y_j) = -\beta\, d(y_i, y_j) = \begin{cases} -\beta & \text{if } y_i = y_j \\ +\beta & \text{if } y_i \neq y_j \end{cases} \quad (7)$$

The β parameter can be estimated empirically using the atlas library (Sabuncu et al., 2010). As β increases the regions become more homogeneous. This is discussed further in Section Parameter Selection. Finally, the posterior probability distribution of the label configuration can be computed using Hammersley-Clifford theorem, and is given by:

$$\begin{aligned} P(y|f) &= \frac{1}{z} exp(-U(y)) \\ P(y|f) &\propto \sum_{i \in C_1} \left(\log\left(\sqrt{2\pi}\sigma_{y_i}\right) + \frac{(f-\mu_{y_i})^2}{2\sigma_{y_i}^2} \right) \\ &\quad + \sum_{i,j \in C_2} \beta d(y_i, y_j) \end{aligned} \quad (8)$$

where Z is the partition function that normalizes configuration energy (U) into a probability distribution. The maximum *a posteriori* (MAP) label distribution is given by:

$$y^{MAP} = \text{argmax}_y\, P(y|f) = \text{argmin}_y\, U(y) \quad (9)$$

The posterior segmentation can be computed using a variety of optimization algorithms as described in the next section.

Inference

This section provides the details of the optimization technique used to compute posterior label distribution. Common iterative inference and learning methods such as Iterated Conditional Modes (ICM) and Expectation Maximization (EM) are computationally intensive, and ICM variants often suffer from greedy behavior that results in local optima. Here, we present an alternative approach that computes the posterior label distribution in a non-iterative, online process, minimizing computational costs. The intuition behind this approach is to mimic manual tracing protocols where the delineation process traverses from higher-confidence regions to lower-confidence regions in a sequential manner. In order to follow such a process, we transform the undirected graph structures defined by the MRF patches into directed spanning trees (see **Figure 1C**). Then we compute the posterior label distributions one voxel at a time as we traverse (*walk*) through the directed tree exhaustively. The directed tree structure mitigates the need for iterative inference over loops within the original undirected graph. The following is a brief outline of the implementation of the inference procedure:

1. Initialize all voxels to the labels given by the mode of baseline label distribution.
2. Transform the graph consisting of S_L nodes within an MRF patch into a directed tree graph, specifically a spanning tree graph with seed voxels as the root of the tree. This transformation is computed using a minimum spanning tree (MST) method (Prim's Algorithm; Prim, 1957), which finds the optimal tree structure based on a predefined edge-weight criterion. In this method, the weights are assigned based on the node adjacency and voxel intensity gradients.

$$w(x_i, x_j) = \begin{cases} (f_i - f_j)^2 & \text{if } d(x_i, x_j) = 1 \\ \infty & \text{if } d(x_i, x_j) \neq 1 \end{cases} \quad (10)$$

   where $d(x_i, x_j)$ is a graph metric representing distance between two vertices.
3. Traverse through the entire ordered sequence of the MST to update the label at each voxel using Equation (9).
4. Repeat this process for all MRF patches.

VALIDATION EXPERIMENTS

Datasets

For complete details please refer to the Supplementary Materials.

Experiment I: ADNI Validation

Data used in this experiment were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). The dataset consists of 60 baseline scans in the ADNI1: Complete 1Yr 1.5T standardized dataset (Jack et al., 2011; Wyman et al., 2013). The demographics of this cohort are summarized in **Table 1**. The manual segmentations for the hippocampus (ADNI-specific) were generated by expert raters following the Pruessner-protocol (Pruessner et al., 2000). These manual segmentations were used for validation and performance comparisons.

Experiment II: First Episode Psychosis (FEP) Validation

Data used in this experiment were obtained from the Prevention and Early Intervention Program for Psychoses (PEPP-Montreal), a specialized early intervention service at the Douglas Mental Health University Institute in Montreal, Canada (Malla et al., 2003). The dataset consists of structural MR images (1.5T) of 81 subjects. The demographics of this cohort are summarized in **Table 2**. The manual segmentations for the hippocampus were generated by expert raters following the Pruessner-protocol (Pruessner et al., 2000).

TABLE 1 | ADNI1 cross-validation subset demographics.

	CN ($N = 20$)	LMCI ($N = 20$)	AD ($N = 20$)	Combined ($N = 60$)
Age (Years)	72.2, 75.5, 80.3	70.9, 75.6, 80.4	69.4, 74.9, 80.1	70.9, 75.2, 80.2
Sex (Female)	50% (10)	50% (10)	50% (10)	50% (30)
Education	14.0, 16.0, 18.0	13.8, 16.0, 16.5	12.0, 15.5, 18.0	13.0, 16.0, 18.0
CDR-SB	0.00, 0.00, 0.00	1.00, 2.00, 2.50	3.50, 4.00, 5.00	0.00, 1.75, 3.62
ADAS 13	6.00, 7.67, 11.00	14.92, 20.50, 25.75	24.33, 27.00, 32.09	9.50, 18.84, 26.25
MMSE	28.8, 29.5, 30.0	26.0, 27.5, 28.2	22.8, 23.0, 24.0	24.0, 27.0, 29.0

CN, Cognitively Normal; LMCI, Late-onset Mild Cognitive Impairment; AD, Alzheimer's Disease; CDR-SB, Clinical Dementia Rating-Sum of Boxes; ADAS, Alzheimer's Disease Assessment Scale; MMSE, Mini-Mental State Examination; Values are presented as lower quartile, median, and upper quartile for continuous variables, or as a percentage (frequency) for discrete variables.

TABLE 2 | First Episode Psychosis subject demographics.

	*N**	*FEP (N = 81)*
Age	80	21 23 26
Gender: M	81	63% (51)
Handedness: ambi	81	6% (5)
Left		5% (4)
Right		89% (72)
Education	81	11 13 15
SES: Lower	81	31% (25)
Middle		54% (44)
Upper		15% (12)
FSIQ	79	88 102 109

Ambi, ambidextrous; SES, Socioeconomic Status score; FSIQ, Full Scale IQ. Values are presented as lower quartile, median, and upper quartile for continuous variables, or as a percentage (frequency) for discrete variables. N is the number of non-missing values.*

Experiment III: Preterm Neonatal Cohort Validation

This cohort consists of 22 premature neonates whose anatomical images (1.5T) were acquired at two time points, once in the first weeks after birth when clinically stable and again at the term-equivalent age (total of 44 images: 22 early-in-life and 22 term-age equivalent). The whole hippocampus was manually segmented by an expert rater using a 3-step segmentation protocol. The protocol adapts histological definitions (Duvernoy et al., 2005), as well as existing whole hippocampal segmentation protocols for MR images (Pruessner et al., 2000; Winterburn et al., 2013; Boccardi et al., 2015) to the preterm infant brain (Guo et al., 2015).

Experiment IV: Hippocampal Volumetry

The volumetric analysis was performed using the standardized ADNI1: Complete Screening 1.5T dataset (Wyman et al., 2013) comprising 811 ADNI T1-weighted screening and baseline MR images of healthy elderly (227), MCI (394), and AD (190) patients. The segmentations were produced using 9 atlases (segmented following the Pruessner-protocol) with each method. For majority vote, STAPLE, and AWoL-MRF the number of templates was set to 19. As mentioned earlier, the use of templates is not possible with JLF due to coupling between image and label volumes from the atlas library. In the first part of analysis, we compared the mean hippocampal volume measurements per diagnosis (AD: Alzheimer's disease patients, MCI: subjects with mild cognitive impairment, CN: cognitively normal). Then in the second part of analysis, we compared the mean hippocampal volume measurements of two MCI sub-groups: MCI-converters (65 subjects converting from MCI to AD diagnosis) and MCI-stable (285 subjects with stable MCI diagnosis) within 1 year from the screening time-point. Furthermore, during both parts, we performed analysis using a linear model predictive of hippocampal volume based on diagnostic category along with "age," "sex," and "total-brain-volume" as covariates (data used from ADNIMERGE table from the ADNI database).

Label Fusion Methods Compared

We compared the performance of AWoL-MRF against MAGeT-Brain majority vote, STAPLE, and JLF. The basic approach of these label fusion methods is described below.

MAGeT-Brain Majority Vote

As described in Section Proposed Label Fusion Method: AWoL-MRF, the MAGeT-Brain pipeline uses a template library sampled from the subject image pool. Consequently, the total number of candidate labels (votes) prior to label fusion equals number of atlases $\times$ number of templates. In the default MAGeT-Brain configuration, these candidate labels are fused based on simple majority vote.

Simultaneous Truth and Performance Level Estimation (STAPLE)

STAPLE (Simultaneous truth and performance level estimation; Warfield et al., 2004), is a probabilistic performance model that tries to estimate underlying ground-truth labels from a set of manual or automatic segmentations generated by multiple raters or methods. Note that STAPLE does not consider the intensity values from the subject image in its model comprising MRF. STAPLE carries out label fusion in an Expectation-Maximization framework and estimates performance of a manual rater or an automatic segmentation method for each label class—which is then used to find the optimal segmentation for the subject image. Software implementation of STAPLE was obtained from the Computational Radiology Laboratory (http://www.crl.med.harvard.edu/software/STAPLE/index.php).

Joint Label Fusion (JLF)

Among the modern label fusion approaches incorporating spatially varying weight-distribution, JLF also accounts for the dependencies within the atlas library (Wang et al., 2012). These

dependencies are estimated based on an intensity similarity measure between a pair of atlases and a target image in a small neighborhood surrounding a voxel. This approach allows mitigation of bias typically incurred by the presence of similar atlases. Software implementation of JLF was obtained from the ANTs repository on Github (https://github.com/stnava/ANTs/blob/master/Scripts/antsJointLabelFusion.sh).

Evaluation Criteria

For experiments I, II, and III, we performed both quantitative and qualitative assessment of the results. The segmentation accuracy was measured using Dice similarity coefficient (DSC), given as follows:

$$DSC = \frac{2\,|A \cap B|}{|A| + |B|} \tag{11}$$

where A and B are the three dimensional label volumes being compared. We also evaluated the level of agreement between automatically computed volumes and manual segmentations using Bland-Altman plots (Bland and Altman, 1986). Bland-Altman plots were created with segmentations generated from 5 to 19 templates configuration. For the ADNI and FEP datasets, we performed three-fold cross validation and obtained the quantitative scores by averaging over all the validation rounds, as well as the left and right hippocampal segmentations. Constrained by the size of the Premature Birth and Neonatal dataset and the quality of certain images which caused difficulties in the registration pipeline, we simply performed a single round of validation to determine if the results that we found in Experiments I and II were generalizable to brains with radically different neuroanatomy. Due to incomplete myelination of the brain, the neonatal MR images have drastically different contrast levels. The intensity values for the hippocampus are reversed relative to T1-weighted images of the adolescent or adult human brains. These distinct attributes make it an excellent "held out sample" or "independent test-set" for performance evaluation. Thus, for this dataset, the quantitative scores are averages over left and right hippocampi over a single validation round.

Additionally, we also verified the segmentation precision using surface based metric analogous to the Hausdorff distance (Chakravarty et al., 2009). The details of this evaluation are reported in the Supplementary Materials.

RESULTS

Experiment I: ADNI Validation

For ADNI dataset, the mean Dice score of AWoL-MRF maximizes at 0.881 with 9 atlases and 19 templates. As seen from **Figure 2**, AWoL-MRF outperforms both majority vote (0.862), STAPLE (0.858), and JLF label fusion (0.873) methods.

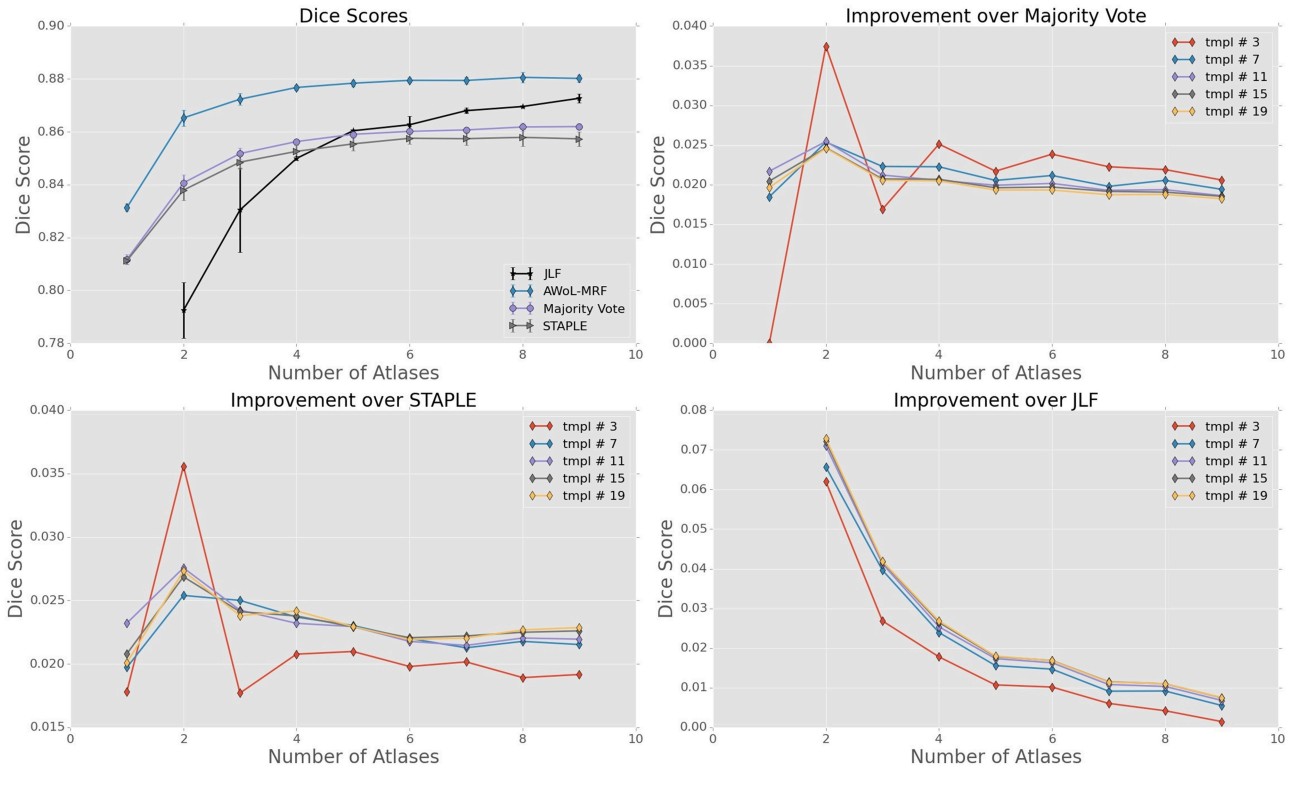

FIGURE 2 | Experiment I DSC: All results show the average performance values of left and right hippocampi over three-fold validation. The top-left subplot shows mean DSC score performance of all the methods. Remaining subplots show the mean DSC score improvement over compared methods for different number of templates (bootstrapping parameter of MAGeT-Brain).

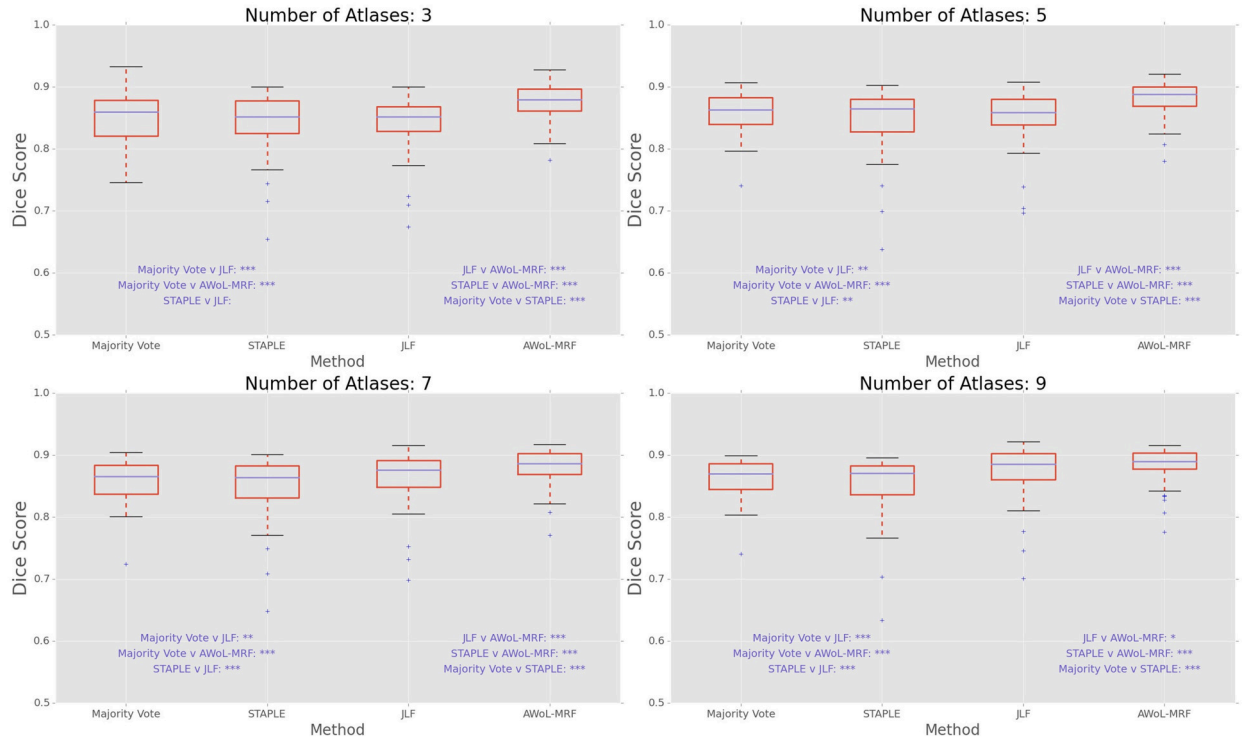

FIGURE 3 | Experiment I DSC: Statistical comparison of the performance of all methods for different atlas library sizes. The statistical significance is reported for pairwise comparisons (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

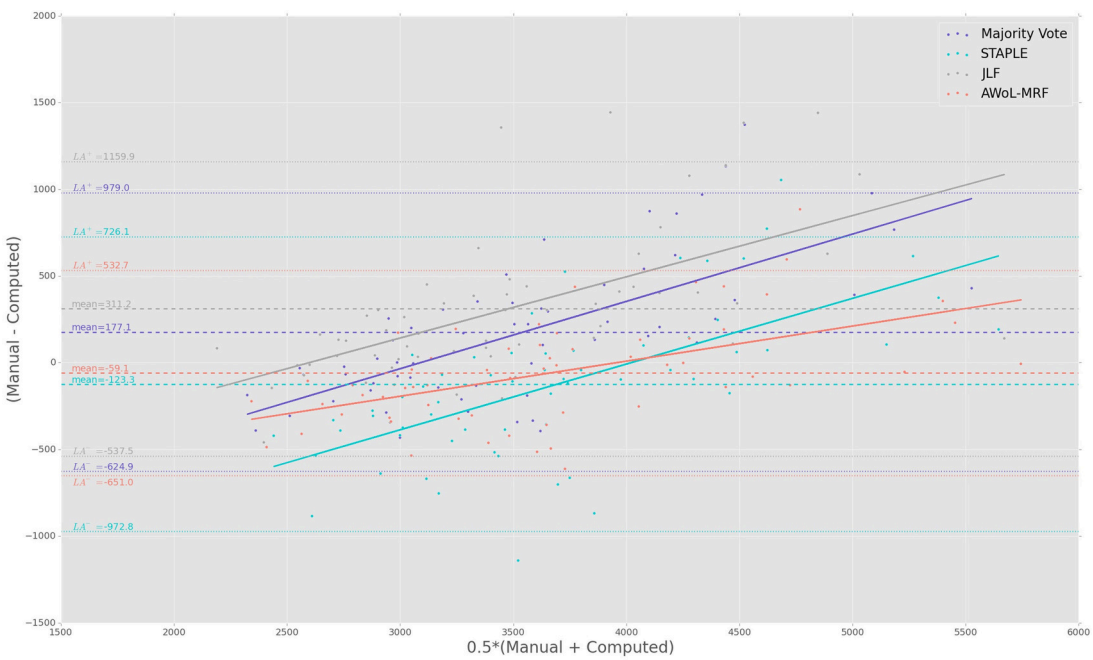

FIGURE 4 | Experiment I Bland-Altman Analysis: Comparison between computed and manual volumes (in mm^3) for single parameter configuration of 9 atlases and 19 templates. The overall mean difference in volume, and limits of agreement (LA+/LA−: 1.96 SD) are shown by dashed horizontal lines. Linear fit lines are shown for each method. Note that the points above the mean difference indicate underestimation of the volume with respect to the manual volume, and vice versa.

Particularly compared to JLF, more improvement is seen with fewer atlases as AWoL-MRF reaches mean Dice score of 0.880 with only 6 atlases. The improvement diminishes with an increasing number of atlases and a smaller number of templates (bootstrapping parameter for generating candidate labels). Additionally, AWoL-MRF helps reduce the bias introduced by certain majority vote techniques while arbitrarily breaking voteties in the cases of even number of atlases, as previously described by our group and others (Heckemann et al., 2006; Pipitone et al., 2014). We find that AWoL-MRF corrects these decreases in performance, which is evident by the extra boosts in accuracy for the cases with an even number of atlases.

DSC distribution comparisons for a four sample configurations (number of atlases = 3, 5, 7, 9; number of templates = 11) are shown in **Figure 3**. These plots reveal that AWoL-MRF provides statistically significant improvement over all other methods regardless of size of the atlas library. As expected, we also notice the reduction in variance with an increasing number of atlases.

The Bland-Altman plots reveal the biases incurred with the application of each automatic segmentation method during volumetric analysis. **Figure 4** shows that all four methods have a proportional bias associated with their volume estimates. Specifically, we see that in all four methods, the volumes of the smaller hippocampi are overestimated, whereas the larger hippocampi are underestimated. Nevertheless, AWoL-MRF shows the smallest magnitude of mean bias, along with tighter limits of agreement across the cohort. STAPLE displays similar mean bias values, but higher variance in volume estimation compared to AWoL-MRF, which is evident by its steeper line-slope and wider limits of agreements. Majority vote and JLF show the highest amount of positive mean bias indicating a tendency toward underestimation of hippocampal volume.

Qualitatively, improvement in segmentations is seen on the surface regions of the hippocampus. As seen in **Figure 5**, spatial homogeneity is improved as well.

Experiment II: FEP Validation

For the FEP dataset, the mean Dice score of AWoL-MRF maximizes at 0.897, with 9 atlases and 19 templates. Similar to Experiment I, the AWoL-MRF consistently outperforms the majority vote (0.891), STAPLE (0.892), and JLF (0.888) methods; however, the improvement is comparatively modest. More improvement is seen with fewer atlases when compared to JLF, as AWoL-MRF surpasses the mean Dice score of 0.890 with only 3 atlases (see **Figure 6**). The improvement diminishes

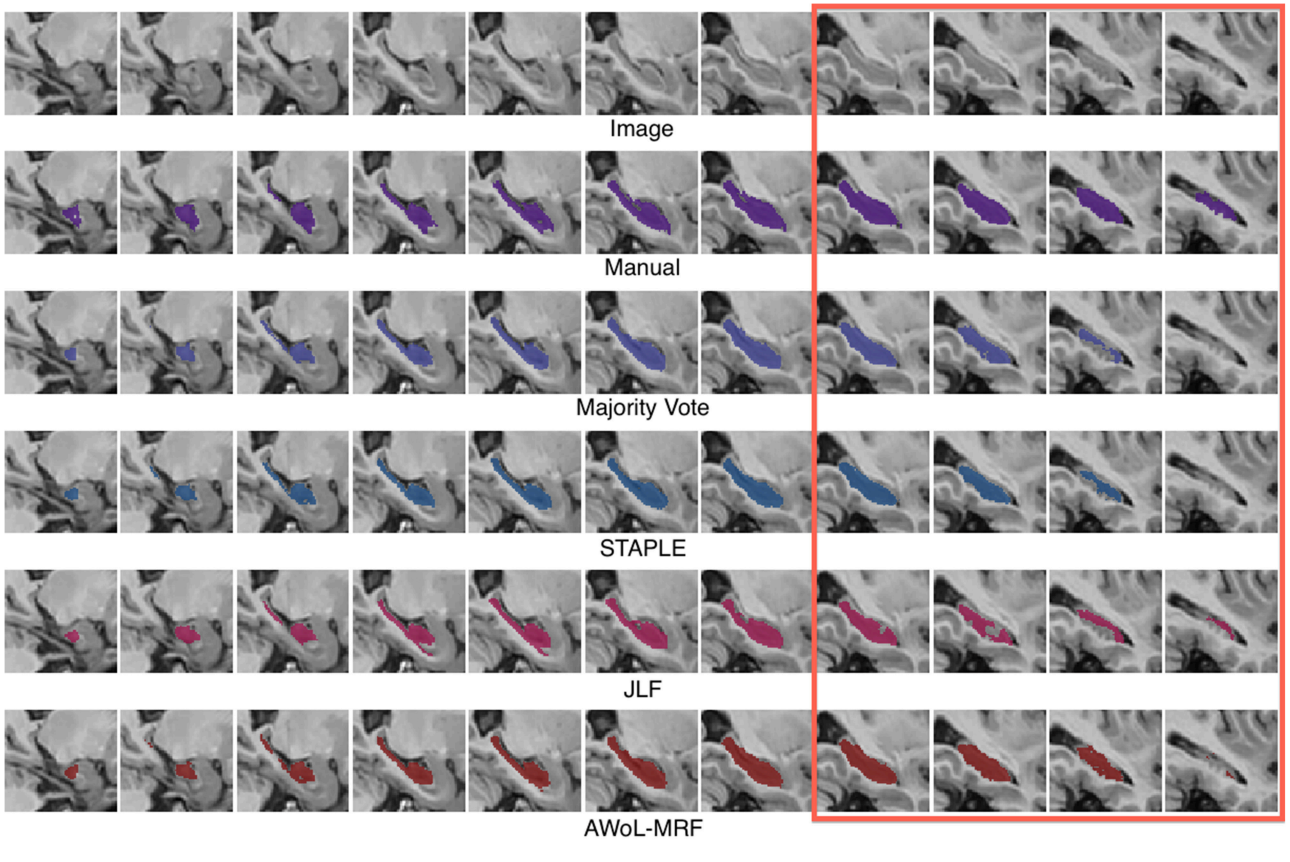

FIGURE 5 | Experiment I Qualitative Analysis: Comparison of manual vs. automatic segmentation methods. The red rectangle illustrates a section where the superiority of the AWoL-MRF approach is particularly apparent. The segmentations are performed using 3 atlases, and the Dice scores are as follows: Majority Vote: 0.806 STAPLE: 0.833 JLF: 0.804 AWoL-MRF: 0.854. The segmentation of the left hippocampus is shown in sagittal view.

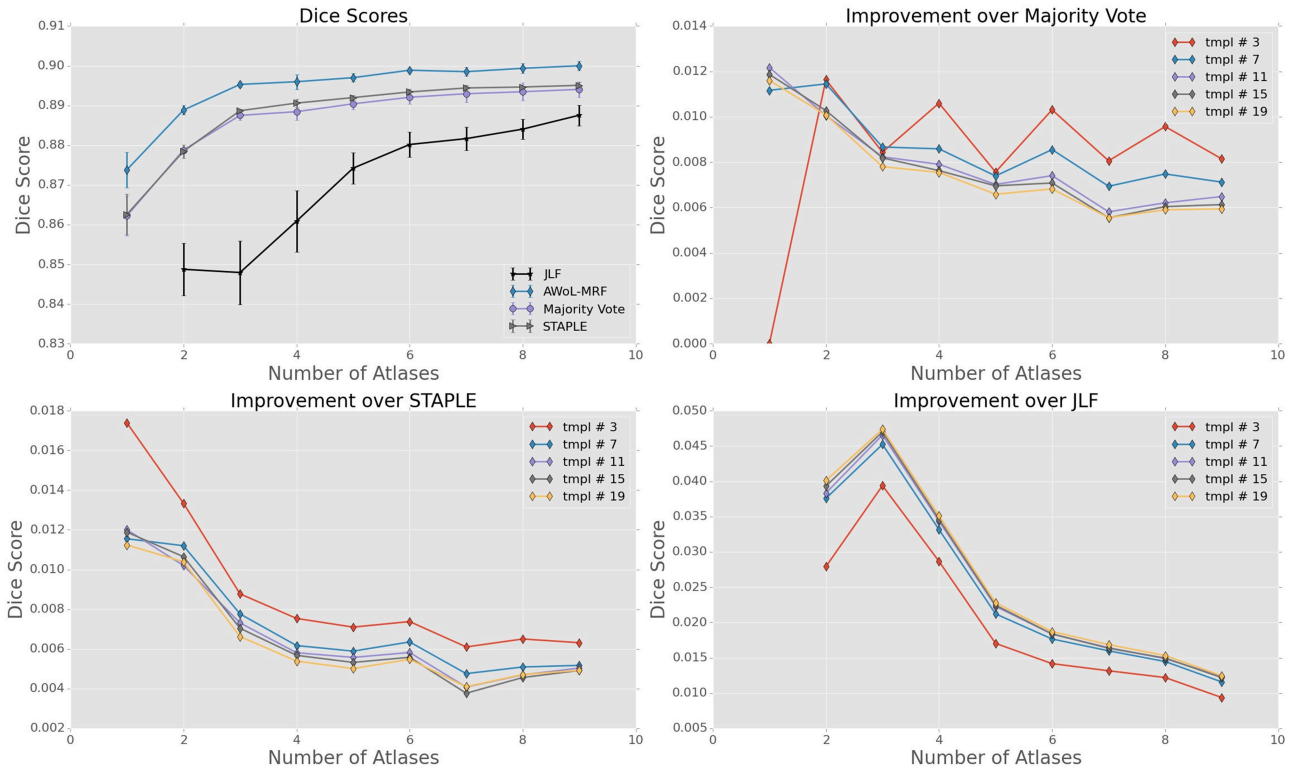

FIGURE 6 | Experiment II DSC: All results show the average performance values of left and right hippocampi over three-fold validation. The top-left subplot shows mean DSC score performance of all the methods. Remaining subplots show the mean DSC score improvement over compared methods for different number of templates (bootstrapping parameter of MAGeT-Brain).

with an increasing number of atlases and a smaller number of templates. In addition to a smaller atlas library requirement, the ability to reduce the bias introduced by the majority vote technique is also observed in this experiment.

DSC distribution comparisons for four sample configurations (number of atlases = 3, 5, 7, 9; number of templates = 11) are shown in **Figure 7**. These plots reveal that AWoL-MRF provides statistically significant improvement over all other methods regardless of the size of the atlas library. Similar to accuracy gains, the variance of the Dice score distribution is also smaller compared to ADNI experiment.

The Bland-Altman plots (see **Figure 8**) show that both AWoL-MRF and majority vote exhibit the smallest mean proportional bias. In comparison, STAPLE and JLF show strong biases characterizing considerable overestimation (negative bias) and underestimation (positive bias) of hippocampal volume across the cohort, respectively. Quantitatively, AWoL-MRF still outperforms the other three methods, as evident from the smaller line-slope and tighter limits of agreement.

Similar to the ADNI experiment, qualitative improvement is seen at the surface regions of the hippocampus (see **Figure 9**).

Experiment III: Preterm Neonatal Cohort Validation

The mean Dice score of AWoL-MRF maximizes at 0.807, with 9 atlases and 19 templates. Similar to the first two experiments, the AWoL-MRF consistently outperforms the majority vote (0.775), STAPLE (0.775), and JLF (0.771) methods by a large amount. More improvement is seen with fewer atlases when compared to JLF, as AWoL-MRF surpasses the mean Dice score of 0.800 with only four atlases (see **Figure 10**). The improvement diminishes as the number of atlases increases the number of templates decreases. Also, due to the single fold experimental design for this dataset, higher performance variability is observed especially with a smaller number of templates.

DSC distribution comparisons for four sample configurations (number of atlases = 3, 5, 7, 9; number of templates = 11) are shown in **Figure 11**. These plots reveal that AWoL-MRF provides statistically significant improvement over all other methods regardless of the size of the atlas library.

The Bland-Altman plots show that both AWoL-MRF and JLF can estimate hippocampal volume with an extremely small proportional bias (see **Figure 12**). Compared to the ADNI and FEP datasets, the magnitude of the bias is significantly lower, with AWoL-MRF producing the best result. In comparison, majority vote consistently underestimates and STAPLE consistently overestimates hippocampal volumes across the cohort.

Similar to the previous two experiments, qualitative improvement is seen at the surface regions of the hippocampus (see **Figure 13**). Note that the intensity values for hippocampus are reversed due to incomplete myelination.

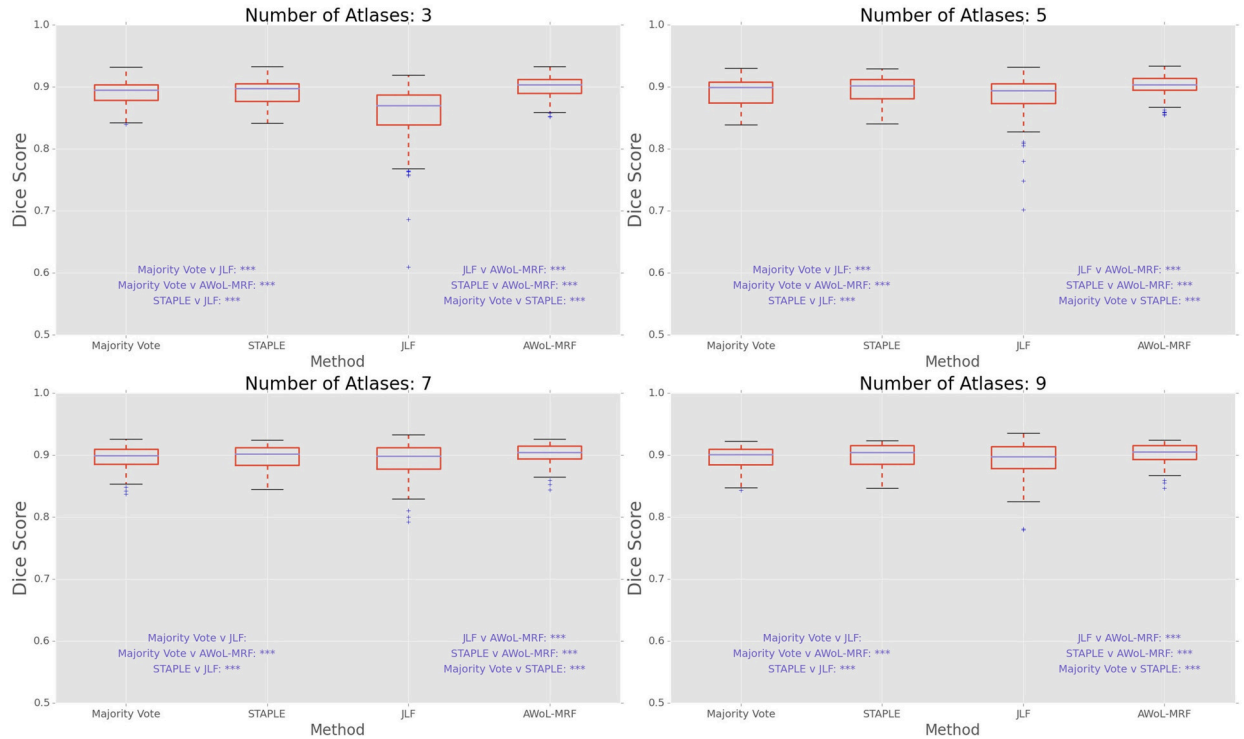

FIGURE 7 | Experiment II DSC: Statistical comparison of the performance of all methods for different atlas library sizes. The statistical significance is reported for pairwise comparisons (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

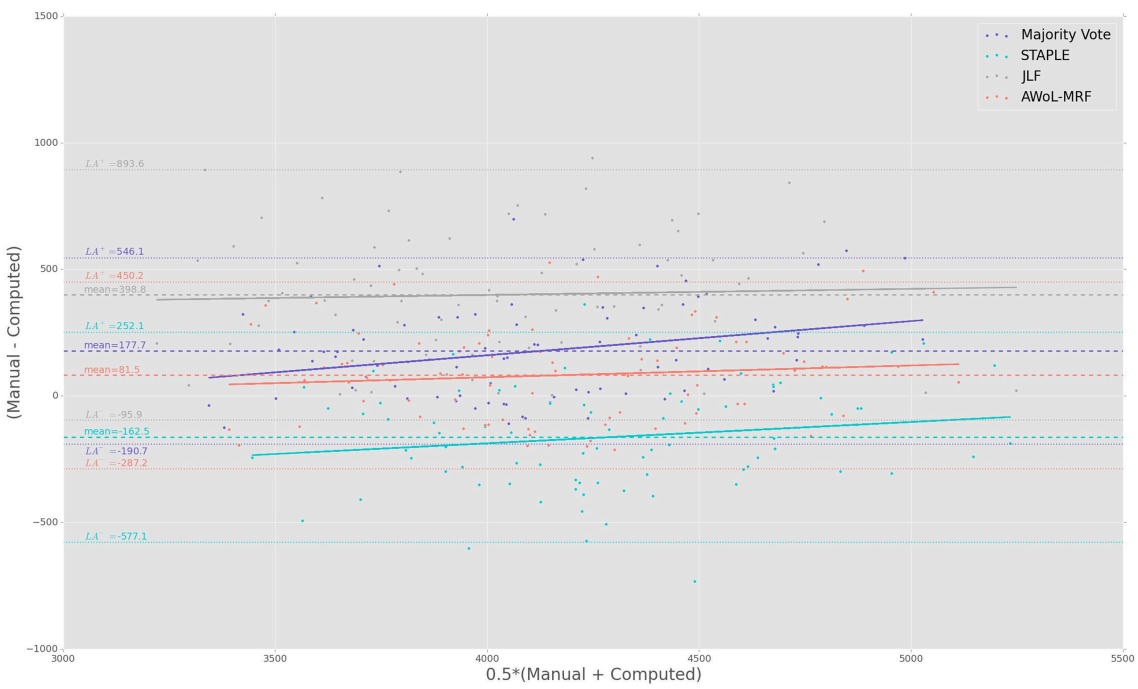

FIGURE 8 | Experiment II Bland-Altman Analysis: Comparison between computed and manual volumes (in mm^3) for single parameter configuration of 9 atlases and 19 templates. The overall mean difference in volume, and limits of agreement (LA+/LA−: 1.96SD) are shown by dashed horizontal lines. Linear fit lines are shown for each method. Note that the points above the mean difference indicate underestimation of the volume with respect to the manual volume, and vice versa.

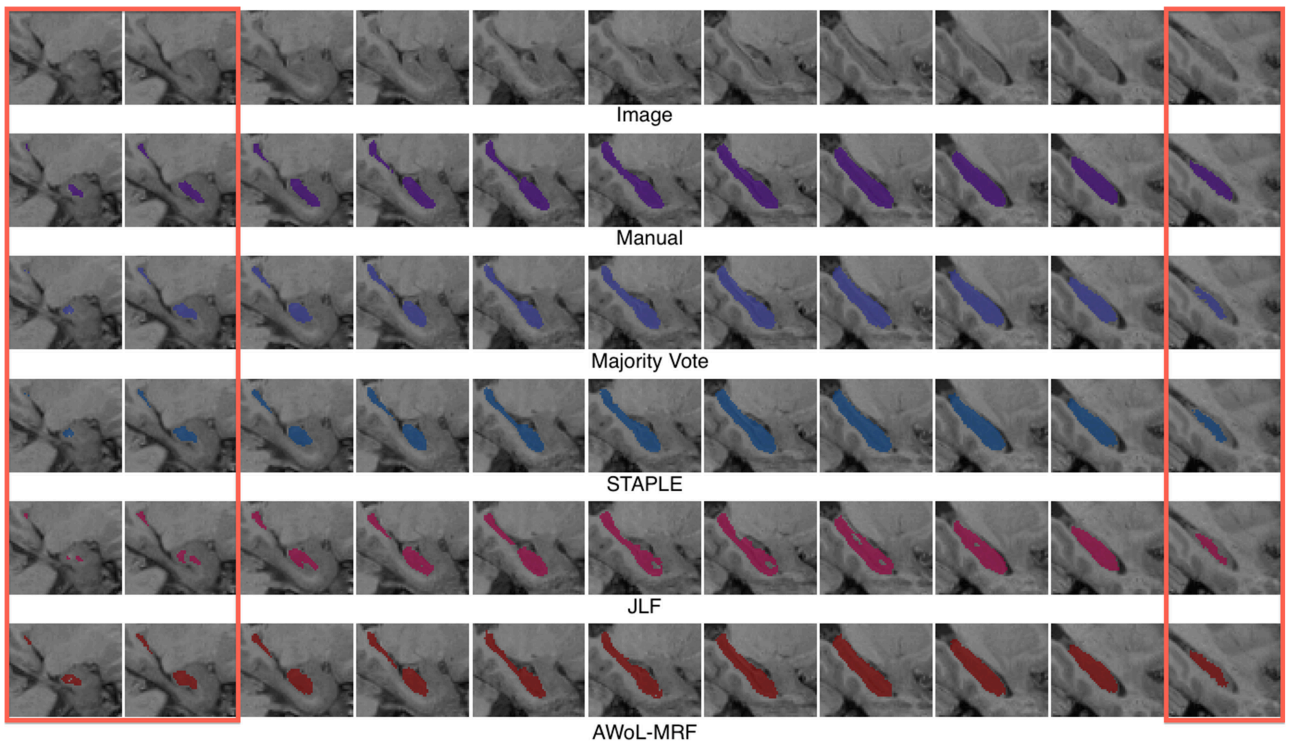

FIGURE 9 | Experiment II Qualitative Analysis: Comparison of manual vs. automatic segmentation methods. The red rectangle illustrates a section where the superiority of the AWoL-MRF approach is particularly apparent. The segmentations are performed using 3 atlases, and the Dice scores are as follows: majority vote: 0.875, STAPLE: 0.878, JLF: 0.856 AWoL-MRF: 0.891. The segmentation of the right hippocampus is shown in sagittal view.

Experiment IV: Hippocampal Volumetry

Group Comparisons between CN, MCI, and AD

As seen from **Figure 14A**, mean volume decreases with the severity of the disease for all methods. The volumetric statistics are summarized in **Table 3**. Based on Cohen's d metric as a measure of effect size, we see the largest separation between "CN vs. AD" diagnostic categories, followed by "CN vs. MCI" categories, and lastly between "MCI vs. AD" categories. The results show that the effect sizes are most pronounced in AWoL-MRF and JLF in all pairwise comparisons. All four methods show strong volumetric differences ($p < 0.001$ or $p < 0.01$) between "CN vs. AD" categories followed by "CN vs. MCI," which show relatively weaker differences. JLF also shows volumetric differences between "MCI vs. AD" categories with a much weaker significance level ($p < 0.05$) compared to the other two pairwise comparisons. In the linear model analysis, all four methods show significant differences ($p < 0.001$ or $p < 0.01$) only between "CN vs. AD" and "CN vs. MCI" comparisons.

Group Comparisons between MCI-Converters and MCI-Stable Cohorts

Figure 14B shows that the MCI-converters have relatively smaller volumes compared to MCI-stable group. The volumetric statistics are summarized in **Table 3**. AWoL-MRF shows statistically significant ($p < 0.05$) volumetric differences between these two groups, with strongest effect size based on Cohen's d metric. In the comparison using a linear model, AWoL-MRF continues to show significant volumetric differences ($p < 0.05$) between these two groups.

Parameter Selection

We studied the impact of parameter selection on the performance of AWoL-MRF with joint consideration of the segmentation accuracy and computational cost. The four parameters that need to be chosen *a priori* are: confidence thresholds (L_T^0 *and* L_T^1), patch-length (L_{patch}), mixing-ratio $(S_H/S_L)_{patch}$, and the β parameter of the MRF model. Recall that the Gaussian distribution parameters in the MRF are estimated for each patch automatically using the S_H nodes in the given patch.

First, the confidence threshold parameters are heuristically derived from the voting distribution. As mentioned before, both ${L_T}^0$ and ${L_T}^1$ values need to be greater than 0.5 to produce non-empty low-confidence voxel set. Based on the assumptions that the high-confidence region (S_H) comprises more structural voxels ($L(x_i) = 1$) than the total number of voxels in the low-confidence region (S_L), we define a following metric:

$$\rho = \frac{|S_L|}{|L(x_i) = 1|} \tag{12}$$

Then we choose confidence thresholds (L_T^0 and L_T^1), which fall in the parameter space bounded by $\rho \in (0.5, 1)$. **Figure 15A** shows an example of these bound values—computed for the ADNI

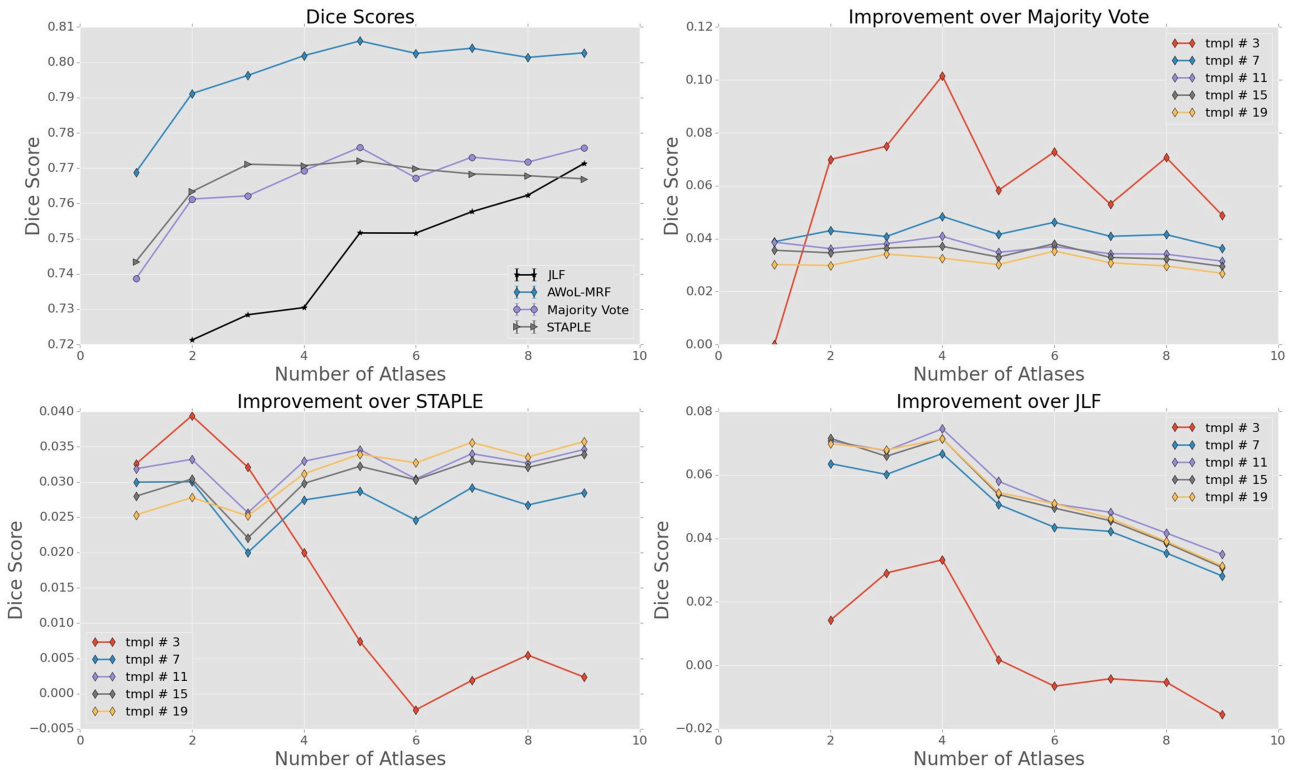

FIGURE 10 | Experiment III DSC: Preterm Neonate Cohort Validation: All results show the average performance values of left and right hippocampi over three-fold validation. The top-left subplot shows mean DSC score performance of all the methods. Remaining subplots show the mean DSC score improvement over compared methods for different number of templates (bootstrapping parameter of MAGeT-Brain).

dataset in Experiment I (left hippocampus). Note that the larger threshold values imply larger S_L region, and consequently higher computational time. Based on this heuristic, we chose ${L_T}^0 = 0.8$ and ${L_T}^1 = 0.6$ for experiments I, II, and IV; and ${L_T}^0 = {L_T}^1 = 0.7$ for the experiment III.

As described in Section Validation Experiments, the patch-length and the mixing ratio parameters are interrelated and directly affect the coverage of S_L region. From a performance perspective, these have higher impact on the computational time than the segmentation accuracy (see **Figures 15B,C**). Higher L_{patch} implies larger MRF model on the sub-volume and therefore requires higher computational time. Conversely, smaller patches would reduce the computational time; but would run a risk of insufficient coverage of S_L region and consequently offer poor accuracy improvement. The third parameter choice of mixing ratio affects the total number of seeds/patches for a given image. A higher ratio necessitates a search for S_L nodes surrounded with a large number of S_H nodes, which reduces the total number of patches as well as the computational time. Based on the accuracy vs. computational cost trade-off analysis with respect to these parameter choices, we selected a patch-length of 11 voxels and a minimum mixing ratio of 0.0075 which translates into seed nodes surrounded by a minimum of 10 S_H nodes in the 26-node neighborhood, for all validation experiments.

Lastly, the β parameter of MRF model controls the homogeneity of the segmentation. It is dependent on the image intensity distribution and the structural properties of the anatomical structure. The large value of β results in more homogeneous regions giving a smoothed appearance to a structure. We selected $\beta = -0.2$ based on the results of training phase where we split the atlas pool into two groups and used one set to segment the other.

DISCUSSION AND CONCLUSIONS

In this work, we presented a novel label fusion method that can be incorporated into any multi-atlas segmentation pipeline for improved accuracy and robustness. We validated the performance of AWoL-MRF over three independent datasets spanning a wide range of demographics and anatomical variations. In Experiment I, we validated AWoL-MRF on an Alzheimer's disease cohort ($N = 60$) with median age of 75. In Experiment II, validation was performed on first episode of psychosis cohort ($N = 81$), with median age of 23. In Experiment III, we applied AWoL-MRF to a unique cohort ($N = 22 \times 2$) comprising preterm neonates scanned in the first weeks after birth and again at term-equivalent age with distinctly different brain sizes and MR scan characteristics from our first two datasets. In all of these exceptionally

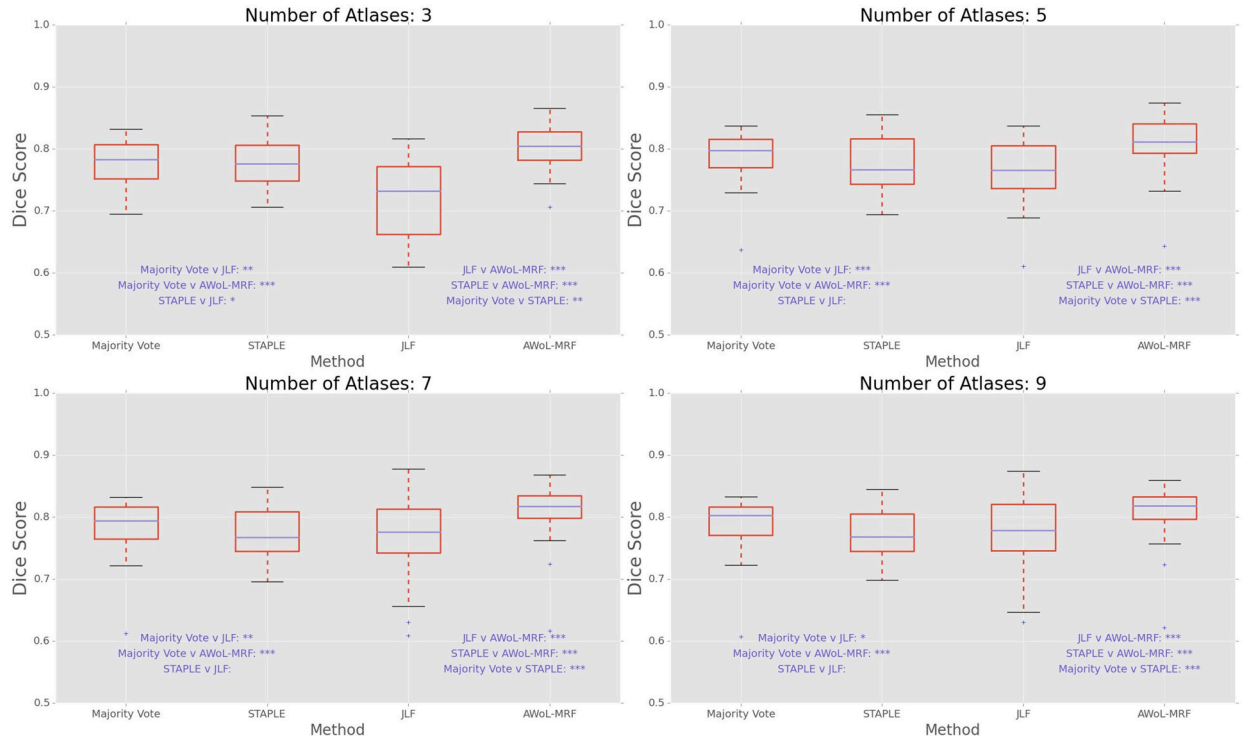

FIGURE 11 | Experiment III DSC: Statistical comparison of the performance of all methods for different atlas library sizes. The statistical significance is reported for pairwise comparisons (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

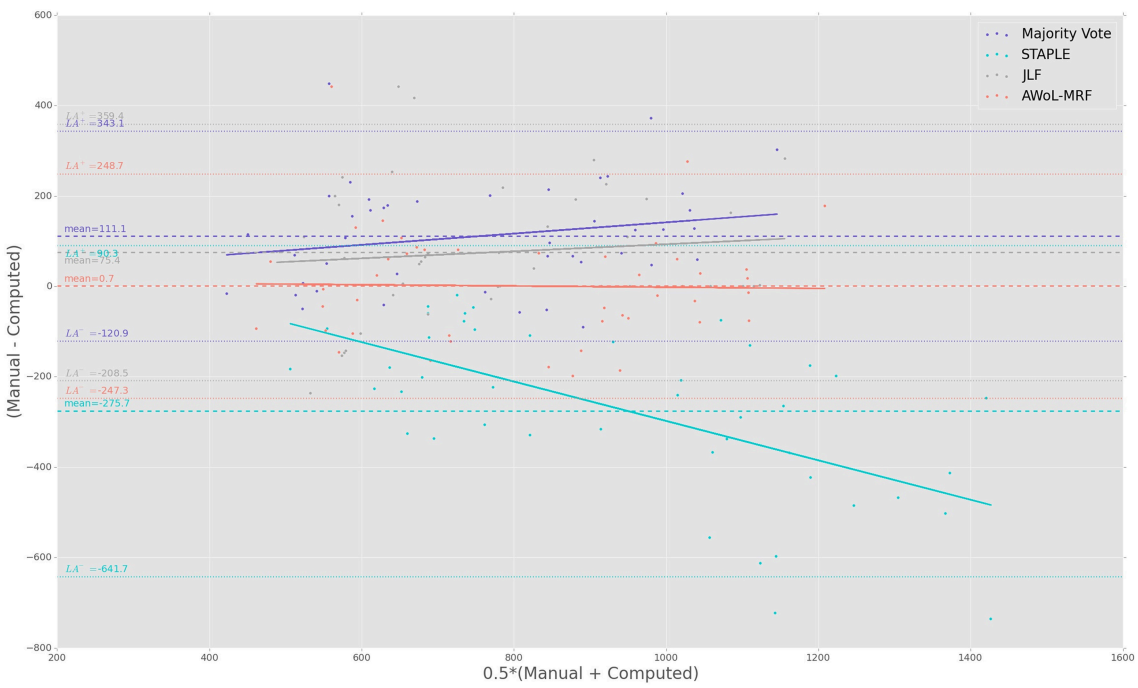

FIGURE 12 | Experiment III Bland-Altman analysis: Comparison between computed and manual volumes (in mm^3) for single parameter configuration of 9 atlases and 19 templates. The overall mean difference in volume, and limits of agreement (LA+/LA−: 1.96 SD) are shown by dashed horizontal lines. Linear fit lines are shown for each method. Note that the points above the mean difference indicate underestimation of the volume with respect to the manual volume, and vice versa.

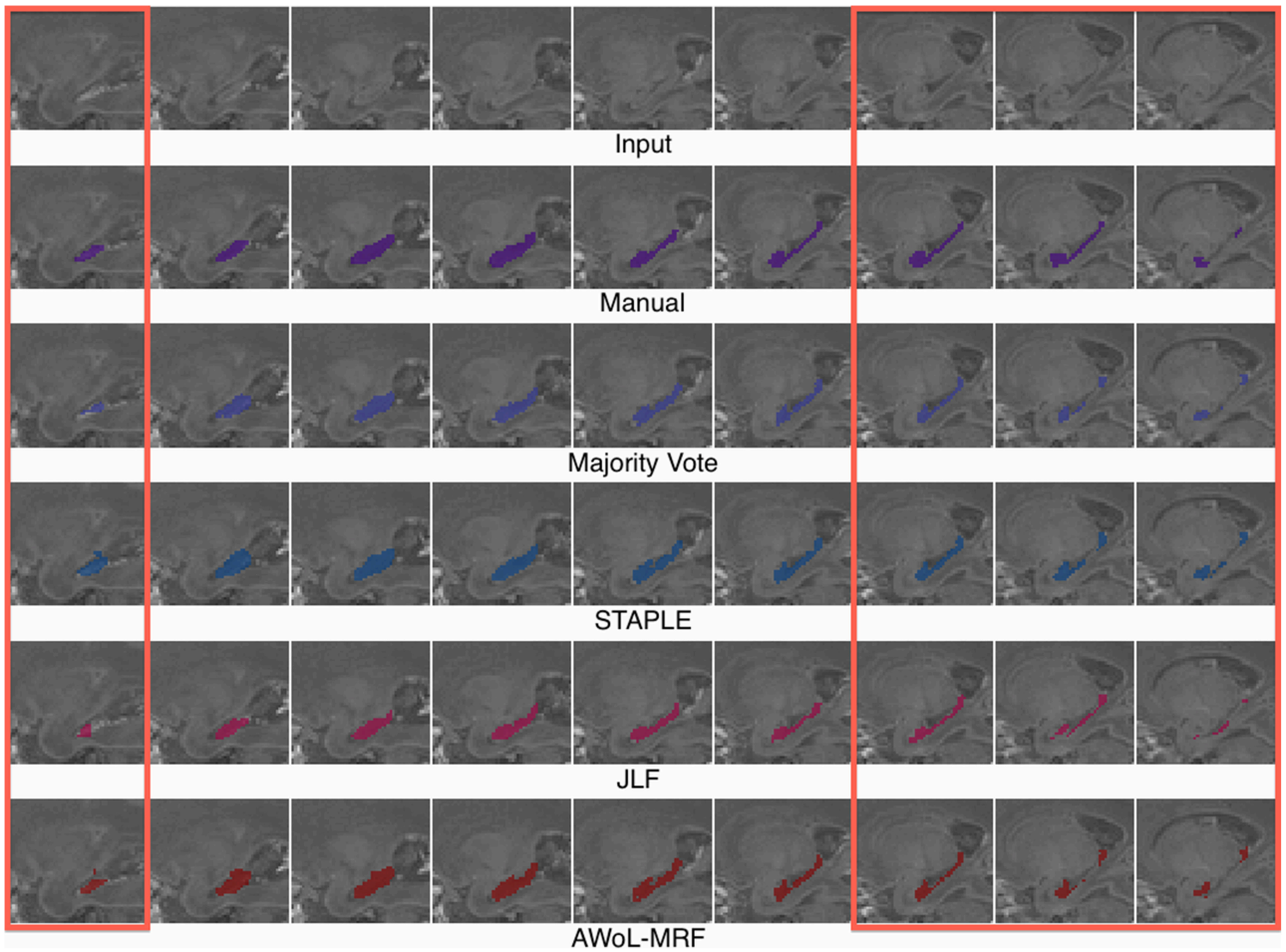

FIGURE 13 | Experiment III Qualitative Analysis: Comparison of manual vs. automatic segmentation methods. The red rectangles illustrate sections where the superiority of the AWoL-MRF approach is particularly apparent. The segmentations are performed using three atlases, and the Dice scores are as follows: majority vote: 0.748, STAPLE: 0.760, JLF: 0.746, AWoL: 0.807. The segmentation of the left hippocampus is shown in sagittal view. Note that for this particular dataset, the brain structures are mostly unmyelinated causing a reversal of the intensity values for the hippocampal structure—as shown in the top row.

heterogeneous subject groups, AWoL-MRF provided superior segmentation results compared to all three competing methods: majority vote, STAPLE, and JLF, based on DSC metric as well as proportional bias measurements. In Experiment IV, we validated the diagnostic utility of AWoL-MRF by analyzing the standardized ADNI1: Complete Screening 1.5T dataset. We found significant volumetric differences between "CN vs. AD" and "CN vs. MCI" groups, as well as, "MCI-converters vs. MCI-stable" groups.

In the first three experiments, we see that AWoL-MRF offers superior performance with a remarkably small atlas library, a very desirable quality in a segmentation pipeline. AWoL-MRF provides mean DSC scores over 0.880 with only six atlases (Experiment I), 0.890 with only three atlases (Experiment 2), and 0.800 with only four atlases (Experiment III) compared to other methods, which require larger atlas libraries to deliver similar performance. This is an important benefit as it reduces the resource expenditure on the manual delineation of MR images and speeds up the analysis pipelines. From a robustness perspective, we notice a reduction in the two types of biases. First, AWoL-MRF mitigates the issue of degenerating accuracy caused by the vote-ties with a small, even number of atlases. Then, more importantly, we see a consistent reduction of proportional bias, as evident by the Bland-Altman analysis.

There are several novel features that distinguish AWoL-MRF from other label fusion algorithms, particularly due to its methodological similarities to manual segmentation procedures. For instance, a manual rater estimates the voxel intensity distribution conditioned on a label class purely from the neighborhood of the target image itself and not from the atlas library. AWoL-MRF translates this into estimating the intensity distributions based on the statistics collected from the high-confidence voxels in a given localized patch in the target image. Thus, one of the key differences between AWoL-MRF and the existing multi-atlas label fusion techniques includes the decoupling from the atlas library in the post-registration stages.

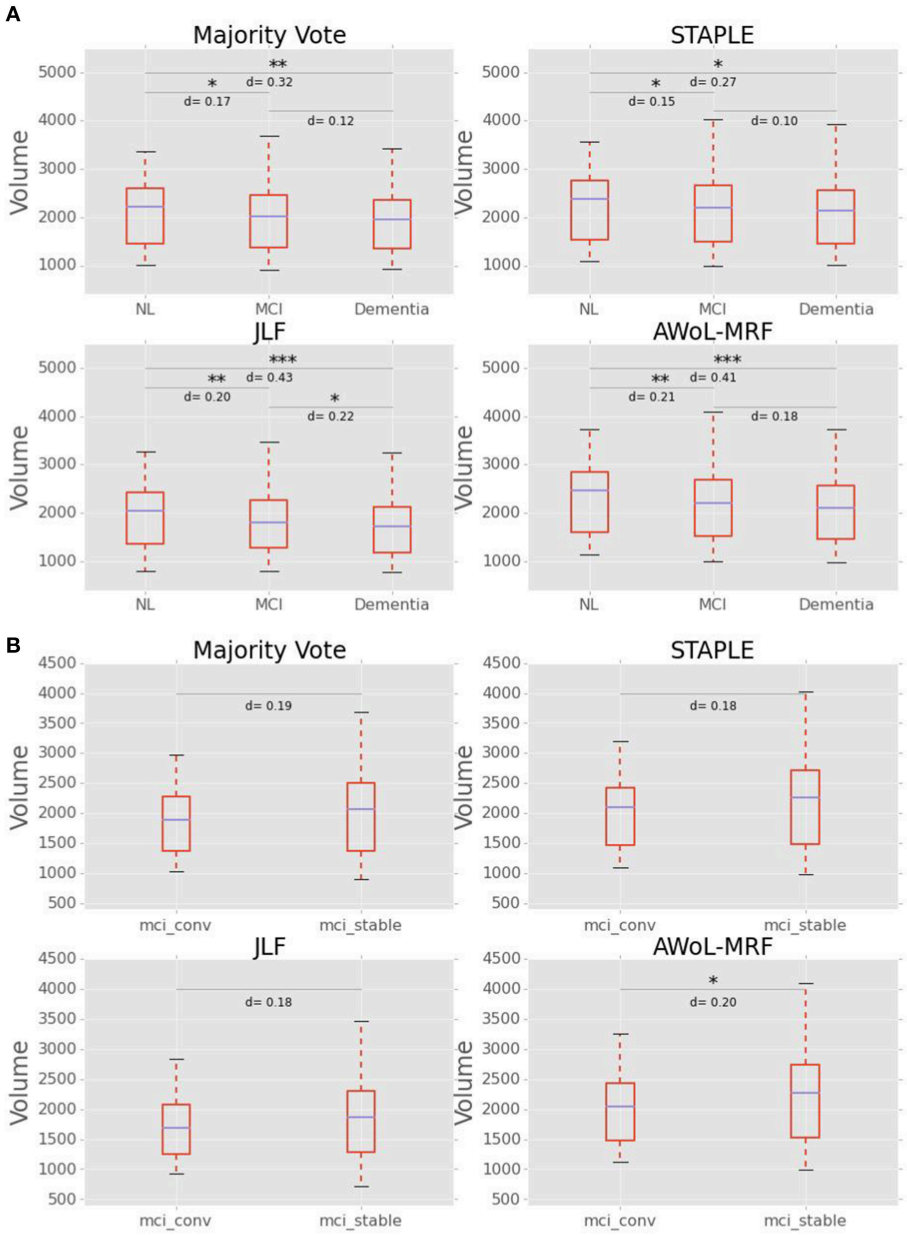

FIGURE 14 | (A) Hippocampal volume (in mm^3) vs. diagnoses (NL vs. MCI vs. AD). Cohen's d scores (effect size) and statistical significance is reported for pairwise comparisons between diagnostic groups. **(B)** Hippocampal Volume (in mm^3) vs. MCI subgroups (converters vs. stable). Cohen's d scores (effect size) and statistical significance is reported for pairwise comparisons between groups.

Once we obtain the initial label-vote distribution, we completely rely on the intensity profile of the target image and avoid any computationally expensive pairwise similarity comparisons with the atlas-library. Additionally, even though we use a commonly used MRF approach to model spatial dependencies, the novel spanning-tree based inference technique that attempts to mimic the delineation process of a manual rater differentiates AWoL-MRF from traditional iterative optimization techniques such as iterative conditional modes or Expectation-Maximization (Van Leemput et al., 2003; Warfield et al., 2004).

The key benefits of the AWoL-MRF implementation are two-fold. First we offer state-of-the-art performance using a small atlas library (<10), whereas most existing segmentation pipelines typically make use of large atlas libraries comprising 30–80 manually segmented image volumes (Pruessner et al., 2000; Heckemann et al., 2006) that require specialized knowledge and experience to generate. Secondly, from a computational perspective, AWoL-MRF mitigates several expensive operations common among many multi-atlas label fusion methods. First, by eliminating the need for pairwise similarity metric estimation,

TABLE 3 | Hippocampal Volumetry Statistics of ADNI1: Complete Screening 1.5T dataset per diagnosis [subjects with Alzheimer's disease (AD), subjects with mild cognitive impairment (MCI), healthy subjects/cognitively normal (CN), as well as, MCI-converters and MCI-stable subgroups].

Volumetric statistics: CN vs. MCI vs. AD Comparisons

	CN			MCI			AD		
Method	**Mean**	**Std. dev**	**Range**	**Mean**	**Std. dev**	**Range**	**Mean**	**Std. dev**	**Range**
Majority Vote	2084.7	615.3	[1010.0, 3364.0]	1960.5	599	[901.5, 3685.5]	1897.2	582.3	[940.0, 3422.0]
STAPLE	2236.6	659	[1097.5, 3557.0]	2124.2	649.3	[988.0, 4026.5]	2068.2	655.4	[1008.5, 3936.0]
JLF	1943.6	593.5	[796.0, 3280.5]	1803.3	572.6	[807.0, 3463.0]	1697.3	551.6	[782.0, 3242.5]
AWoL-MRF	2312.9	676.3	[1133.5, 3736.5]	2147.5	652	[991.0, 4094.5]	2047.7	631.3	[982.5, 3731.0]

	Cohen's d			Linear Model		
	CN vs. MCI	**CN vs. AD**	**MCI vs. AD**	**CN vs. MCI**	**CN vs. AD**	**MCI vs. AD**
Majority Vote	0.1727	0.3194	0.123	−3.875***	−3.662***	−0.402
STAPLE	0.1463	0.2688	0.1005	−3.424***	−3.001**	−0.101
JLF	0.202	0.4343	0.2155	−4.195***	−4.884***	−1.451
AWoL-MRF	0.2092	0.4111	0.1783	−4.424***	−4.657***	−0.987

Volumetric Statistics: MCI-converters vs. MCI-stable Comparisons

	MCI-converters			MCI-stable		
Method	**Mean**	**Std. dev**	**Range**	**Mean**	**Std. dev**	**Range**
Majority Vote	1846.2	489.6	[1036.0, 2980.0]	1995.7	619.3	[901.5, 3685.5]
STAPLE	2000.7	542.8	[1093.0, 3205.5]	2163.6	668.0	[988.0, 4026.5]
JLF	1686.8	483.9	[932.5, 2831.0]	1842.2	586.3	[807.0, 3463.0]
AWoL-MRF	2007.4	534.6	[1115.0, 3251.0]	2186.6	672.6	[991.0, 4094.5]

	MCI-converters vs. MCI-stable: Cohen's d	MCI-converters vs. MCI-stable: Linear Model
Majority Vote	0.185	−1.708
STAPLE	0.181	−1.616
JLF	0.192	−1.844
AWoL-MRF	0.204	−1.965*

Effect sizes of pairwise differences between groups are based on Cohen's d metric. All t-values and significance levels from a linear model comprising "Age," "Sex," and "total-brain-volume" as covariates. ($p < 0.05$, **$p < 0.01$,***$p < 0.001$).*

we avoid computationally expensive registration operations that increase rapidly with the size of the atlas library. Furthermore, several extensions based on patch-based comparisons between an atlas library and a target image make use of a variant of a local search algorithm or a supervised learning approach (Coupé et al., 2011; Rousseau et al., 2011; Wang et al., 2013; Hao et al., 2014; Wu et al., 2014). For instance, Coupé et al. (2011) uses a non-local means approach to carry out label transfer based on multiple patch comparisons; Hao et al. (2014) uses a supervised machine-learning method to train a classifier using similar patches from an atlas library. Computationally these patch-based approaches, especially the implementations that incorporate non-local means, are expensive (Wang et al., 2013) and require a considerable number of labeled images (Hao et al., 2014; Wu et al., 2014). Moreover, compared to the single unified MRF models, the localized MRF model reduces the computational complexity while maintaining the spatial homogeneity constraints in the given neighborhood. It also allows the label fusion step to capture local characteristics of the image based on high-confidence regions without requiring the iterative parameter estimation and inference methods such as EM. Lastly, other confidence based label fusion methods such as Zhang et al. (2011) utilize local image appearance based metric estimated from forward and backward matching procedures involving computationally expensive k-NN search. In contrast, AWoL-MRF simply uses label-vote distribution at each voxel to compute the confidence estimate.

We believe that the performance improvements provided by AWoL-MRF can be explained by two major factors. First, we argue that the utilization of intensity values and local neighborhood constraints act as regularizers, which helps avoid over-fitting to the hippocampal model represented by the atlas library. Both majority vote and STAPLE do not consider intensity values in their label fusion stage and thus are more likely to ignore minute variations near the surface areas of the structure, which are not well represented within the atlas library.

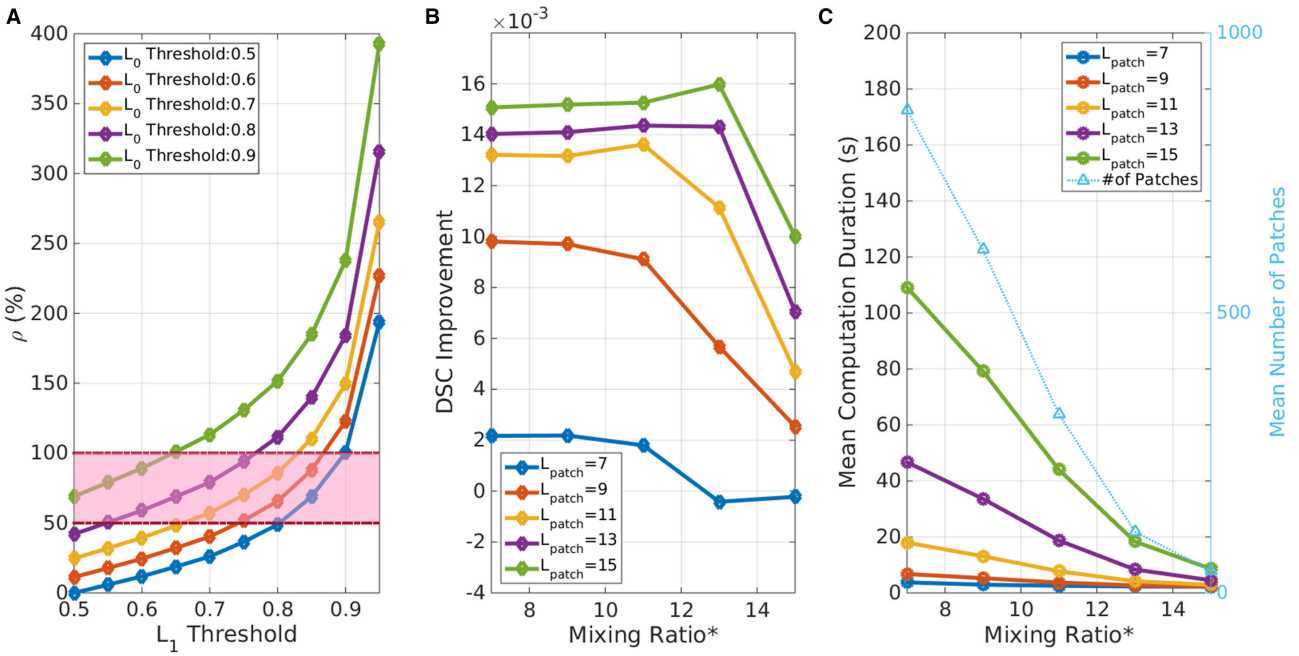

FIGURE 15 | Parameter selection. (A) Effect of confidence threshold values on image partitioning. ρ represents the ratio of low-confidence voxels over high-confidence structural voxels. The highlighted region denotes the heuristically "good" region for threshold selection. **(B)** Effect of mixing ratio and L_{patch} on DSC performance. Mixing Ratio* is the minimum required number of S_H nodes in the 26-node neighborhood for a given seed voxel. Note that with a larger L_{patch} performance improves. Whereas, with a smaller L_{patch} or a higher mixing ratio* the performance worsens due to poor coverage over S_L region. **(C)** Effect of mixing ratio and L_{patch} on computational cost. The light blue line shows the number of patches for given configuration as a reference. Note that the computation time increases exponentially with a higher L_{patch} and a smaller mixing ratio* (Note: mixing ratio* represents the equivalent number of minimum S_H node requirement in the 26-node neighborhood for the seed node selections).

JLF, which does take intensity information into account and implements a patch-based approach, tends to perform better than majority vote and STAPLE with a relatively higher number of atlases: >4 in Experiment I and >6 in Experiment III. Therefore, we speculate that JLF is more likely to deliver superior performance in cases with larger atlas library availability, which again comes with the cost of generating manual segmentations. Second, the spanning tree based inference method tries to mimic the manual delineation process by starting with regions with strong neighborhood label information and moving progressively toward more uncertain areas. Compared to iterative methods (e.g., EM) or graph-cut based approaches (Wolz et al., 2009; Lötjönen et al., 2010) the sequential inference process may not be optimal in a theoretical sense; i.e., spanning-tree does not guarantee the global minimum for the MRF energy function. Nevertheless, we argue that the procedural similarity between the automatic and manual labeling process provides more accurate results, since the ground truth is defined by the latter.

Additionally, decoupling of label fusion process from similarity comparisons with the atlas library allows AWoL-MRF to utilize bootstrapping techniques that augment the pool of candidate labels as used by the baseline segmentation pipeline (MAGeT-Brain) in this work (Pipitone et al., 2014). Use of such techniques is not trivial with approaches using intensity information from the atlas library.

From a diagnostics perspective, the volumetric assessment of all four methods shows significant differences ($p < 0.001$ or $p < 0.01$) between "CN vs. AD" and "CN vs. MCI" groups. Consistent with the Bland-Altman analysis (see **Figure 4**), JLF and majority vote underestimate the volume compared to AWoL-MRF and STAPLE across all diagnostic categories. Even though the direct volumetric comparisons based on JLF yield significant differences ($p < 0.05$) between "MCI vs. AD" groups, these differences vanish in the linear model that includes "age," "sex," and "total-brain-volume" as covariates. These findings are consistent with a variety of studies (Mouiha and Duchesne, 2011; Sabuncu et al., 2011; La Joie et al., 2013) highlighting the heterogeneity in hippocampal volume within MCI subjects, which results in smaller differences between MCI and AD groups. This is particularly typical in the ADNI-1 cohort MCI subjects used in this analysis, which were recently re-classified under more progressed stages of MCI or late-MCI (Aisen et al., 2010). The volumetric comparison between MCI-converters and MCI-stable groups reveals that the subjects from latter group comprise relatively larger hippocampal volumes at the screening time-point. These findings are consistent with a previous study conducted on the ADNI baseline cohort (Risacher et al., 2009). We also find that these differences remain statistically significant in the linear model that includes "age," "sex," and "total-brain-volume" as covariates.

TABLE 4 | Summary of automated segmentation methods of the hippocampus.

No of atlases	DSCmean	Reference study	Validation	Dataset (ground-truth)
9	0.881	AWoL-MRF	Three-Fold MCCV on 60 subjects	ADNI (Pruessner)
9	0.897	AWoL-MRF	Three-Fold MCCV on 81 subjects	FEP (Pruessner)
9	0.807	AWoL-MRF	One-Fold MCCV on 44 subjects	3-step segmentation protocol[b]
9	0.869	Pipitone et al., 2014	10-Fold MCCV on 60 subjects	ADNI (Pruessner)
9	0.892	Pipitone et al., 2014	Five-Fold MCCV on 81 subjects	FEP subjects
9	0.79	Guo et al., 2015	One-Fold MCCV on 44 subjects	3-step segmentation protocol[b]
30	0.82	Heckemann et al., 2006	LOOCV	Controls
21	0.862	Morra et al., 2008	LOOCV	ADNI (SNT)
55	0.86	Barnes et al., 2007	LOOCV	Controls and AD
275	0.835	Aljabar et al., 2009	LOOCV	Controls
80	0.89	Collins and Pruessner, 2010	LOOCV	Controls
30	0.885	Lötjönen et al., 2010	Segmentation of 60 subjects	ADNI (SNT)
55	0.89	Leung et al., 2010	Segmentation of 30 subjects	ADNI (SNT)
30	0.848	Wolz et al., 2010	Segmentation of 182 subjects	ADNI (SNT)
16	0.861	Coupé et al., 2011[a]	LOOCV	ADNI (Pruessner)
20	0.897 (L–HC)	Wang et al., 2012	10-Fold MCCV on 20 of 139 subjects	Landmark based semi-automatic segmentation + manual correction
	0.888 (R–HC)			
15	0.862(L–HC)	Wang et al., 2013[c]	Segmentation of 20 subjects (JLF)	BrainCOLOR
	0.861(R–HC)			
15	0.872(L–HC)	Wang et al., 2013[c]	Segmentation of 20 subjects (With corrective learning)	BrainCOLOR
	0.871(R–HC)			
9	0.841	Pipitone et al., 2014	10-Fold MCCV on 69 subjects	ADNI (SNT)

AD, Alzheimer's Disease; MCI, Mild Cognitive Impairment; CN, Cognitively Normal; FEP, First Episode of Psychosis; LOOCV, Leave-one-out cross-validation; MCCV, Monte Carlo cross-validation; SNT, Surgical Medtronic Navigation Technologies semi-automated labels; L–HC, Left hippocampus; R–HC, Right hippocampus.

[a] *AD: 0.838, MCI: n/a, CN: 0.883.*

[b] *See Guo et al. (2015) for manual segmentation protocol details.*

[c] *The method were applied in the 2012 MICCAI Multi-Atlas Labeling Challenge.*

A direct comparison against other methods from the current literature is difficult due to differences in the choices for gold standards, evaluation metrics, and hyper-parameter configuration, among other variables. Nevertheless, **Table 4** shows a brief survey of several segmentation studies. Note that many of these studies have relied on SNT—labels provided by ADNI—for the ground-truth (manual) segmentations. A performance comparison of the baseline method based on SNT labels is discussed in our previous work (Pipitone et al., 2014), where we noticed several shortcomings of the SNT protocol (Winterburn et al., 2013; Pipitone et al., 2014), and therefore we have evaluated the presented method against the manual label based on the Pruessner protocol (Pruessner et al., 2000). Moreover, we would like to emphasize that the quality and consistency of an anatomical gold-standard is an important consideration when assessing the accuracy of an automated segmentation methodology. The Pruessner protocol used in this work reports reliabilities (Dice kappa) of 0.94 for both intra- and inter-rater over 40 subjects. Other methods (Winston et al., 2013) report the intra- and inter-rater reliability for manual segmentations to be 0.891 and between 0.82 and 0.84, respectively, using 18 subjects. Thus, depending on the choice of gold-standard for assessment of automatic methods, the expected upper bound for performance measures is likely to be different. However, the inter-rater reliabilities in Winston et al. (2013) underscore the need for a reliable segmentation methodology that is not subject to the same confounds as a manual rater in terms of consistency across raters. Our method, like many others, will always provide the same output for the automated segmentation given the same input and parameter configuration.

Despite the differences in the experimental designs, comparisons with the other methods show that AWoL-MRF delivers superior performance with a significantly smaller atlas library requirement. For ADNI cohort validation, barring the ground-truth label dissimilarities, methods presented by Leung et al. (2010), Lötjönen et al. (2010) have equivalent DSC scores; however, the atlas library sizes for these methods are 30 and 55, respectively. Moreover, Lötjönen et al. (2010) use atlas selection procedure that adds another computational step to their pipeline. It may be possible that using similar number of atlases would improve our automated segmentation procedure. However, this is unlikely given the plateau effect on the number of atlases used. Moreover, to the best of our knowledge, no other study has used three drastically different datasets spanning the entire human lifespan to validate the robustness of its method. Other recent approaches (Tong et al., 2013; Zikic et al., 2014)

make use of machine-learning based techniques also report similar performances. Specifically, Tong et al. (2013) make use of sparse coding and dictionary learning techniques that yield Dice scores of 0.864–0.879 depending on atlas library sizes (10–30), atlas selection, and offline training configurations. More recently, similar learning based approaches comprising sparse multimodal representations and random forests have been proposed for infant brain segmentation (Wang et al., 2014, 2015) for tissue-based classification. Nevertheless, the training phases of these methods are computationally involved requiring substantial number of atlases, making them more suitable in the context of large-scale or multimodal studies.

The computational cost of the algorithm implementation, as described in the previous section, depends on the parameter selection. From a theoretical perspective, MST transformation is the most expensive task in this method. The current implementation of MST uses Prim's algorithm with simple adjacency matrix graph representation, which requires $O(|V|^2)$ running time ($|V|$: number of uncertain voxels in the patch). However, this can be reduced down to $O(|E|log|V|)$ or $O(|E| + |V|log|V|)$ using a binary heap or Fibonacci heap data structures, respectively, ($|E|$: number of edges in the patch). The computational times for Experiment I with current implementation for different parameter configurations are shown in **Figure 15C**. The code was implemented in Matlab R2013b and run on a single CPU (Intel x86_64, 3.59 GHz). A direct computational time comparison with other methods is not practical due to hardware and software implementation differences. However, the non-iterative nature of AWoL-MRF provides considerably faster run times compared to EM based approaches, where the convergence of the algorithm is dependent on the agreement between candidate labels and can be highly variable (Van Leemput et al., 2003; Warfield et al., 2004).

In conclusion, AWoL-MRF attempts to mimic the behavior of a manual segmentation protocol in a multi-atlas segmentation framework. We validated its performance over three independent datasets comprising significantly different subject cohorts. Even though this work focuses on hippocampal segmentations, AWoL-MRF can be easily applied to other structures and scenarios with multiple label classes, which will be a part of future studies. Moreover as per the scope of this work, we only performed volumetric comparisons across groups in ADNI. While we do not perform homologous comparisons in FEP and preterm neonate cohorts, we believe the increased accuracy and precision of our method will allow us to better characterize the neuroanatomy of these groups in subsequent studies. Our validations indicate that the method delivers the state-of-the-art performance with a remarkably small library of manually labeled atlases, which motivates its use as a highly efficient label fusion method for rapid deployment of automatic segmentation pipelines.

AUTHOR CONTRIBUTIONS

NB: AWoL-MRF algorithm development, software implementation, and experimental evaluation. JP: Software implementation of MAGeT Brain (Baseline multi-atlas segmentation pipeline). JW: Expert rater—developed Winterburn manual segmentation protocol and corresponding atlases for Experiment 3. TG, ED, and SM: Data acquisition and preprocessing for Experiment III. AV: Co-superviser of NB and JP, ML: Data acquisition and preprocessing for Experiment II. JCP: Expert rater—developed Winterburn manual segmentation protocol and corresponding atlases for Experiment I, II. MC: Superviser of NB.

ACKNOWLEDGMENTS

NB receives support from the Alzheimer's Society. MC is funded by the Weston Brain Institute, the Alzheimer's Society, the Michael J. Fox Foundation for Parkinson's Research, Canadian Institutes for Health Research, National Sciences and Engineering Research Council Canada, and Fondation de Recherches Santé Québec. AV is funded by the Canadian Institutes of Health Research, Ontario Mental Health Foundation, the Brain and Behavior Research Foundation, and the National Institute of Mental Health (R01MH099167 and R01MH102324). FEP data was supported by a CIHR (#68961) to Dr. Martin Lepage and Dr. Ashok Malla. The preterm neonate cohort is supported by Canadian Institutes for Health Research (CIHR) operating grants MOP-79262 (SPM) and MOP-86489 (Dr. Ruth Grunau). SPM is supported by the Bloorview Children's Hospital Chair in Pediatric Neuroscience. The authors thank Drs. Ruth Grunau, Anne Synnes, Vann Chau, and Kenneth J. Poskitt for their contributions in studying the preterm neonatal cohort and providing access to the MR images. Computations were performed on the GPC supercomputer at the SciNet HPC Consortium (Loken et al., 2010). SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto. In addition, computations were performed on the CAMH Specialized Computing Cluster. The SCC is funded by the Canada Foundation for Innovation, Research Hospital Fund. ADNI Acknowledgments: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research Development, LLC.; Johnson & Johnson Pharmaceutical Research Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the

Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is Rev March 26, 2012 coordinated by the Alzheimer's disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514. We would also like to thank Curt Johnson and Robert Donner for inspiring some of the ideas in this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnins.2016.00325

REFERENCES

Aisen, P. S., Petersen, R. C., Donohue, M. C., Gamst, A., Raman, R., Thomas, R. G., et al. (2010). Clinical core of the Alzheimers Disease neuroimaging initiative: progress and plans. *Alzheimers Dement. 6*, 239–246. doi: 10.1016/j.jalz.2010.03.006

Akhondi-Asl, A., and Warfield, S. K. (2012). Estimation of the prior distribution of ground truth in the STAPLE algorithm: an empirical bayesian approach. *Med. Image Comput. Comput. Assist. Interv*. 15, 593–600. doi: 10.1007/978-3-642-33415-3_73

Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage 46*, 726–738. doi: 10.1016/j.neuroimage.2009.02.018

Barnes, J., Boyes, R. G., Lewis, E. B., Schott, J. M., Frost, C., Scahill, R. I., et al. (2007). Automatic calculation of hippocampal atrophy rates using a hippocampal template and the boundary shift integral. *Neurobiol. Aging*. 28, 1657–1663. doi: 10.1016/j.neurobiolaging.2006.07.008

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 327, 307–310. doi: 10.1016/S0140-6736(86)90837-8

Boccardi, M., Bocchetta, M., Ganzola, R., Robitaille, N., Redolfi, A., Duchesne, S., et al. (2015). Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. *Alzheimers Dement.* 11, 184–194. doi: 10.1016/j.jalz.2013.03.001

Chakravarty, M. M., Sadikot, A. F., Germann, J., Bertrand, G., and Collins, D. L. (2008). Towards a validation of atlas warping techniques. *Med. Image Anal*. 12, 713–726. doi: 10.1016/j.media.2008.04.003

Chakravarty, M. M., Sadikot, A. F., Germann, J., Hellier, P., Bertrand, G., and Collins, D. L. (2009). Comparison of piece-wise linear, linear, and nonlinear atlas-to-patient warping techniques: analysis of the labeling of subcortical nuclei for functional neurosurgical applications. *Hum. Brain Mapp*. 30, 3574–3595. doi: 10.1002/hbm.20780

Chakravarty, M. M., Steadman, P., van Eede, M. C., Calcott, R. D., Gu, V., Shaw, P., et al. (2013). Performing label fusion-based segmentation using multiple automatically generated templates. *Hum. Brain Mapp. 34*, 2635–2654. doi: 10.1002/hbm.22092

Collins, D. L., Holmes, C. J., Peters, T. M., and Evans, A. C. (1995). Automatic 3-D model-based neuroanatomical segmentation. *Hum. Brain Mapp. 3*, 190–208. doi: 10.1002/hbm.460030304

Collins, D. L., and Pruessner, J. C. (2010). Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *Neuroimage 52*, 1355–1366. doi: 10.1016/j.neuroimage.2010.04.193

Commowick, O., Akhondi-Asl, A., and Warfield, S. K. (2012). Estimating a reference standard segmentation with spatially varying performance parameters: local MAP STAPLE. *IEEE Trans. Med. Imaging 31*, 1593–1606. doi: 10.1109/TMI.2012.2197406

Commowick, O., and Warfield, S. K. (2010). "Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori STAPLE," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Vol. 13, 25–32. Available online at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3079852&tool=pmcentrez&rendertype=abstract

Coupé, P., Manjón, J. V., Fonov, V., Pruessner, J., Robles, M., and Collins, D. L. (2011). Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. *Neuroimage 54* 940–954. doi: 10.1016/j.neuroimage.2010.09.018

Duvernoy, H. M., Cattin, F., Risold, P., and SpringerLink (2005).*The Human Hippocampus: Functional Anatomy, Vascularization and Serial Sections with MRI, 4th Edn.* 2013. Berlin; Heidelberg: Springer-Verlag.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron 33*, 341–355. doi: 10.1016/S0896-6273(02)00569-X

Frey, B. N., Andreazza, A. C., Nery, F. G., Martins, M. R., Quevedo, J., Soares, J. C., et al. (2007). The role of hippocampus in the pathophysiology of bipolar disorder. *Behav. Pharmacol. 18*, 419–430. doi: 10.1097/FBP.0b013e3282df3cde

Guo, T., Winterburn, J. L., Pipitone, J., Duerden, E. G., Park, M. T. M., Chau, V., et al. (2015). Automatic Segmentation of the hippocampus for preterm neonates from early-in-life to term-equivalent age. *Neuroimage Clin.* 9, 176–193. doi: 10.1016/j.nicl.2015.07.019

Hao, Y., Wang, T., Zhang, X., Duan, Y., Yu, C., Jiang, T., et al. (2014). Local label learning (LLL) for subcortical structure segmentation: application to hippocampus segmentation. *Hum. Brain Mapp. 35*, 2674–2697. doi: 10.1002/hbm.22359

Harrison, P. J. (2004). The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications. *Psychopharmacology (Berl). 174*, 151–162. doi: 10.1007/s00213-003-1761-y

Heckemann, R. A., Keihaninejad, S., Aljabar, P., Gray, K. R., Nielsen, C., Rueckert, D., et al. (2011). Automatic morphometry in Alzheimers disease and mild cognitive impairment. *Neuroimage 56*, 2024–2037. doi: 10.1016/j.neuroimage.2011.03.014

Heckemann, R. A, Hajnal, J. V., Aljabar, P., Rueckert, D., and Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *Neuroimage 33*, 115–126. doi: 10.1016/j.neuroimage.2006.05.061

Jack, C. R., Barkhof, F., Bernstein, M. A., Cantillon, M., Cole, P. E., Decarli, C., et al. (2011). Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimers disease. *Alzheimers Dement.* 7, 474–485. doi: 10.1016/j.jalz.2011.04.007

Jorge Cardoso, M., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., et al. (2013). STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. *Med. Image Anal.* 17, 671–684. doi: 10.1016/j.media.2013.02.006

Kempton, M. J., Salvador, Z., Munafò, M. R., Geddes, J. R., Simmons, A., Frangou, S., et al. (2011). Structural neuroimaging studies in major depressive disorder. *Meta-analysis and comparison with bipolar disorder. Arch. Gen. Psychiatry* 68, 675–690. doi: 10.1001/archgenpsychiatry.2011.60

La Joie, R., Perrotin, A., de La Sayette, V., Egret, S., Doeuvre, L., Belliard, S., et al. (2013). Hippocampal subfield volumetry in mild cognitive impairment, Alzheimers disease and semantic dementia. *NeuroImage. Clin. 3*, 155–162. doi: 10.1016/j.nicl.2013.08.007

Lerch, J. P., Pruessner, J., Zijdenbos, A. P., Collins, D. L., Teipel, S. J., Hampel, H., et al. (2008). Automated cortical thickness measurements from MRI can accurately separate Alzheimers patients from normal elderly controls. *Neurobiol. Aging 29*, 23–30. doi: 10.1016/j.neurobiolaging.2006.09.013

Leung, K. K., Barnes, J., Ridgway, G. R., Bartlett, J. W., Clarkson, M. J., Macdonald, K., et al. (2010). Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *NeuroImage 51*, 1345–1359. doi: 10.1016/j.neuroimage.2010.03.018

Loken, C., Gruner, D., Groer, L., Peltier, R., Bunn, N., Craig, M., et al. (2010). SciNet: lessons learned from building a power-efficient Top-20 System and data centre. *J. Phys.* 256:012026. doi: 10.1088/1742-6596/256/1/012026

Lötjönen, J. M., Wolz, R., Koikkalainen, J. R., Thurfjell, L., Waldemar, G., Soininen, H., et al. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage 49*, 2352–2365. doi: 10.1016/j.neuroimage.2009.10.026

Malla, A., Norman, R., McLean, T., Scholten, D., and Townsend, L. (2003). A Canadian programme for early intervention in non-affective psychotic disorders. *Aust. N. Z. J. Psychiatry.* 37, 407–413. doi: 10.1046/j.1440-1614.2003.01194.x

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. (1995). A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). NeuroImage 2, 89–101. doi: 10.1006/nimg.1995.1012

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al. (2001). A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 356, 1293–1322. doi: 10.1098/rstb.2001.0915

Meda, S. A., Koran, M. E. I., Pryweller, J. R., Vega, J. N., and Thornton-Wells, T. A. (2013). Genetic interactions associated with 12-month atrophy in hippocampus and entorhinal cortex in Alzheimers Disease neuroimaging initiative. *Neurobiol. Aging* 34, 1518.e9–1518.e18. doi: 10.1016/j.neurobiolaging.2012.09.020

Morra, J. H., Tu, Z., Apostolova, L. G., Green, A. E., Avedissian, C., Madsen, S. K., et al. (2008). Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. *Neuroimage* 43, 59–68. doi: 10.1016/j.neuroimage.2008.07.003

Mouiha, A., and Duchesne, S. (2011). Hippocampal atrophy rates in Alzheimers disease: automated segmentation variability analysis. *Neurosci. Lett. 495*, 6–10. doi: 10.1016/j.neulet.2011.02.065

Park, M. T. M., Pipitone, J., Baer, L. H., Winterburn, J. L., Shah, Y., Chavez, S., et al. (2014). Derivation of high-resolution MRI atlases of the human cerebellum at 3T and segmentation using multiple automatically generated templates. *Neuroimage* 95, 217–231. doi: 10.1016/j.neuroimage.2014.03.037

Pausova, Z., Paus, T., Abrahamowicz, M., Almerigi, J., Arbour, N., Bernard, M., et al. (2007). Genes, maternal smoking, and the offspring brain and body during adolescence: design of the saguenay youth study. *Hum. Brain Mapp. 28*, 502–518. doi: 10.1002/hbm.20402

Pipitone, J., Park, M. T. M., Winterburn, J., Lett, T. A., Lerch, J. P., Pruessner, J. C., et al. (2014). Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *Neuroimage 101*, 494–512. doi: 10.1016/j.neuroimage.2014.04.054

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell Syst. Tech. J.* 36, 1389–1401. doi: 10.1002/j.1538-7305.1957.tb01515.x

Pruessner, J. C., Li, L. M., Serles, W., Pruessner, M., Collins, D. L., Kabani, N., et al. (2000). Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. *Cereb. Cortex* 10, 433–442. doi: 10.1093/cercor/10.4.433

Risacher, S. L., Saykin, A. J., West, J. D., Shen, L., Firpi, H. A., McDonald, B. C., et al. (2009). Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr. Alzheimer Res.* 6, 347–361. doi: 10.2174/156720509788929273

Rousseau, F., Habas, P. A., and Studholme, C. (2011). A supervised patch-based approach for human brain labeling. *IEEE Trans. Med. Imaging 30*, 1852–1862. doi: 10.1109/TMI.2011.2156806

Sabuncu, M. R., Desikan, R. S., Sepulcre, J., Yeo, B. T. T., Liu, H., Schmansky, N. J., et al. (2011). The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Arch. Neurol. 68*, 1040–1048. doi: 10.1001/archneurol.2011.167

Sabuncu, M. R., Yeo, B. T. T., Van Leemput, K., Fischl, B., and Golland, P. (2010). A generative model for image segmentation based on label fusion. *IEEE Trans. Med. Imaging 29*, 1714–1729. doi: 10.1109/TMI.2010.2050897

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23(Suppl 1), S208–S219. doi: 10.1016/j.neuroimage.2004.07.051

Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognit. 32*, 71–86.

Tong, T., Wolz, R., Coupé P., Hajnal, J. V., Rueckert, D. (2013). ADNI. segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. *Neuroimage* 76, 11–23. doi: 10.1016/j.neuroimage.2013.02.069

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (2003). A unifying framework for partial volume segmentation of brain MR images. *IEEE Trans. Med. Imaging 22*, 105–119. doi: 10.1109/TMI.2002.806587

Wang, H., Pouch, A., Takabe, M., Jackson, B.,Gorman, J. Gorman, R.,et al. (2013). "Multi-atlas segmentation with robust label transfer and label fusion," in *Proceedings of the 23rd International Conference on Information Processing in Medical Imaging,* Vol. 7917, eds J. C. Gee, S. Joshi, K. M. Pohl, W. M. Wells, and L. Zöllei (Berlin; Heidelberg: Springer Berlin Heidelberg), 548–559.

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., and Yushkevich, P. A. (2012). Multi-Atlas Segmentation with Joint Label Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 611–623. doi: 10.1109/TPAMI.2012.143

Wang, L., Gao, Y., Shi, F., Li, G., Gilmore, J. H., Lin, W., et al. (2015). Learning-based multi-source IntegratioN frameworK For Segmentation of infant brain images. *Neuroimage* 108, 160–172. doi: 10.1016/j.neuroimage.2014.12.042

Wang, L., Shi, F., Gao, Y., Li, G., Gilmore, J. H., Lin, W., et al. (2014). Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation. *Neuroimage* 89, 152–164. doi: 10.1016/j.neuroimage.2013.11.040

Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging 23*, 903–921. doi: 10.1109/TMI.2004.828354

Weiner, M. W. (2013). Dementia in 2012: further insights into Alzheimer disease pathogenesis. *Nat. Rev. Neurol.* 9, 65–66. doi: 10.1038/nrneurol.2012.275

Winston, G. P., Cardoso, M. J., Williams, E. J., Burdett, J. L., Bartlett, P. A., Espak, M., et al. (2013). Automated hippocampal segmentation in patients with epilepsy: available free online. *Epilepsia* 54, 2166–2173. doi: 10.1111/epi.12408

Winterburn, J. L., Pruessner, J. C., Chavez, S., Schira, M. M., Lobaugh, N. J., Voineskos, A. N., et al. (2013). A novel *in vivo* atlas of human hippocampal subfields using high-resolution 3 T magnetic resonance imaging. *NeuroImage 74*, 254–265. doi: 10.1016/j.neuroimage.2013.02.003

Wolz, R., Aljabar, P., Hajnal, J. V., Hammers, A., and Rueckert, D. (2010). LEAP: learning embeddings for atlas propagation. *NeuroImage 49*, 1316–1325. doi: 10.1016/j.neuroimage.2009.09.069

Wolz, R., Aljabar, P., Rueckert, D., Heckemann, R. A., and Hammers, A. (2009). "segmentation of subcortical structures and the hippocampus in brain mri using graph-cuts and subject-specific a-priori information," in *Proceedigs of IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (Boston, MA), 470–473.

Wu, G., Wang, Q., Zhang, D., Nie, F., Huang, H., and Shen, D. (2014). A generative probability model of joint label fusion for multi-atlas based brain segmentation. *Med. Image Anal. 18*, 881–890. doi: 10.1016/j.media.2013.10.013

Wyman, B. T., Harvey, D. J., Crawford, K., Bernstein, M. A., Carmichael, O., Cole, P. E., et al. (2013). Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimers Dement.* 9, 332–337. doi: 10.1016/j.jalz.2012.06.004

Yushkevich, P. A., Wang, H., Pluta, J., and Avants, B. B. (2012). From label fusion to correspondence fusion: a new approach to unbiased groupwise registration. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern. Recognit*. 956–963. doi: 10.1109/CVPR.2012.6247771

Zhang, D., Wu, G., Jia, H., and Shen, D. (2011). "Confidence-guided sequential label fusion for multi-atlas based segmentation," in *Medical Image Computing and Computer-Assisted Intervention* (Toronto, ON: Springer).

Zikic, D., Glocker, B., and Criminisi, A. (2014). Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. *Med. Image Anal.* 18, 1262–1273. doi: 10.1016/j.media.2014.06.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.