Fast and Sequence-Adaptive Whole-Brain Segmentation Using Parametric Bayesian Modeling

Puonti, Oula; Iglesias, Juan Eugenio; Van Leemput, Koen

Published in:
NeuroImage

Link to article, DOI:
10.1016/j.neuroimage.2016.09.011

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

NeuroImage 143 (2016) 235–249

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling

Oula Puonti [a,*], Juan Eugenio Iglesias [b,d], Koen Van Leemput [a,c]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
[b] Basque Center on Cognition, Brain and Language (BCBL), Paseo Mikeletegi, 20009 San Sebastian - Donostia, Gipuzkoa, Spain
[c] Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, 149 13th St, Charlestown, MA 02129, USA
[d] Department of Medical Physics and Biomedical Engineering, University College London, Gower St, London WC1E 6BT, United Kingdom

ARTICLE INFO

Article history:
Received 10 June 2016
Accepted 5 September 2016
Available online 7 September 2016

Keywords:
MRI
Segmentation
Atlases
Parametric models
Bayesian modeling

ABSTRACT

Quantitative analysis of magnetic resonance imaging (MRI) scans of the brain requires accurate automated segmentation of anatomical structures. A desirable feature for such segmentation methods is to be robust against changes in acquisition platform and imaging protocol. In this paper we validate the performance of a segmentation algorithm designed to meet these requirements, building upon generative parametric models previously used in tissue classification. The method is tested on four different datasets acquired with different scanners, field strengths and pulse sequences, demonstrating comparable accuracy to state-of-the-art methods on T1-weighted scans while being one to two orders of magnitude faster. The proposed algorithm is also shown to be robust against small training datasets, and readily handles images with different MRI contrast as well as multi-contrast data.

1. Introduction

So-called *whole-brain segmentation* techniques aim to automatically label a multitude of cortical and subcortical regions from brain MRI scans. Recent years have seen tremendous advances in this field, enabling, for the first time, fine-grained comparisons of regional brain morphometry between large groups of subjects. Current state-of-the-art whole-brain segmentation algorithms are typically based on supervised models of image appearance in T1-weighted scans, in which the relationship between intensities and neuroanatomical labels is learned from a set of manually annotated training images.

This approach suffers from two fundamental limitations. First, segmentation performance often degrades when the algorithms are applied to T1-weighted data acquired on different scanner platforms or using different imaging sequences, due to subtle changes in the obtained image contrast (Han and Fischl, 2007; Roy et al., 2013). And second, the exclusive focus on only T1-weighted images hinders the ultimate translation of whole-brain segmentation techniques into clinical practice, where they hold great potential to support personalized treatment of patients suffering from brain diseases. This is because clinical imaging uses additional MRI contrast mechanisms to show clinically relevant information, including T2-weighted or fluid attenuated inversion recovery (FLAIR) images that are much more sensitive to certain pathologies than T1-weighted scans (e.g., white matter lesions or brain tumors). Although incorporating models of lesions into whole-brain segmentation techniques is an open problem in itself, a first necessary step towards bringing these techniques into clinical practice is to make them capable of handling the multi-contrast images that are acquired in standard clinical routine.

In this article, we present and validate the performance of a fast, sequence-independent whole-brain segmentation algorithm. The method, which is based on a mesh-based computational atlas combined with a Gaussian appearance model, yields segmentation accuracies comparable to the state of the art; automatically adapts to different MRI contrasts (even if multimodal); requires only a small amount of training data; and achieves computational times comparable to those of the fastest algorithms in the field (Zikic et al., 2014; Ta et al., 2014).

1.1. Current state of the art in whole-brain segmentation

Early methods for the segmentation of brain structures often relied on *parametric* models, in which the available training data were summarized in relevant statistics that were subsequently used to inform the segmentation of previously unseen subjects.

* Corresponding author.
E-mail address: oupu@dtu.dk (O. Puonti).

http://dx.doi.org/10.1016/j.neuroimage.2016.09.011

Because many distinct brain structures have similar intensity characteristics in MRI, these methods were typically built around detailed probabilistic models of the expected shape and relative positioning of different brain regions, using surface-based (Kelemen et al., 1998; Pizer et al., 2003; Patenaude et al., 2011; Cootes et al., 1998) or volumetric (Fischl et al., 2002; Pohl et al., 2006b) models. These anatomical models were then combined with supervised models of appearance to encode the typical intensity characteristics of the relevant structures in the training data, often using Gaussian models for either the intensity of individual voxels (Fischl et al., 2002; Pohl et al., 2006b) or for entire regional intensity profiles (Kelemen et al., 1998; Pizer et al., 2003; Patenaude et al., 2011; Cootes et al., 1998). The segmentation problem was then formulated in a Bayesian setting, in which segmentations were sought that satisfy both the shape and appearance constraints.

More recently, *non-parametric* methods[1] have gained increasing attention in the field of whole-brain segmentation, mostly in the form of multi-atlas label fusion (Rohfling et al., 2004a; Heckemann et al., 2006; Isgum et al., 2009; Artaechevarria et al., 2009; Sabuncu et al., 2010; Rohfling et al., 2004b; Wang et al., 2013; Manjón et al., 2011; Rousseau et al., 2011; Tong and Wolz, 2013; Wu et al., 2014; Asman and Landman, 2013; Zikic et al., 2014; Iglesias and Sabuncu, 2015). In these methods, each of the manually annotated training scans is first deformed onto the target image using an image registration algorithm. Then, the resulting deformation fields are used to warp the manual annotations, which are subsequently fused into a final consensus segmentation. Although early methods used a simple majority voting rule (Rohfling et al., 2004a; Heckemann et al., 2006), recent developments have concentrated on exploiting local intensity information to guide the atlas fusion process. This is particularly helpful in cortical areas, for which accurate inter-subject registration is challenging (Sabuncu et al., 2010; Ledig et al., 2012). Label fusion methods have been shown to yield very accurate whole-brain segmentations (Landman and Warfield, 2012), but their accuracy comes at the expense of a high computational cost as a result of the multiple non-linear registrations that are required. Efforts to alleviate this issue include a local search using entire image patches, such that much faster *linear* registrations can be used (Manjón et al., 2011; Ta et al., 2014), as well as using rich contextual features so that only a single non-linear warp is needed (Zikic et al., 2014).

1.2. Existing methods that handle changes in MRI contrast

With the exception of simple majority voting (Rohfling et al., 2004a; Heckemann et al., 2006), all the methods reviewed above use *supervised* intensity models, in the sense that they explicitly exploit the specific image contrast properties of the dataset used for training. This poses limitations on their ability to segment images that were acquired with different scanners or imaging sequences than the training scans.

A generic way of making such methods work across imaging platforms is histogram matching (also known as intensity normalization), in which the intensity profiles of new images are altered so as to resemble those of the images used for training (Nyúl et al., 2000; Roy et al., 2013). However, histogram matching can only be used when the training and target data have been acquired with the same type of MRI sequence (e.g., T1-weighted), and it does not completely cancel the negative effects that intensity mismatches have on segmentation accuracy (Roy et al., 2013).

Another approach is to have the training dataset include images that are representative of all the scanners and protocols that are expected to be encountered in practice. However, this approach quickly becomes impractical due to the large number of possible combinations of MRI hardware and acquisition parameters. The situation is exacerbated for clinical data, due to the lack of standardized protocols to acquire multi-contrast MRI data across clinical imaging centers.

In contrast synthesis (Roy et al., 2013), the original scan is not directly segmented, but rather used to generate a new scan with the desired intensity profile, which is then segmented instead. The premise of this technique is that a database of scans acquired with both the source and target contrast is available, so that the relationship between the two can be learned (Iglesias et al., 2013a; Roy et al., 2013). This approach makes it unnecessary to manually annotate additional training data for each new set-up that is considered – a considerable advantage given that a manual whole-brain segmentation often takes several days per scan (Fischl et al., 2002). However, it still requires that additional example subjects are scanned with both the source and target scanner and protocol, which is not always practical.

Finally, a more fundamental way to address the problem is to perform whole-brain segmentation in the space of intrinsic MRI tissue parameters (Fischl et al., 2004b). However, this requires the usage of specific MRI sequences for which a physical forward model is available, which are not widely implemented on MRI scanning platforms, and particularly not on clinical systems.

1.3. Contribution: validation of a fast, sequence-adaptive whole-brain segmentation algorithm

In contrast to the aforementioned approaches to whole-brain segmentation, which rely on supervised models of the specific intensity profiles seen in the training data, in this paper we validate an *unsupervised* approach that automatically learns appropriate intensity models from the images being analyzed. At the core of the method is an intensity clustering algorithm (a Gaussian mixture model) that derives its independence of specific image contrast properties by simply grouping together voxels with similar intensities. This approach is well-established for the purpose of tissue classification (aimed at extracting the white matter, gray matter and cerebrospinal fluid) where it is typically augmented with models of MRI imaging artifacts (Wells et al., 1996a; Van Leemput et al., 1999a; Ashburner and Friston, 2005) and spatial models such as probabilistic atlases (Ashburner and Friston, 1997; Van Leemput et al., 1999a; Ashburner and Friston, 2005) or Markov random fields (Van Leemput et al., 1999b; Zhang et al., 2001).

Here we validate a method for whole-brain segmentation that is rooted in this type of approach, building on prior work from our group including a proof-of-concept demonstration in whole-brain segmentation (Van Leemput, 2009), as well as the automated segmentation methods for hippocampal subfields (Iglesias et al., 2015a) and subregions of the brainstem (Iglesias et al., 2015b) that are distributed with the FreeSurfer software package (Fischl et al., 2002). The method we validate here uses a mesh-based probabilistic atlas to provide whole-brain segmentation accuracy at the level of the state of the art, both within and across scanner platforms and pulse sequences. Unlike many other techniques, the method does not need any preprocessing such as skull stripping, bias field correction or intensity normalization. Furthermore, because the method is parametric, only a single non-linear registration (of the atlas to the target image) is required, yielding a very fast overall computational footprint.

An early version of this work, with a preliminary validation, was presented in Puonti et al. (2013). The current article adds a

[1] Note that the distinction between parametric vs. non-parametric methods here only refers to the overall segmentation approach that is taken – the pair-wise registrations in non-parametric segmentation methods can still be either parametric (e.g., B-splines, Rueckert et al. (1999)) or non-parametric (e.g., Demons, Thirion (1998)).

more detailed explanation of our modeling approach, quantitative comparisons with additional state-of-the-art label fusion algorithms, and more extensive experiments – particularly regarding test-retest reliability, segmentation of multi-contrast and non-T1-contrast data, and the sensitivity of the method to the size of the training dataset.

2. Modeling framework

Let $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_I)$ denote a matrix collecting the intensities in a multi-contrast brain MRI scan with I voxels, where the vector $\mathbf{d}_i = (d_i^1, \ldots, d_i^N)^T$ contains the intensities in voxel i for each of the available N contrasts. Furthermore, let $\mathbf{l} = (l_1, \ldots, l_I)$ be the corresponding segmentation, where $l_i \in \{1, \ldots, K\}$ denotes the one of K possible segmentation labels assigned to voxel i.

In order to estimate $\mathbf{l}$ from $\mathbf{D}$, i.e., to compute automated segmentations, we use a generative modeling approach: a forward probabilistic model of MRI images is defined, and subsequently "inverted" to obtain the segmentation. The model consists of two parts: a prior and a likelihood. The prior is a probability distribution over segmentations $p(\mathbf{l})$ that encodes prior knowledge on human neuroanatomy. The likelihood is a probability distribution over image intensities that is conditioned on the segmentation $p(\mathbf{D}|\mathbf{l})$, which models the imaging process through which a certain segmentation yields the observed MRI scan. This type of model is generative because it provides a mechanism to generate data through the forward model: in our case, we could generate a random brain MRI scan by first sampling the prior to obtain a segmentation, and then sampling the likelihood conditioned on the resulting segmentation.

Within this framework, the posterior distribution of image segmentations given an input brain MRI scan is given by Bayes' rule:

$$p(\mathbf{l}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{l})p(\mathbf{l}). \tag{1}$$

Maximizing Eq. (1) with respect to $\mathbf{l}$ then yields the maximum a posteriori (MAP) estimate of the segmentation.

In the rest of this Section, we will describe in depth the prior (Section 2.1) and likelihood (Section 2.2); we will propose an inference algorithm to approximately maximize Eq. (1) (Section 2.3); and finally we will describe the details of the implementation of this algorithm (Section 2.4).

2.1. Prior

For the prior $p(\mathbf{l})$ we use a generalization of the probabilistic brain atlases often used in brain MRI segmentation (Ashburner and Friston, 1997; Van Leemput et al., 1999b, 1999a, 2001; Zijdenbos et al., 2002; Fischl et al., 2002; Ashburner and Friston, 2005; Prastawa et al., 2005; Pohl et al., 2006b; D'Agostino et al., 2006; Awate et al., 2006; Bouix et al., 2007). This model, detailed in Van Leemput (2009), is based on a deformable tetrahedral mesh, the properties of which are learned automatically from a set of manual example segmentations made on MRI scans of training subjects. Each of the vertices of the mesh has an associated set of label probabilities specifying how frequently each of the K labels occurs at the vertex. The resolution of the mesh is locally adaptive, being sparse in large uniform regions and dense around the structure borders. This automatically introduces a locally varying amount of spatial blurring in the resulting atlas, aiming to avoid over-fitting of the model to the available training samples (Van Leemput, 2009). During training, the topology of the mesh and the position of its vertices in atlas space (henceforth "reference position") is computed along with the label probabilities in a non-linear, group-wise registration of the labeled training data. An example of the resulting probabilistic brain atlas, computed from manual parcellations in 20 subjects, is displayed in its reference position in Fig. 1; note the irregularity in the shapes and sizes of the tetrahedra.

The positions of the mesh nodes $\mathbf{x}$ can change according to their prior distribution $p(\mathbf{x})$:

$$p(\mathbf{x}) \propto \exp\left(-\beta \sum_{t=1}^{T} \phi_t(\mathbf{x}, \mathbf{x}_{ref})\right) \tag{2}$$

where T and $\mathbf{x}_{ref}$ denote the number of tetrahedra and the reference position of the mesh, respectively; $\phi_t(\mathbf{x}, \mathbf{x}_{ref})$ is a penalty for deforming tetrahedron t from its reference to its actual position; and $\beta > 0$ is a scalar that controls the global stiffness of the mesh. We use the penalty term proposed in Ashburner et al. (2000), which goes to infinity when the Jacobian determinant of the deformation approaches zero. This choice prevents the mesh from tearing or folding onto itself, thus preserving its topology.

Given a deformed mesh with node positions $\mathbf{x}$, the probability $p_i(k|\mathbf{x})$ of observing label k at a voxel i is obtained by barycentric interpolation of the label probabilities at the vertices of the tetrahedron containing the voxel. Moreover, we assume conditional

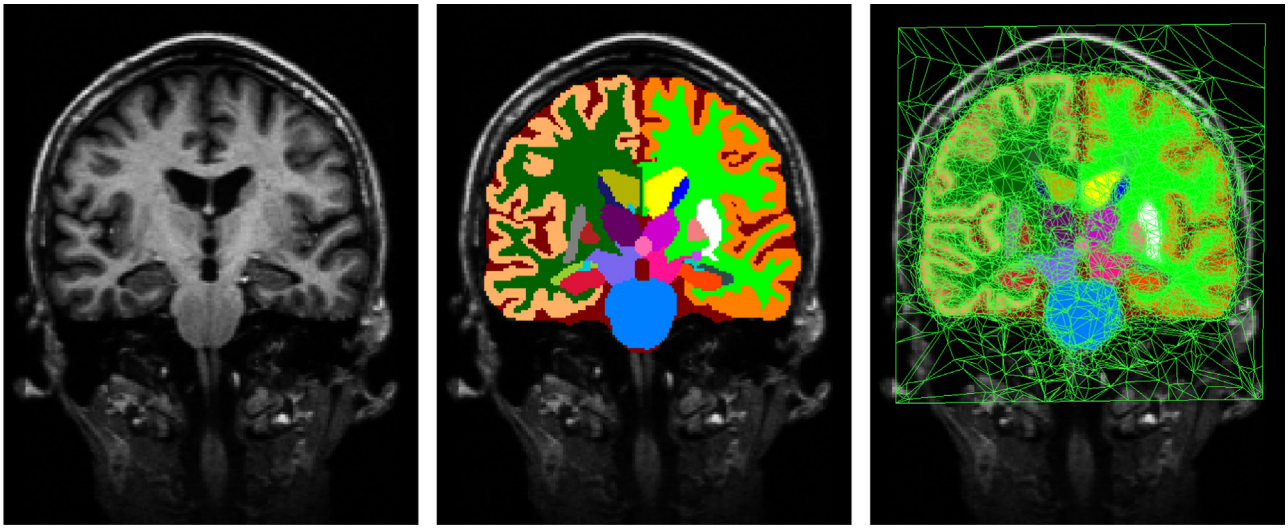

Fig. 1. Left: T1-weighted scan from the training data. Center: corresponding manual segmentation. Right: atlas mesh built from 20 randomly selected subjects from the training data.

independence of the labels of the different voxels given the mesh node positions, such that

$$p(\mathbf{l}|\mathbf{x}) = \prod_{i=1}^{I} p_i(l_i|\mathbf{x}). \tag{3}$$

The expression for the prior distribution over segmentations is finally:

$$p(\mathbf{l}) = \int_{\mathbf{x}} p(\mathbf{l}|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{4}$$

2.2. Likelihood

The likelihood $p(\mathbf{D}|\mathbf{l})$ models the relationship between segmentation labels and image intensities. For this purpose, we associate a mixture of Gaussian distributions with each label (Ashburner and Friston, 2005), and assume that the bias field imaging artifact typically seen in MRI can be modeled as a multiplicative and spatially smooth effect (Wells et al., 1996a). For computational reasons, we use log-transformed image intensities in $\mathbf{D}$, and model the bias field as a linear combination of spatially smooth basis functions that is *added* to the local voxel intensities (Van Leemput et al., 1999a).

Specifically, letting $\boldsymbol{\theta}$ denote all bias field and Gaussian mixture parameters, with uniform prior $p(\boldsymbol{\theta}) \propto 1$, the likelihood is defined by

$$p(\mathbf{D}|\mathbf{l}) = \int_{\boldsymbol{\theta}} p(\mathbf{D}|\mathbf{l}, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{5}$$

where

$$\begin{aligned} p(\mathbf{D}|\mathbf{l}, \boldsymbol{\theta}) &= \prod_{i=1}^{I} p_i(\mathbf{d}_i|l_i, \boldsymbol{\theta}), \\ p_i(\mathbf{d}|k, \boldsymbol{\theta}) &= \sum_{g=1}^{G_k} w_{k,g}\mathcal{N}\left(\mathbf{d} - \mathbf{C}\boldsymbol{\phi}^i|\boldsymbol{\mu}_{k,g}, \boldsymbol{\Sigma}_{k,g}\right), \end{aligned} \tag{6}$$

and

$$\mathcal{N}(\mathbf{d}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^N|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{d} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{d} - \boldsymbol{\mu})\right).$$

Here, G_k is the number of Gaussian distributions in the mixture associated with label k; and $\boldsymbol{\mu}_{k,g}$, $\boldsymbol{\Sigma}_{k,g}$, and $w_{k,g}$ are the mean, covariance matrix, and weight of component $g \in \{1, \ldots, G_k\}$ in the mixture model of label k (satisfying $w_{k,g} \geq 0$ and $\sum_g w_{k,g} = 1$). Furthermore,

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_N^T \end{pmatrix}, \quad \mathbf{c}_n = \begin{pmatrix} c_{n,1} \\ \vdots \\ c_{n,P} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\phi}^i = \begin{pmatrix} \phi_1^i \\ \phi_2^i \\ \vdots \\ \phi_P^i \end{pmatrix},$$

where P denotes the number of bias field basis functions, ${\phi_p}^i$ is the basis function p evaluated at voxel i, and $\mathbf{c}_n$ holds the bias field coefficients for MRI contrast n.

The entire forward model is summarized in Table 1.

Table 1
Equations for the forward probabilistic model of MRI brain scans.

$\mathbf{x}$	$\sim$	$p(\mathbf{x})$	(Eq. (2))
$\mathbf{l}$	$\sim$	$p(\mathbf{l}\|\mathbf{x})$	(Eq. (3))
$\boldsymbol{\theta}$	$\sim$	$p(\boldsymbol{\theta}) \propto 1$	
$\mathbf{D}$	$\sim$	$p(\mathbf{D}\|\mathbf{l}, \boldsymbol{\theta})$	(Eq. (6))

2.3. Inference

Using the model described above, the MAP segmentation for a given MRI scan is obtained by maximizing Eq. (1) with respect to $\mathbf{l}$:

$$\hat{\mathbf{l}} = \arg\max_{\mathbf{l}} p(\mathbf{l}|\mathbf{D}) = \arg\max_{\mathbf{l}} p(\mathbf{D}|\mathbf{l})p(\mathbf{l}), \tag{7}$$

which is intractable due to the integrals over the parameters $\mathbf{x}$ and $\boldsymbol{\theta}$ that appear in the expressions for $p(\mathbf{l})$ (Eq. (4)) and $p(\mathbf{D}|\mathbf{l})$ (Eq. (5)), respectively. This difficulty can be side-stepped if the posterior distribution of the model parameters in light of the data is heavily peaked around its mode:

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{D}) \simeq \delta\left(\mathbf{x} - \hat{\mathbf{x}}, \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right),$$

where $\delta(\cdot)$ is Dirac's delta and the point estimates $\left\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right\}$ are given by:

$$\left\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right\} = \arg\max_{\{\mathbf{x},\boldsymbol{\theta}\}} p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{D}). \tag{8}$$

In that scenario, we can approximate:

$$\begin{aligned} p(\mathbf{l}|\mathbf{D}) &= \int_{\mathbf{x}}\int_{\boldsymbol{\theta}} p(\mathbf{l}|\mathbf{D}, \mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{D})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta} \\ &\simeq p\left(\mathbf{l}|\mathbf{D}, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right), \end{aligned} \tag{9}$$

which no longer involves intractable integrals. The resulting inference algorithm then involves two distinct phases, detailed below: first, computing the point estimates by maximizing Eq. (8); and subsequently computing the segmentation by maximizing Eq. (9) with respect to $\mathbf{l}$.

Computation of point estimates. Applying Bayes' rule to Eq. (8), we obtain:

$$\begin{aligned} p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{D}) &\propto p(\mathbf{D}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x})p(\boldsymbol{\theta}) \\ &\propto \left(\sum_{\mathbf{l}} p(\mathbf{D}|\mathbf{l}, \boldsymbol{\theta})p(\mathbf{l}|\mathbf{x})\right)p(\mathbf{x}) \\ &= \prod_{i=1}^{I}\left(\sum_{k=1}^{K} p_i(\mathbf{d}_i|k, \boldsymbol{\theta})p_i(k|\mathbf{x})\right)p(\mathbf{x}). \end{aligned}$$

Taking the logarithm, we can rewrite the problem as the maximization of the following objective function:

$$\left\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right\} = \operatorname*{argmax}_{\{\mathbf{x},\boldsymbol{\theta}\}}\left[\sum_{i=1}^{I} \log\left(\sum_{k=1}^{K} p_i(\mathbf{d}_i|k, \boldsymbol{\theta})p_i(k|\mathbf{x})\right) + \log p(\mathbf{x})\right]. \tag{10}$$

We solve this problem with a coordinate ascent scheme, in which the mesh node positions $\mathbf{x}$ and likelihood parameters $\boldsymbol{\theta}$ are iteratively updated, by alternately optimizing one while keeping the other fixed.

To optimize the mesh node positions $\mathbf{x}$ with fixed $\boldsymbol{\theta}$, we use a standard conjugate gradient optimizer (Shewchuk, 1994). To optimize the likelihood parameters $\boldsymbol{\theta}$ with fixed $\mathbf{x}$, we use a generalized expectation-maximization (GEM) algorithm (Dempster et al., 1977) similar to the one proposed in Van Leemput et al. (1999a). In particular, the GEM optimization involves iteratively computing the following soft assignments of each voxel to each of the Gaussian distributions, based on the current parameter estimates:

$$q_i^{k,g} = \frac{w_{k,g}\mathcal{N}\left(\mathbf{d}_i - \mathbf{C}\boldsymbol{\phi}^i|\boldsymbol{\mu}_{k,g}, \boldsymbol{\Sigma}_{k,g}\right)p_i(k|\mathbf{x})}{\sum_{k'=1}^{K} p_i(\mathbf{d}_i|k', \boldsymbol{\theta})p_i(k'|\mathbf{x})}, \tag{11}$$

and subsequently updating the parameters accordingly:

$$\boldsymbol{\mu}_{k,g} \leftarrow \frac{\sum_{i=1}^{I} q_i^{k,g}(\mathbf{d}_i - \mathbf{C}\boldsymbol{\phi}^i)}{\sum_{i=1}^{I} q_i^{k,g}}, \quad w_{k,g} \leftarrow \frac{\sum_{i=1}^{I} q_i^{k,g}}{\sum_{i=1}^{I}\sum_{g'=1}^{G_k} q_i^{k,g'}},$$

$$\boldsymbol{\Sigma}_{k,g} \leftarrow \frac{\sum_{i=1}^{I} q_i^{k,g}(\mathbf{d}_i - \boldsymbol{\mu}_{k,g} - \mathbf{C}\boldsymbol{\phi}^i)(\mathbf{d}_i - \boldsymbol{\mu}_{k,g} - \mathbf{C}\boldsymbol{\phi}^i)^T}{\sum_{i=1}^{I} q_i^{k,g}},$$

$$\begin{pmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_N \end{pmatrix} \leftarrow \begin{pmatrix} \mathbf{A}^T\mathbf{S}_{1,1}\mathbf{A} & \dots & \mathbf{A}^T\mathbf{S}_{1,N}\mathbf{A} \\ \vdots & \ddots & \vdots \\ \mathbf{A}^T\mathbf{S}_{N,1}\mathbf{A} & \dots & \mathbf{A}^T\mathbf{S}_{N,N}\mathbf{A} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \mathbf{A}^T\left(\mathbf{S}_{1,1}\mathbf{r}_{1,1} + \dots + \mathbf{S}_{1,N}\mathbf{r}_{1,N}\right) \\ \vdots \\ \mathbf{A}^T\left(\mathbf{S}_{N,1}\mathbf{r}_{N,1} + \dots + \mathbf{S}_{N,N}\mathbf{r}_{N,N}\right) \end{pmatrix},$$

where

$$\mathbf{A} = \begin{pmatrix} \phi_1^1 & \dots & \phi_P^1 \\ \vdots & \ddots & \vdots \\ \phi_1^I & \dots & \phi_P^I \end{pmatrix}, \quad \mathbf{S}_{m,n} = \mathrm{diag}\left(s_i^{m,n}\right)$$

and $\mathbf{r}_{m,n} = \left(r_1^{m,n}, \dots, r_I^{m,n}\right)^T$, with

$$s_i^{m,n} = \sum_{k=1}^{K}\sum_{g=1}^{G_k} s_{i,k,g}^{m,n}, \quad s_{i,k,g}^{m,n} = q_i^{k,g}\left(\boldsymbol{\Sigma}_{k,g}^{-1}\right)_{m,n}$$

$$r_i^{m,n} = d_i^n - \frac{\sum_{l=1}^{K}\sum_{g=1}^{G_l} s_{i,k,g}^{m,n}\left(\boldsymbol{\mu}_{k,g}\right)_n}{\sum_{l=1}^{K}\sum_{g=1}^{G_k} s_{i,k,g}^{m,n}}.$$

It can be shown that this process is guaranteed to increase the objective function of Eq. (10) with respect to $\boldsymbol{\theta}$ in each GEM iteration (Dempster et al., 1977; Van Leemput et al., 1999a).

Computation of the final segmentation. Given the point estimates of the model parameters, the conditional posterior distribution of the segmentation $\mathbf{l}$ factorizes over voxels:

$$p\left(\mathbf{l}|\mathbf{D}, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right) = \prod_{i=1}^{I} p_i\left(l_i|\mathbf{d}_i, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right), \quad p_i\left(k|\mathbf{d}_i, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right) = \sum_{g=1}^{G_k} q_i^{k,g}.$$

The optimal segmentation for each voxel is therefore given by:

$$\hat{l}_i = \underset{k}{\mathrm{argmax}} \sum_{g=1}^{G_k} q_i^{k,g}.$$

2.4. Implementation

In practice, we have found that modeling substructures with similar intensity properties (e.g., all white matter structures) with the same Gaussian mixture model improves the robustness of the algorithm while giving faster execution times. Letting f denote a set of structures that share the same mixture model, this is accomplished by altering the GEM update equations for the Gaussian mixture parameters as follows:

$$\boldsymbol{\mu}_{k,g} \leftarrow \frac{\sum_{i=1}^{I} q_i^{f,g}(\mathbf{d}_i - \mathbf{C}\boldsymbol{\phi}^i)}{\sum_{i=1}^{I} q_i^{f,g}} \quad \forall\, k \in f,$$

$$w_{k,g} \leftarrow \frac{\sum_{i=1}^{I} q_i^{f,g}}{\sum_{i=1}^{I}\sum_{g'=1}^{G_f} q_i^{f,g'}} \quad \forall\, k \in f,$$

$$\boldsymbol{\Sigma}_{k,g} \leftarrow \frac{\sum_{i=1}^{I} q_i^{f,g}(\mathbf{d}_i - \boldsymbol{\mu}_{k,g} - \mathbf{C}\boldsymbol{\phi}^i)(\mathbf{d}_i - \boldsymbol{\mu}_{k,g} - \mathbf{C}\boldsymbol{\phi}^i)^T}{\sum_{i=1}^{I} q_i^{f,g}} \quad \forall\, k \in f,$$

where

$$q_i^{f,g} = \sum_{k \in f} q_i^{k,g}.$$

The details of which structures share the same mixture models will be given in Section 3.3.

To initialize the algorithm, we first affinely align the atlas to the target image using the registration method described in D'Agostino et al. (2004), which uses atlas probabilities – rather than an intensity template – to drive the registration process. After the initial registration we mask out non-brain tissues by excluding voxels that have a prior probability lower than 0.01 of belonging to any of the brain structures.

The image intensities are then log-transformed to accommodate the additive bias field that is employed (cf. Section 2.2). For the bias field modeling, we use the lowest frequency components of the 3D discrete cosine transform (DCT) as basis functions (for the number of components see Section 3.3).

The subsequent optimization is done at two resolution levels. In the first level, the atlas probabilities are smoothed using a Gaussian kernel with a standard deviation of 2.0 mm in order to fit large scale mesh deformations. No smoothing is used in the second level, which refines the registration on a smaller scale.

The stopping criteria for the different components of the algorithm are as follows: the likelihood parameters $\boldsymbol{\theta}$ are updated until the relative change in the objective function (Eq. (10)) falls under 10^{-5}; the mesh node positions are updated until the maximum deformation across vertices falls under 10^{-3} mm; and the GEM and conjugate gradient optimizers are iteratively interleaved until the decrease in the cost function falls under 10^{-6}.

The algorithm is implemented in Matlab except for the computationally demanding optimization of the mesh node positions, which is implemented in C++, and which involves computing the mesh node deformation prior $p(\mathbf{x})$ (Eq. (2)), the interpolated prior probabilities $p(\mathbf{l}|\mathbf{x})$ (Eq. (3)) and the gradient of the objective function (Eq. (10)) with respect to the mesh node positions.

3. Experiments

In this section, we first describe the brain MRI datasets used in this study (Section 3.1). Then, we outline four methods that our algorithm is benchmarked against (Section 3.2). Next, we detail how the free parameters of each method are set (Section 3.3). Finally, we describe the setups for four different experiments in which the different methods are tested (Section 3.4).

3.1. MRI data

In the experiments, we use five different sets of scans: one exclusively for training the segmentation methods, and the other four for testing the performance on unseen data. For training, we use a dataset of 39 T1-weighted MRI scans and corresponding expert segmentations. The expert segmentations were obtained using a validated semi-automated protocol developed at the Center for Morphometric Analysis (CMA), MGH, Boston (Caviness et al., 1989, 1996; Kennedy et al., 1989). All raters had to pass tests measuring intra- and inter-rater reliability before they were allowed to perform segmentations. The resulting training data consists of 28 healthy subjects and 11 subjects with questionable or probable Alzheimer's disease with ages ranging from under 30 years old to over 60 years old (Sabuncu et al., 2010). The scans were acquired on a 1.5T Siemens Vision scanner using an MPRAGE sequence with parameters: TR=9.7 ms, TE=4 ms, TI=20 ms, flip angle=10° and voxel size=$1.0 \times 1.0 \times 1.5$ mm^3 (128 sagittal slices), where the scan parameters were empirically optimized for gray-white matter contrast (Buckner et al., 2004). This is the same dataset used for training in the publicly available software package FreeSurfer (Fischl et al., 2002). An example scan and a corresponding manual segmentation are shown in Fig. 1.

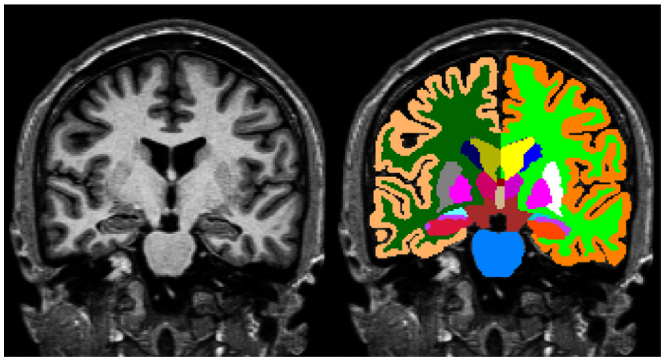

Fig. 2. On the left an example slice from the intra-scanner dataset and on the right a corresponding manual segmentation.

For testing, we use four different datasets acquired on scanners from different manufacturers, with different field strengths and pulse sequences. For three of the datasets, including a total of 35 subjects, we have access to expert manual segmentations, enabling quantitative comparisons of automated segmentation accuracy. All these manual segmentations were performed using the same protocol as was used for the training data. The fourth test dataset consists of 40 subjects scanned at two time points; it does not have expert segmentations but will be used to assess test-retest reliability instead. Below we provide details on each of these four test datasets.

The first test dataset consists of T1-weighted scans of 13 individuals with age and disease status matching those of the training dataset, acquired on a 1.5T Siemens Sonata scanner with the same sequence and parameters as the training data (Han and Fischl, 2007). Given the similarity with the training data (vendor, field strength, pulse sequence), we will refer to this dataset as the "**intra-scanner dataset**". An example scan and a corresponding manual segmentation are shown in Fig. 2.

The second test dataset consists of T1-weighted scans of 14 individuals with age and disease status matching those of the training dataset, acquired on a 1.5T GE Signa Scanner using an SPGR sequence with parameters: TR=35 ms, TE=5 ms, flip angle=45° and voxel size = $0.9375 \times 0.9375 \times 1.5$ mm^3 (124 coronal slices) (Han and Fischl, 2007). This dataset will be referred to as the "**cross-scanner dataset**". An example scan and a corresponding manual segmentation are shown in Fig. 3.

The third test dataset consists of multi-echo FLASH scans from 8 healthy subjects acquired on a 1.5T Siemens Sonata scanner. The acquisition parameters were: TR=20 ms, TE=min, flip angle=3°, 5°, 20° and 30°, and voxel size=1.0 mm^3 isotropic (Fischl et al., 2004b; Iglesias et al., 2012). The different flip angles correspond to different contrast properties, with the smallest angle having contrast similar to proton density (PD) weighting and the largest one having a contrast similar to T1-weighting. These data will be referred to as the "**multi-echo dataset**". A sample slice from this dataset, with flip angles 30° and 3°, is shown in Fig. 4.

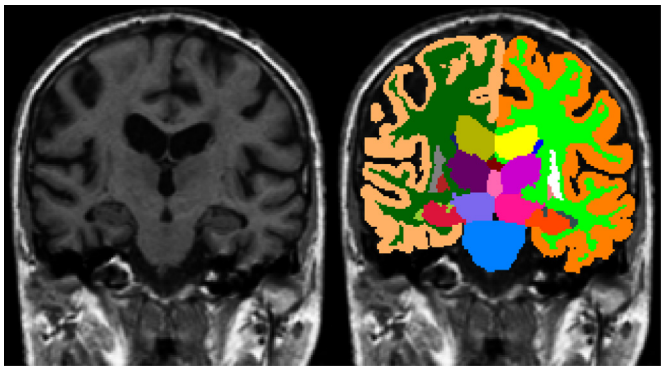

Fig. 3. On the left an example slice from the cross-scanner dataset and on the right a corresponding manual segmentation.

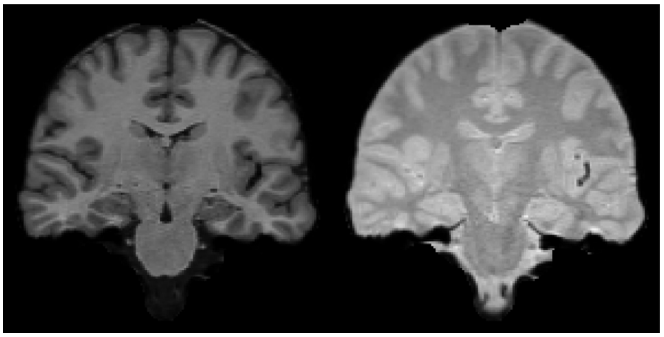

Fig. 4. An example of the T1- (flip angle=30°) and PD-weighted (flip angle=3°) scans of the same subject from the multi-echo dataset.

The fourth and final test dataset consists of 40 healthy subjects scanned at two different time points at different facilities, with scan intervals ranging from 2 days to six months, amounting to a total of 80 T1- and T2-weighted scans for the whole dataset (Holmes et al., 2012). The scans were all acquired with 3T Siemens Tim Trio scanners using identical multi-echo MPRAGE sequences for the T1 and 3D T2-SPACE sequences for the T2, with voxel size=$1.2 \times 1.2 \times 1.2$ mm^3. Note that the acquisition protocol was highly optimized for speed, with a total acquisition time for both scans of under 5 minutes. This dataset will be referred to as the "**test-retest dataset**". One of the scans had to be excluded because of motion artifacts. Moreover, some of the T2-weighted scans have minor artifacts not present in the T1-weighted scans. These scans were however included in the experiments. Manual segmentations were not available for this dataset; however, these scans are still useful in test-retest experiments quantifying the differences between the two time points. Ideally, as all the subjects are healthy, the biological variations should be small and the segmentations between the two time points should be identical. An example of the T1- and T2-weighted scans is shown in Fig. 5.

3.2. Benchmark methods

In order to gauge the performance of the proposed algorithm with respect to the state of the art in brain MRI segmentation, we compare its performance against four representative methods:

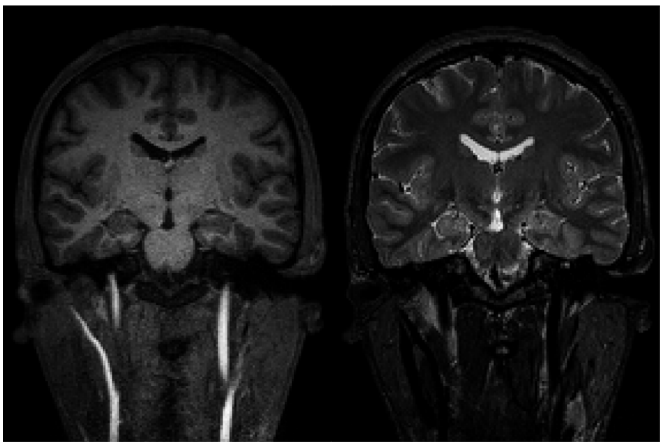

Fig. 5. An example of the T1- and T2-weighted scans of the same subject from the test-retest dataset.

- **BrainFuse**[2] (Sabuncu et al., 2010) is a multi-atlas segmentation method that uses an intensity-based label fusion approach to merge a set of propagated training labelings into a final segmentation of a target scan. More specifically, it assumes a generative model in which the joint intensity-label space is modeled with a Parzen density estimator (using a logOdds-based kernel (Pohl et al., 2006a) for the labels, and a Gaussian kernel for the intensities); with an optional Markov random field prior enforcing spatial consistency. Segmentation is carried out through Bayesian inference, effectively giving more weight to atlases that have locally similar intensities to the target scan. In the publicly available implementation, the Markov random field prior is not included – however it does not yield a significant increase in segmentation accuracy (Sabuncu et al., 2010). For computing the registrations between the training and target subjects, BrainFuse employs asymmetric bidirectional registrations based on an efficient Demons-style algorithm that uses a one parameter sub-group of diffeomorphisms combined with a sum of squared intensity differences (SSD) similarity measure (Sabuncu et al., 2010). The freely available implementation of BrainFuse is optimized to work with data that has been preprocessed (skull-stripped, bias field corrected, intensity normalized and re-sampled to a $1mm^3$ grid) with FreeSurfer. In the experiments, we follow these preprocessing requirements. The free parameters of the registration method are set to the values reported in Sabuncu et al. (2010), where the authors cross-validated the parameter values on the same training dataset that we use in this study.
- **PICSL MALF**[3] (Wang et al., 2013) assumes that the segmentation errors of the propagated training labelings can be correlated, as opposed to BrainFuse, in which independence of the errors of the different labelings is assumed. PICSL MALF formulates a weighted voting problem in terms of trying to minimize the expectation of the labeling error, i.e., the error between the fused labels and the true segmentation in every voxel. To achieve this, it approximates the expected pairwise joint label differences between the training scans and the target scan using intensity similarity information. The intensity similarities are computed within a patch around each voxel. The patch intensities are normalized to have zero mean and a constant norm, making the similarity measure robust against linear intensity change, which is often enough to correct for small differences in MRI contrast. Moreover, PICSL MALF also performs a local search to try to find the voxel that is most similar to the corresponding target image voxel patch-wise. This can be interpreted as additional refinement of the pre-computed pairwise registrations. For computing the initial pair-wise registrations between the training and target subjects PICSL MALF uses ANTs/SyN[4] (Avants et al., 2008), which is a diffeomorphic registration algorithm. We follow the implementation details that were used in the implementation of PICSL MALF that won the MICCAI 2012 Grand Challenge on Multi-Atlas Labeling (Landman and Warfield, 2012). Specifically, for computing the pair-wise registrations, we use the cross-correlation (CC) similarity metric, which adapts naturally to situations where locally varying intensities occur (Avants et al., 2008); and we set the registration parameters to the values reported in Landman and Warfield (2012). The authors use no specific preprocessing steps such as bias field correction; however, the ANTs/SyN registration algorithm has been shown to be robust to quite severe bias field effects when the CC similarity metric is used (Avants et al., 2008). We note that the PICSL MALF software also provides a post-processing procedure to correct systematic segmentation errors based on corrective learning (Wang et al., 2011); however since this is an independent module that is equally applicable to the other benchmark methods as well it was not used in this study.
- **FreeSurfer**[5] (Fischl et al., 2002) is based on a statistical atlas of neuroanatomy, along with an intensity atlas in which a Gaussian distribution is associated with each voxel and class. The parameters of these Gaussians are estimated in a supervised fashion from training data. The model is completed by a Markov random field model that ensures spatial smoothness of the segmentation, which is computed as the MAP estimate in a Bayesian framework. We note that FreeSurfer was trained on the same training data that we are using in this study, which makes direct comparison with our approach and the multi-atlas methods feasible.
- **Majority Voting** (Rohfling et al., 2004a; Heckemann et al., 2006) is a simple multi-atlas segmentation method, where the propagated training labelings are fused into a final segmentation by picking, in each voxel, the most frequent label across the propagated labelings. We include this method as a reference against which we can compare the performance of the more sophisticated label fusion approaches. For our implementation of majority voting, we use the same pair-wise registrations as for PICSL MALF.

These benchmark methods cover a wide spectrum of modern brain MRI segmentation algorithms. Majority voting, BrainFuse and PICSL MALF represent multi-atlas segmentation, which is arguably the most popular segmentation paradigm at the moment. Moreover, they are non-parametric methods, whereas our method and FreeSurfer represent parametric approaches.

3.3. Cross-validation experiments on training data for parameter tuning

The free parameters of the different methods are determined using the 39-subject training dataset as follows:

Proposed Algorithm. We use 20 randomly picked subjects out of the available 39 to build our probabilistic atlas. Only 20 subject are chosen, because the atlas building process is very computationally expensive (several weeks to build an atlas with 20 subjects) and the results show that the segmentation performance does not increase any further when more subjects are added (see Section 4.3). The remaining 19 subjects are used to find suitable values for the free parameters in our algorithm: the global stiffness of the mesh β, the number of bias field basis functions P, the groups of structures f that share the same GMM parameters, and the number of mixture components associated with each structure group. The parameters are tuned based on a visual inspection of the automatic segmentations in the 19 training subjects. The chosen values for the mesh stiffness and number of bias field basis functions are: $\beta = 0.1$ and $P=5$ per dimension, amounting to a total of $P = 5^3 = 125$ basis functions in 3D. The choice of which sets of structures share the Gaussian mixture parameters, as well as the number of Gaussians for each mixture, is summarized in Table 2.

BrainFuse. We use the optimal parameters listed in the original publication (Sabuncu et al., 2010); this choice is appropriate because the authors cross-validate the parameter values on the same training dataset as used in this study.

PICSL MALF. For this method we need to determine the optimal values for the patch radius over which the intensity similarity is

[2] http://people.csail.mit.edu/msabuncu/sw/bfl/index.html.
[3] http://www.nitrc.org/projects/picsl_malf/.
[4] http://stnava.github.io/ANTs/.
[5] http://surfer.nmr.mgh.harvard.edu/.

Table 2
Details of the parameter sharing between structure classes. The groups of structures that share their Gaussian mixture parameters are shown in the first column, and the corresponding amount of Gaussians in the mixture in the second column.

Structures with shared parameters	Number of Gaussians
Non-brain tissues	3
L/R Cerebral White Matter (WM) L/R Cerebellum White Matter (CWM) Brain Stem (BS) L/R Ventral Diencephalon Optic Chiasm	2
L/R Cerebral Cortex (CT) L/R Cerebellum Cortex (CCT) L/R Caudate (CA) L/R Hippocampus (HP) L/R Amygdala (AM) L/R Accumbens Area	3
L/R Lateral Ventricle (LV) L/R Inferior Lateral Ventricle 3rd Ventricle Cerebro-Spinal Fluid (CSF) 5th Ventricle 4th Ventricle Vessel L/R Choroid Plexus	3
L/R Thalamus (TH)	2
L/R Putamen (PU)	2
L/R Pallidum (PA)	2

calculated; a constant controlling the inverse distance function which maps the intensity difference to the joint error; and the size of the local search window (Wang et al., 2013). For this purpose, we randomly select 10 subjects as test data and use the remaining 29 subjects as training data, and perform a cross-validation grid search using similarity patch radii of $r_p = [1, 2, 3]$, local search radii of $r_s = [0, 1, 2, 3]$ and inverse mapping constants of $\beta = [0.5, 1, 1.5, 3, 6]$. As a measure of goodness we use the mean Dice overlap score[6] (which is the main performance metric used in the experiments below) over the structures listed in Section 3.4 below. The resulting optimal values are: $r_p = 1$, $r_s = 2$ and $\beta = 3$.

FreeSurfer. We use the standard processing pipeline with default parameters. No cross-validation needs to be performed as FreeSurfer is trained on the same training dataset (using all 39 subjects) we use in this study.

Majority Voting. Given the pre-computed registrations, majority voting has no parameters to tune.

3.4. Experimental setup

We perform a comprehensive evaluation consisting of four sets of experiments:

I. In a first experiment, we use models trained on the training dataset to segment the scans from the intra-scanner and the cross-scanner datasets, comparing each method's segmentations with the corresponding manual annotations using the Dice overlap score. This experiment enables us not only to compare the performance of the different methods, but also to assess how much their performance degrades when the image intensity properties of the training and test datasets are not matched.
II. In a second experiment, we evaluate the computational efficiency of the various methods on the intra- and inter-scanner datasets. We compute the running time of the different algorithms on a cluster where each node has two quad-core Xeon 5472 3.0 GHz CPUs and 32 GB of RAM; we only use one core in the experiments in order to make fair comparisons, even though all the algorithms can potentially be parallelized. We also record the execution time of a multi-threaded implementation of our method, using 8 cores on a computer with 8 dual-cores with 3.4 GHz CPU and 64 GB of RAM. This setup represents a realistic scenario that enables us to compare the running time of our algorithm with those reported by other studies in the literature.
III. In a third experiment, we study the effect of the number of training subjects on the segmentation performance. To achieve accurate segmentations, a representative training set is needed to capture all the structural variation one might see within the subjects to be segmented (Aljabar and Heckemann, 2009). However, some algorithms require less training data than others to approach their asymptotic performance, which represents a saving in manual labeling effort. We therefore randomly pick 5 sets of 5, 10 and 15 subjects from the training data, and re-evaluate the segmentation performance of the proposed method, BrainFuse, PICSL MALF and majority voting on the intra- and cross-scanner datasets.
IV. In a final experiment, we evaluate the ability of the proposed algorithm to segment non-T1-contrast and multi-contrast MR scans using the multi-echo and the test-retest datasets. Given a training set consisting only of T1-weighted scans, using multi-contrast or non-T1-contrast information is out of reach for the four specific benchmark methods we compare against in this article, although we note that several multi-atlas label fusion techniques exist that could potentially be used in this context (cf. discussion in Section 5). For the multi-echo dataset we first run the proposed method using only the T1-weighted images (i.e., flip angle 30°), then only the PD-weighted images (i.e., flip angle 3°), and finally using both the T1- and PD-weighted images simultaneously. The resulting automated segmentations are then compared to the expert segmentations using Dice scores. For the test-retest dataset, we first segment the two time points using only the T1-weighted images, and subsequently using both T1- and T2-weighted images together. Because no manual segmentations are available for this dataset, we use absolute symmetrized percent change (ASPC) (Reuter et al., 2012) to quantify the differences in the automatic segmentations between the two time points. This metric is defined as the absolute value of the difference in volume, normalized by the mean volume:

$$\mathrm{ASPC} = \frac{2|V_2 - V_1|}{V_1 + V_2},$$

where V_1,V_2 are the volumes at the two time points. Ideally this number should be small, as the subjects are all healthy and the time between the scans is not so long.

We report the Dice scores and the ASPC on a representative subset of 23 relevant structures that is also used in other studies (e.g., Fischl et al., 2002; Sabuncu et al., 2010): left and right cerebral white matter (WM), cerebellum white matter (CWM), cerebral cortex (CT), cerebellum cortex (CCT), lateral ventricle (LV), hippocampus (HP), thalamus (TH), putamen (PU), pallidum (PA), caudate

[6] Dice = $2|\mathbf{l}_A \cap \mathbf{l}_M|/(|\mathbf{l}_A| + |\mathbf{l}_M|)$, where $\mathbf{l}_A$ and $\mathbf{l}_M$ are the automatic and manual segmentations respectively and $|\cdot|$ is the cardinality of a set.

(CA), amygdala (AM) and brain stem (BS). We will refer to these structures as the "regions of interest" (ROIs); note that for clarity of presentation we report the average Dice score of the left and right hemisphere for all structures except for the brain stem.

4. Results

4.1. Intra-scanner and cross-scanner segmentation performance

The Dice scores between the manual and automated segmentations of the ROIs, obtained using the different methods, are shown for the intra-scanner dataset in Fig. 6 (top). Table 3 (first column) summarizes the scores in average over the ROIs and subjects, and reports statistically significant differences between the methods. The significance testing was done using paired, two-sided t-tests, by stacking the individual Dice scores in each ROI and subject for a given method. Corresponding scores and significant differences for each ROI separately are reported in Supplementary material, Table 1. All of the methods perform well on the intra-scanner dataset, which was expected, as the contrast properties of the training data are identical to those of this dataset. The multi-atlas segmentation methods achieve the highest mean scores, with PICSL MALF being the best method for this dataset. Majority voting also obtains a very high mean score despite its simple fusion strategy. This is likely due to the accurate ANTs/SyN registration framework, which has been shown to perform very well on intra-scanner data (Klein et al., 2009). We note that each of the benchmark methods is specifically trained for this type of data, whereas the proposed method is not.

For the cross-scanner data, where the contrast properties of the target data are different from the training data, the ROI Dice scores are shown in Fig. 6 (bottom) and the mean scores over the ROIs and subjects in Table 3 (third column). Corresponding scores and significant differences for each ROI separately are reported in Supplementary material, Table 2. Compared to the intra-scanner dataset, the overall segmentation accuracy of all methods decreases, which is likely due to the lower intrinsic image contrast of the SPGR pulse sequence as noted in Han and Fischl (2007) and as also visible from Fig. 3. In this dataset, the proposed method

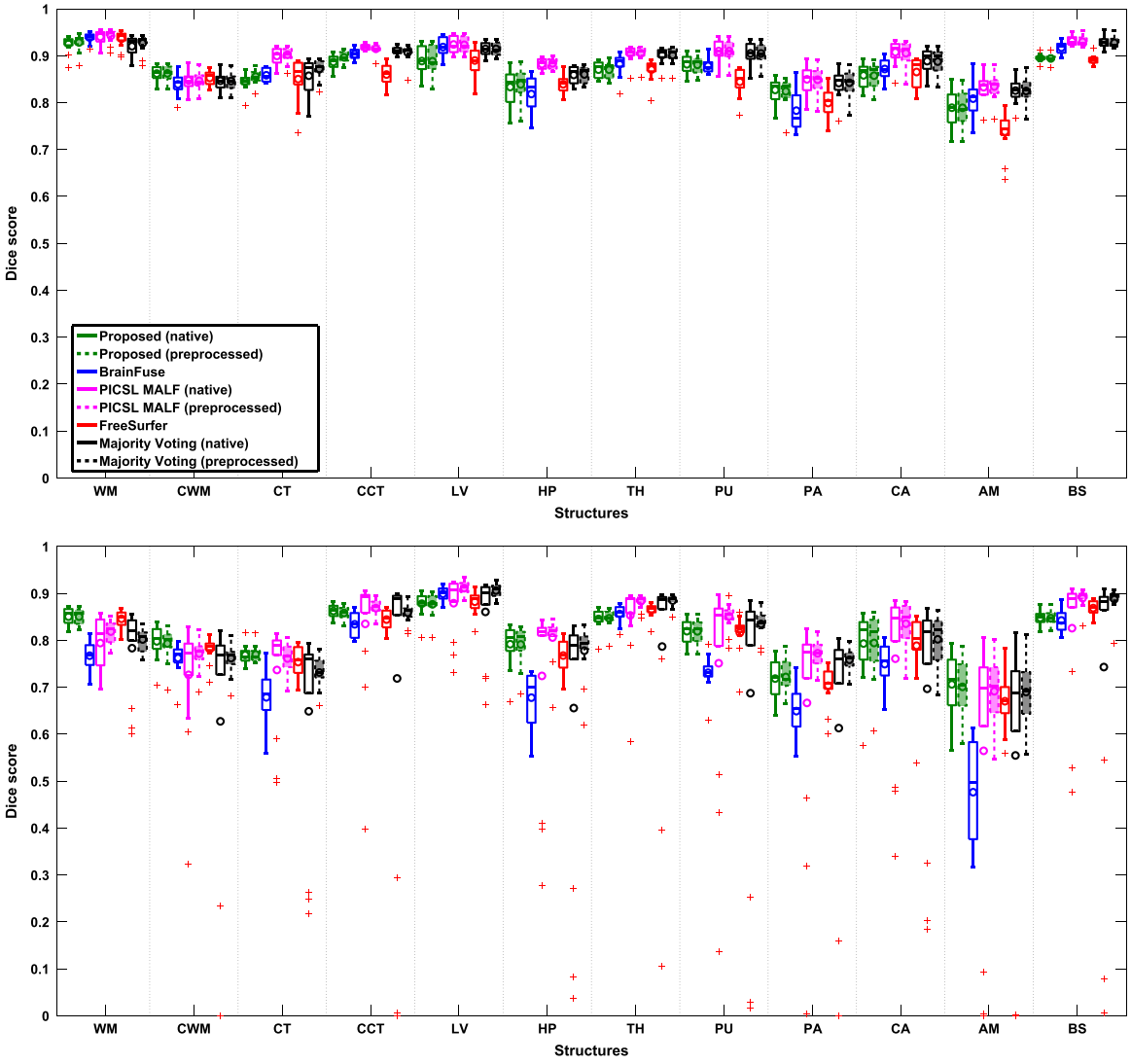

Fig. 6. The Dice scores of the different methods for the intra-scanner (top) and cross-scanner (bottom) data. The proposed method=green, BrainFuse=blue, PICSL MALF=magenta, FreeSurfer=red and Majority Voting=black. Additional results, obtained by preprocessing the input data using the FreeSurfer pipeline, are also shown (filled boxes with broken lines). On each box, the central horizontal line is the median, the circle is the mean, and the edges of the box are the 25th and 75th percentiles. Data points falling outside of the range covered by scaling the box four times are considered outliers, and are plotted individually. The whiskers extend to the most extreme data points that are not considered outliers. See Section 3.4 for the acronyms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Mean Dice scores of the different methods over the ROIs for the intra-scanner (first column) and cross-scanner (third column) datasets. Additional results, obtained by preprocessing the input data using the FreeSurfer pipeline, are also shown (in italics, second and fourth columns). The superscript lists the methods that obtain significantly lower scores compared to a given method. The significance was tested using a paired two-sided t-test with a 5% significance level.

	Intra-scanner data		Cross-scanner data	
Method	Native	*Preprocessed*	Native	*Preprocessed*
Proposed (P)	0. 863^{FS}	0. *865*^{FS}	**0. 807**^{BF,PM,FS,MV}	0. *806*^{BF,FS}
BrainFuse (BF)	0. 868^{FS}	0. *868*^{FS}	0. 744^{MV}	*0.744*
PICSL MALF (PM)	**0. 895**^{P,BF,FS,MV}	**0. *895***^{P,BF,FS,MV}	0. 760^{BF,MV}	**0. *822***^{P,BF,FS,MV}
FreeSurfer (FS)	0.853	*0.853*	0. 799^{BF,PM,MV}	0. *799*^{BF}
Majority Voting (MV)	0. 883^{P,BF,FS}	0. *885*^{P,BF,FS}	0.698	0. *808*^{BF,FS}

achieves the highest mean score, demonstrating its robustness against changes in contrast. Although FreeSurfer explicitly encodes the contrast properties of the training scans, its performs relatively well on this un-matched data; this can be explained by its in-built renormalization procedure for T1 acquisitions, which applies a multi-linear atlas-image registration and a histogram matching step to update the class-conditional densities for each structure (Han and Fischl, 2007). In contrast, the label fusion methods, which directly rely on the image intensities of the training data in their registration and fusion steps, are clearly affected by the changes in the MRI contrast. The pair-wise registrations are especially more challenging for this dataset, leading to mis-registrations that are the principal error source in multi-atlas segmentation.

The segmentation accuracy of BrainFuse, which uses FreeSurfer-preprocessed images, varies between different structures. In general it seems to perform well on some of the larger structures (LV, BS), whereas the performance is not so good on some of the smaller structures (AM, HP, PA). This is likely explained by the choice of registration algorithm and especially the SSD similarity measure, which is not invariant against small intensity changes. Although the PICSL MALF and majority voting methods use the more robust CC similarity measure in the ANTs/SyN registration framework, there are some subjects in the cross-scanner dataset for which computing the registrations without preprocessing is very difficult, resulting in the segmentation outliers shown in Fig. 6. We note that although majority voting does not rely on intensity information when fusing the labels, its usage of CC in the registration step indirectly assumes that the training and target scans have similar properties (linear local intensity transformation). Compared to PICSL MALF, its simple fusion rule makes majority voting much more dependent on the quality of the pair-wise registrations, as the effect of poorly registered subjects can not be downplayed.

In order to further analyze the relative performance of the various methods without the influence of outlier subjects in majority voting and PICSL MALF, we performed an additional, *post hoc* analysis with the explicit aim of avoiding ANTs/SyN registration failures in the cross-scanner data. For this purpose, we re-ran majority voting and PICSL MALF, as well as the proposed method, on data that had been preprocessed with the FreeSurfer pipeline (which includes skull-stripping, bias field correction, intensity normalization, and re-sampling to a 1mm^3 grid), although we note that this preprocessing is not part of the default implementations of these algorithms.

The resulting average Dice scores for the intra- and the cross-scanner data are shown (in italics) in the second and the fourth column of Table 3 – note that BrainFuse already depends on FreeSurfer so that all five methods effectively use the same preprocessing pipeline in this scenario. The Dice scores obtained this way for each ROI individually are also displayed in Fig. 6. Comparing the results obtained with and without FreeSurfer preprocessing, it can be seen that the additional preprocessing effectively avoids ANTs/SyN registration failures in the cross-scanner data, resulting in a strong performance for majority voting with only a relatively minor improvement for the more advanced label fusion of PICSL MALF, which obtains the strongest overall segmentation accuracy. Unlike in the intra-scanner data, however, majority voting no longer outperforms the proposed method in the cross-scanner data even though all its pair-wise registrations are successful. It can also be seen that the proposed method does not benefit from FreeSurfer preprocessing in either the intra-scanner or the cross-scanner data, an indirect demonstration of its intrinsic bias field correction and skull stripping performance.

Finally, Table 3 and Table 4 of the Supplementary material list the Dice scores and significant differences for each ROI separately for the intra- and cross-scanner data preprocessed with FreeSurfer. It can be seen that, although after preprocessing PICSL MALF outperforms other techniques in seven structures on the cross-scanner data, the proposed technique remains the best method for four other structures, especially those in cortical areas (WM, CT, CWM).

A limitation of the comparisons presented in this Section is that the proposed method uses an atlas built from 20 randomly selected subjects, potentially introducing a bias when comparing to benchmark methods that use all 39 subjects without selection as training set. However, as shown in Section 4.3, PICSL MALF, BrainFuse and majority voting all benefit from using all the available training subjects compared to random subsets of various sizes, whereas the proposed method saturates around 10 subjects with very little further gains from larger training sets.

4.2. Running time

The approximate mean computation time for a single scan using the different methods is shown in Table 4. The proposed method is approximately 7 times faster than FreeSurfer, 12 times faster than BrainFuse and 100 times faster than PICSL MALF and majority voting.

In general, the parametric methods (i.e., FreeSurfer and the proposed method) are significantly faster than the label fusion approaches. This is because only a single non-linear registration is needed, as opposed to the multiple pair-wise registrations used in the non-parametric methods. Moreover, in PICSL MALF the local search is especially time consuming with large search windows. Compared with FreeSurfer, which is also parametric, our method is faster due to the sparse encoding of the mesh prior. Encoding this sparsity is computationally expensive, but needs to be done only once (in an offline fashion). Furthermore, in the proposed approach, no special post or preprocessing of the target scans is needed.

In its multi-threaded setup, the proposed method has an execution time of 23.5 min per scan on average. The fastest whole-brain segmentation method to our knowledge is presented in Zikic

Table 4
Mean computational time for the different methods (single core). For label fusion methods the computation times for registration (Reg.) and label fusion (Fusion) are listed separately.

	Mean time per subject (single core)		
Method	Reg.	Fusion	Full time
BrainFuse	16 h	1 h	17 h
Majority voting	143.9 h	0.1 h	144 h
PICSL MALF	143.9 h	3.8 h	147.7 h
FreeSurfer	–	–	9.5 h
Proposed	–	–	**1.4 h**

et al. (2014) with execution times in the range of 5 to 13 minutes; however this method is not designed to handle image contrast differences.

4.3. Effect of the number of training subjects

Fig. 7 shows the effect on each method's Dice scores, averaged across all ROIs, of training on randomly selected subsets of the entire training pool, both for the intra-scanner and the cross-scanner datasets. In order to compare the different methods' performance without the influence of gross registration failures in majority voting and PICSL MALF, results obtained after preprocessing the data with FreeSurfer are also provided. The figure shows that adding more training subjects generally yields more accurate segmentations for all methods, but that the proposed method reaches its maximum performance faster than the multi-atlas methods: Already with 10 training subjects the segmentation accuracy of the proposed method is above 99% of its maximal performance in all experiments, regardless of the specific subjects included in the training set. This is especially useful for populations where expert segmentations are expensive or difficult to obtain, such as infants. The fact that the performance of the proposed method is not more dependent on the specific subjects included in the training set is likely due to the atlas construction process that explicitly avoids over-fitting to training data (Van Leemput, 2009), yielding sparser tetrahedral meshes (and therefore blurrier probabilistic atlases) when fewer training subjects are available. This effect is illustrated in Table 5, where the average number of mesh vertices for the 5, 10 and 15 training subject groups are reported.

Table 5
Average number of vertices in the proposed atlas mesh for different numbers of training subjects.

Number of subjects	Average number of vertices
5	33,606
10	44,614
15	51,258

The effect of FreeSurfer preprocessing appears to be minimal for the proposed method across the different training set sizes, showing a similar performance in both the intra-scanner and the cross-scanner data compared to when no preprocessing is applied. In contrast, preprocessing is crucial for both majority voting and PICSL MALF in the cross-scanner setting, as ANTs/SyN registration failures otherwise severely compromise segmentation performance. Compared to the other multi-atlas methods working on the same (i.e., preprocessed) data, as well as the proposed method (with or without preprocessing), BrainFuse appears to be much more sensitive to small training datasets, both in terms of the average Dice scores that it obtains as well as its sensitivity to the specific random subjects that are used for training.

4.4. Multi-contrast performance

Fig. 8 shows the Dice scores of the proposed method on the multi-echo dataset, for various combinations of single- (T1-weighted only or PD-weighted only) and multi-contrast (T1- and PD-weighted simultaneously) input data. The results are very similar between T1-weighted only and multi-contrast input data, whereas using the PD-weighted contrast alone often yields reduced performance. This indicates that the PD-weighted contrast does not add much useful information to the T1-weighted scan when healthy brains are segmented. Example segmentations of the multi-echo dataset using T1-weighted only and multi-contrast scans are shown in Fig. 9.

The volume differences between the two time points in the 39 subjects of the T1/T2 test-retest dataset are shown in Fig. 10. In general, they are quite similar and small for both single- (only T1) and multi-contrast (both T1 and T2) segmentations, with the median ASPC in the 1–2% range. There are some larger differences

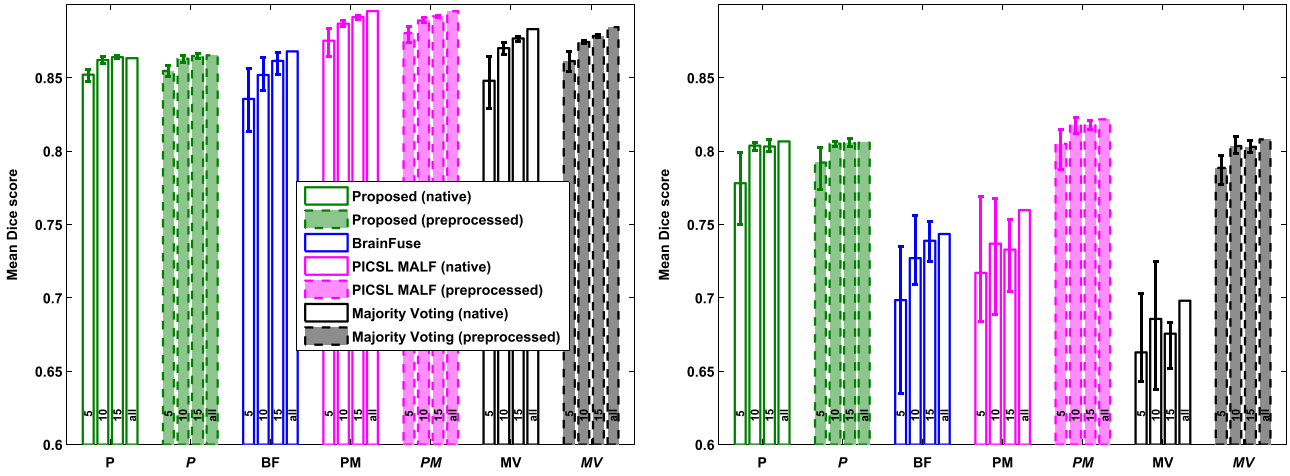

Fig. 7. Mean Dice scores over the ROIs for the intra-scanner (left) and the cross-scanner (right) data when the different methods are trained using randomly picked subsets of only 5, 10 and 15 training subjects. The error bars correspond to the lowest and highest obtained mean Dice score across the random subsets. The score obtained when all subjects in the training pool are used is also shown for reference (fourth bar of each method). The proposed method (P) is shown in green, BrainFuse (BF) in blue, PICSL MALF (PM) in magenta and majority voting (MV) in black. Additional results, obtained by preprocessing the input data using the FreeSurfer pipeline, are also shown (filled bars with broken lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

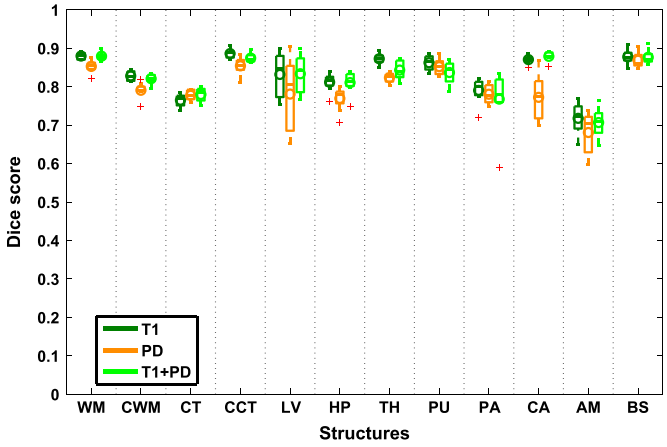

Fig. 8. Dice scores for the multi-echo dataset. Performance on T1-weighted data is shown in dark green, on PD-weighted data in orange, and on multi-contrast input data in light green. The box plots are drawn in the same way as explained in Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

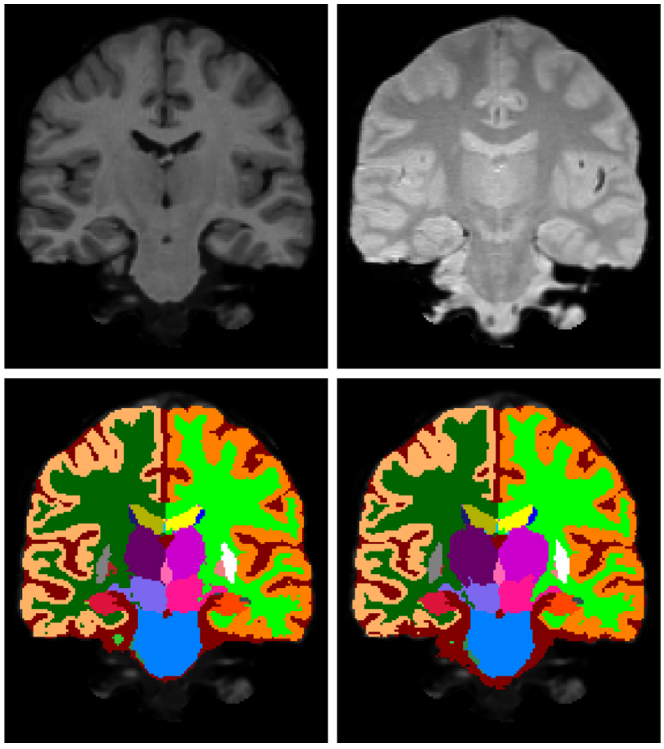

Fig. 9. Top row: target scans, T1-weighted on the left and PD-weighted on the right. Bottom row: automatic segmentation using **only** the T1-weighted scan on the left, automatic segmentation using **both** scans on the right.

– especially in the thalamus and pallidum – when using multi-contrast data. This appears to be mostly due to imaging artifacts in the T2-scans, an example of which is shown in Fig. 11. We note that this dataset has the lowest resolution of all the datasets we tested the method on, and therefore is affected the most by partial volume segmentation errors.

In order to put the ASPC test/retest results of Fig. 10 in perspective, we also report the ASPC scores for the benchmark methods when applied to the T1-weighted scans of the two time points. Because of the heavy computational burden of some of the methods (e.g., PICSL MALF occupies a CPU core for more than six days per scan, cf. Table 4), we only report the results on 10 randomly chosen subjects (20 scans in total) out of the available 39.

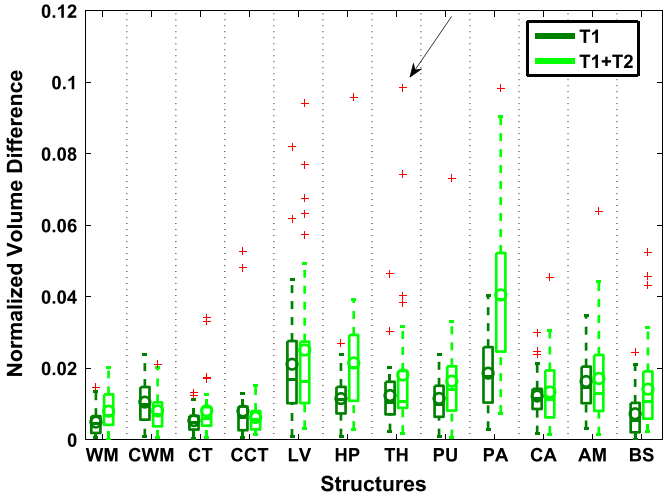

Fig. 10. The ASPC scores for the test-retest dataset. Volume differences between the time points on multi-contrast input data is shown in light green, and on T1-weighted data only in dark green. The box plots are drawn in the same way as explained in Fig. 6. The outlier marked by an arrow is the one shown in Fig. 11. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The benchmark methods' ASPC scores are shown in Fig. 12, along with those obtained with the proposed method on the same subjects (both T1-only and multi-contrast). The figure shows that the proposed method, PICSL MALF and majority voting perform most reliably across the time points, while BrainFuse and FreeSurfer have more variance in their segmentations. As discussed before, the weaker performance of BrainFuse compared to the other label fusion methods is likely a combination of the chosen registration framework and sub-optimal similarity measure used for the registrations; the reasons for FreeSurfer's weaker performance are not immediately clear. On the selected 10 subjects we did not observe problems with the pair-wise registrations when using the ANTs/SyN registration framework, leading to a robust performance of the PICSL MALF and majority voting methods in this experiment.

5. Discussion and conclusion

In this paper we have validated a whole-brain segmentation method that builds upon the parametric, unsupervised intensity clustering models commonly used in tissue classification. We have demonstrated that these type of models are capable of achieving state-of-the-art segmentation performance, while being very fast, adaptive to changes in tissue contrast, and able to handle multi-contrast data. We emphasize that the exact same algorithm was used for all datasets in this paper, without any parameter retuning or configuration changes, demonstrating the robustness of the approach.

Our experiments indicate that, in the general cross-scanner scenario, the proposed method yields a robust segmentation performance on par with the very best competitors, while being orders of magnitude faster and without requiring any form of preprocessing. The method's accuracy is outperformed only when the image intensities of the training and test data are perfectly matched; however we believe this scenario will seldom occur in practice because manual whole-brain segmentation is so time-consuming (e.g., taking hundreds of days for the training data used in this paper) that the available training data will seldom be acquired on the exact same imaging system as the images being segmented.

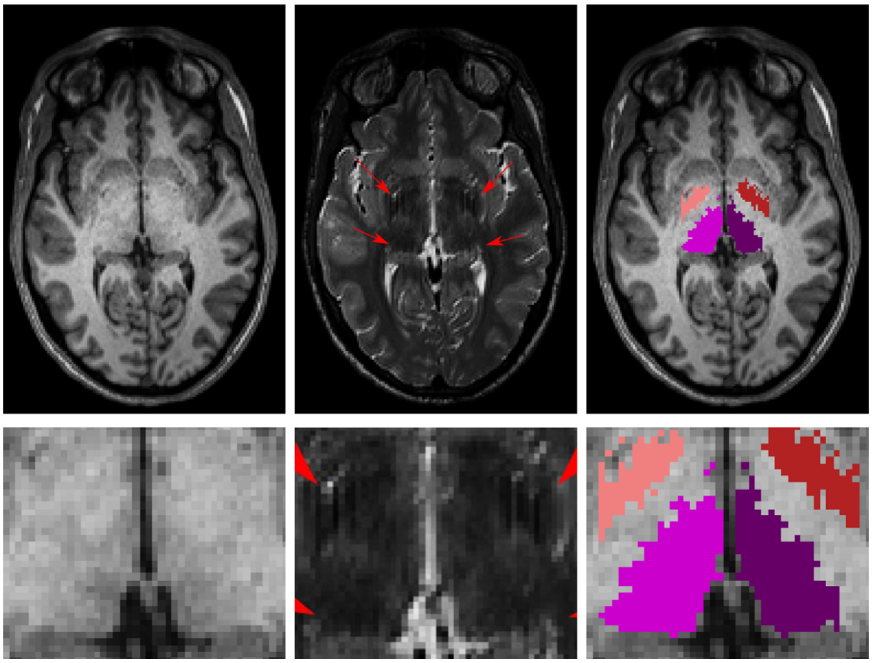

Fig. 11. An example of an outlier subject marked by the arrow in Fig. 10. Top row from left to right: a T1-weighted scan with no visible artifacts, a T2-weighted scan with a line-like artifact in the pallidum and thalamus area marked by red arrows, and an automated segmentation of pallidum and thalamus showing the segmentation error caused by the artifact. The bottom row shows zoomed figures of the affected area, highlighting vertical lines in the T2-scan that cause jagged borders in the automatic segmentation, resulting in a poor ASPC score for this subject. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

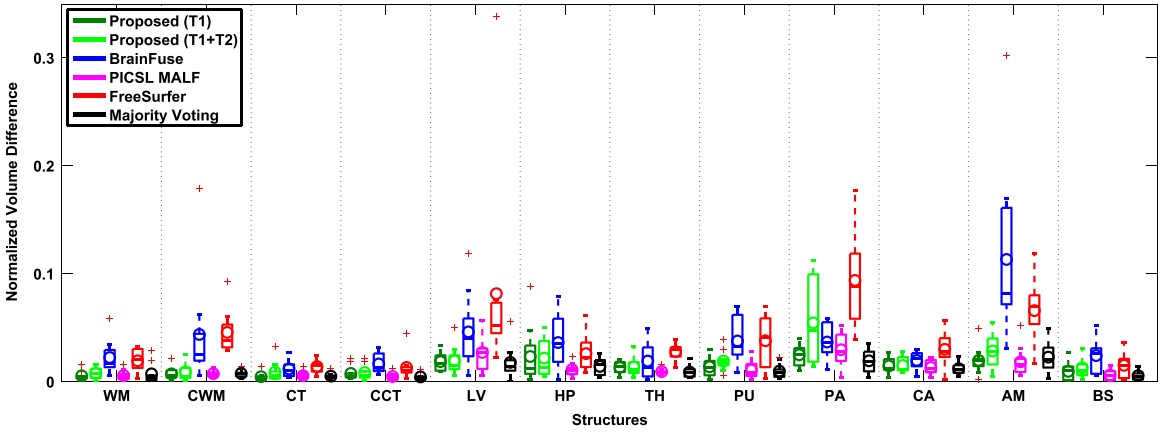

Fig. 12. The ASPC scores of the different methods for 10 randomly chosen subjects from the test-retest dataset. The performance of the proposed method when using only T1-weighted data in dark green and when using both T1- and T2-weighted scans in light green, BrainFuse in blue, PICSL MALF in magenta, FreeSurfer in red and Majority Voting in black. The box plots are drawn in the same way as explained in Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Since the method we have validated here combines Gaussian mixture modeling with MRI bias field correction and probabilistic atlas deformation, it is closely related to the unified segmentation framework described in Ashburner and Friston (2005); however only basic tissue classification on T1-weighted images was attempted in that work. A related method based on fuzzy c-means clustering and a topological atlas was described in Bazin and Pham (2008), but that only segmented a handful of structures, and relied on the availability of pre-defined centroid initializations for each type of MRI sequence the method is expected to encounter.

An early attempt at whole-brain segmentation using a deformable probabilistic atlas combined with unsupervised intensity clustering was described in Babalola et al. (2009); however, the atlas registration was performed independently of the segmentation process, using relatively coarse deformations, and the resulting segmentation performance was found to trail that of label fusion methods. Subsequent methods showing better performance (Ledig et al., 2012a, 2015; Makropoulos et al., 2014; Iglesias et al., 2013b; Tang et al., 2013) have used the non-parametric paradigm instead, where a probabilistic atlas is computed in the space of the target scan, i.e., after warping each of the training scans onto the target image using pairwise registration. We note that well-known majority voting methods (Rohfling et al., 2004a; Heckemann et al., 2006) using mutual information (Maes et al., 1997; Wells et al., 1996b; Studholme et al., 1999) as registration criterion also implicitly combine the non-parametric paradigm (multi-atlas label fusion) with unsupervised intensity clustering, since mutual information-based registration can be understood as jointly estimating registration parameters and class-conditional densities (Roche et al., 2000). In general, however, such non-parametric approaches are computationally much more expensive than the parametric method we evaluated here.

In the current paper, we only analyzed images of healthy subjects, and our experiments on multi-contrast images showed no benefit in terms of segmentation accuracy compared to when only T1-weighted scans are used. However, the ability to seamlessly handle multi-contrast data becomes essential when analyzing diseased populations, since many brain lesions are much better visualized in T2-weighted and FLAIR scans than in T1-weighted contrast. In future work we will therefore include models of pathologies in the proposed framework, enabling simultaneous whole-brain segmentation and pathology detection (Puonti and Van Leemput, 2016).

The proposed method has been evaluated on a set of structures in which the cerebral cortex was considered a single structure, without attempting to further parcellate it into neuroanatomical subregions. However, we note that the volumetric white matter segmentations generated by the method can be used to build and label cortical surface models using FreeSurfer (Dale et al., 1999; Fischl et al., 2004a). Exploring this direction remains as future work.

The segmentation software used in this paper, including the source code, the sparse probabilistic atlases and the code to build such atlases from training data, will be made publicly available.

Acknowledgements

This research was supported by the NIH NCRR (P41-RR14075, 1S10RR023043), NIBIB (R01EB013565), the Lundbeck foundation (R141-2013-13117) and financial contributions from the Technical University of Denmark. JEI acknowledges financial support from the Gipuzkoako Foru Aldundia (Fellows Gipuzkoa Program), the European Union's Horizon 2020 Research and innovation program under the Marie Sklodowska-Curie grant agreement No 654911, as well as from the Spanish Ministry of Economy and Competetiveness (MINECO, TEC2014-51882-P).

Appendix A. Supplementary material

Supplementary material associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neuroimage.2016.09.011.

References

Aljabar, P., Heckemann, A.R., Hammers, A., Hajnal, V.J., Rueckert, D., 2009. Multi-atlas based segmentation of brain images atlas selection and its effect on accuracy. NeuroImage 46, 726–738.
Artaechevarria, X., Munõz Barrutia, A., Ortiz-de Solórzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imaging 28, 1266–1277.
Ashburner, J., Andersson, R.L.J., Friston, J.K., 2000. Image registration using a symmetric prior-in three dimensions. Hum. Brain Mapp. 9, 212–225.
Ashburner, J., Friston, J.K., 1997. Multimodal image coregistration and partitioning – a unified framework. NeuroImage 6, 209–217.
Ashburner, J., Friston, J.K., 2005. Unified segmentation. NeuroImage 26, 839–885.
Asman, J.A., Landman, A.B., 2013. Non-local statistical label fusion for multi-atlas segmentation. Med. Image Anal. 17, 194–208.
Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation:evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41.
Awate, S.P., Tasdizen, T., Whitaker, R.T., Foster, N., 2006. Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification. Med. Image Anal. 10, 726–739.
Babalola, K.O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M., Rueckert, D., 2009. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. NeuroImage 47, 1435–1447.
Bazin, P.L., Pham, D.L., 2008. Homeomorphic brain image segmentation with topological and statistical atlases. Med. Image Anal. 12, 616–625.
Buckner, R., Head, D., Parker, J., Fotenos, A., Marcus, D., Morris, J., Snyder, A., 2004. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volumee. NeuroImage 23, 724–738.
Caviness, V.S., Meyer, J., Makris, N., Kennedy, D.N., 1996. MRI-based topographic parcellation of human neocortex: ananatomically specified method with estimate of reliability. J. Cogn. Neurosci. 8, 566–587.
Caviness Jr., V., Filipek, P., Kennedy, D., 1989. Magnetic resonance technology in human brain science: blueprint for a program based upon morphometry. Brain Dev. 11, 1–13.
Cootes, F.T., Edwards, J.G., Taylor, J.C., 1998. Active appearance models, In: Proceedings of the 5th European Conference on Computer Vision, volume II, pp. 484–498.
Coupé, P., Manjón, V.J., Fonov, V., Pruessner, J., Robles, M., Collins, L.D., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54, 940–954.
D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., 2004. Non-rigid atlas-to-image registration by minimization of class-conditional image entropy. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2004, pp. 745–753.
D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., 2006. A unified framework for atlas based brain image segmentation and registration. In: Biomedical Image Registration, vol. 4057, pp. 136–143.
Dale, M.A., Fischl, B., Sereno, I.M., 1999. Cortical surface-based analysis Isegmentation and surface reconstruction. NeuroImage 9, 179–194.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39, 1–38.
Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, H.D., Busa, E., Seidman, J.L., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, M.A., 2004a. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14, 11–22.
Fischl, B., Salat, D.H., van der Kouwe, A.J.W., Makris, N., Ségonne, F., Quinn, B.T., Dale, A.M., 2004b. Sequence-independent segmentation of magnetic resonance images. NeuroImage 23, S69–S84.
Fischl, B., Salat, H.D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, M.A., 2002. Whole brain segmentation automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.
Han, X., Fischl, B., 2007. Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE Trans. Med. Imaging 26, 479–486.
Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115–126.
Holmes, A.J., Lee, P.H., Hollinshead, M.O., Bakst, L., Roffman, J.L., Smoller, J.W., Buckner, R.L., 2012. Individual differences in amygdala-medial prefrontal anatomy link negative affect, impaired social functioning, and polygenic depression risk. J. Neurosci. 32, 18087–18100.
Iglesias, J.E., Augustinack, J.C., Nguyen, K., Player, C.M., Player, A., Wright, M., Roy, N., Frosch, M.P., McKee, A.C., Wald, L.L., Fischl, B., Van Leemput, K., 2015a. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution mri: application to adaptive segmentation of in vivo mri. NeuroImage 115, 117–137.
Iglesias, J.E., Konukoglu, E., Zikic, D., Glocker, B., Van Leemput, K., Fischl, B., 2013a. Is synthesizing MRI contrast useful for inter-modality analysis? In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, pp. 631–638.
Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images a survey. Med. Image Anal. 24, 205–219.
Iglesias, J.E., Sabuncu, R.M., Van Leemput, K., 2012. A generative model for multi-atlas segmentation across modalities. In: Proceedings of the 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 888–891.
Iglesias, J.E., Sabuncu, R.M., Van Leemput, K., 2013b. A unified framework for cross-modality multi-atlas segmentation of brain MRI. Med. Image Anal. 17, 1181–1191.
Iglesias, J.E., Van Leemput, K., Bhatt, P., Casillas, C., Dutt, S., Schuff, N., Truran-Sacrey, D., Boxer, A., Fischl, B., 2015b. Bayesian segmentation of brainstem structures in mri. NeuroImage 113, 184–195.
Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M.A., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion – application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imaging 28, 1000–1010.
Kelemen, A., Székely, G., Gerig, G., 1998. Three-dimensional model-based segmentation of brain MRI. In: Workshop on Biomedical Image Analysis, pp. 4–13.
Kennedy, D., Filipek, P., Caviness, V., 1989. Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging. IEEE Trans. Med. Imaging 8, 1–7.
Klein, A., Andersson, J., Ardekani, A.B., Ashburner, J., Avants, B., Chiang, M.C., Christensen, E.G., Collins, L.D., Gee, J., Hellier, P., Song, H.J., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, P.R., Mann, J.J., Parsey, V.R., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786–802.
Landman, A.B., Warfield, K.S., 2012. Miccai 2012 workshop on multi-atlas labeling. In: Proceedings of the 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2012), pp. 91–95.
Ledig, C., Heckemann, A.R., Aljabar, P., Wolz, R., Hajnal, V.J., Hammers, A., Rueckert, D., 2012a. Segmentation of MRI brain scans using MALP-EM. In: MICCAI 2012

Grand Challenge and Workshop on Multi-Atlas Labeling, pp. 79–82.
Ledig, C., Heckemann, R.A., Hammers, A., Lopez, J.C., Newcombe, V.F.J., Makropoulos, A., Lötjönen, J., Menon, D.K., Rueckert, D., 2015. Robust whole-brain segmentation application to traumatic brain injury. Med. Image Anal. 21, 40–58.
Ledig, C., Wolz, R., Aljabar, P., Lötjönen, J., Heckemann, R.A., Hammers, A., Rueckert, D., 2012b. Multi-class brain segmentation using atlas propagation and EM-based refinement. In: Proceedings of the 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 896–899.
Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imaging 16, 187–198.
Makropoulos, A., Gousias, I.S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J.V., Edwards, A. D., Counsell, S.J., Rueckert, D., 2014. Automatic whole brain MRI segmentation of the developing neonatal brain. IEEE Trans. Med. Imaging 33, 1818–1831.
Nyúl, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging 19, 143–150.
Patenaude, B., Smith, M.S., Kennedy, N.D., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56, 907–922.
Pizer, M.S., Fletcher, T.P., Sarang, J., Thall, A., Chen, Z.J., Fridman, Y., Fritsch, S.D., Gash, G.A., Glotzer, M.J., Jiroutek, R.M., Lu, C., Muller, E.K., Tracton, G., Yushkevich, P., Chaney, L.E., 2003. Deformable M-reps for 3D medical image segmentation. Int. J. Comput. Vision 55, 85–106.
Pohl, K., Fisher, J., Shenton, M., McCarley, R., Grimson, W., Kikinis, R., Wells, W., 2006a. Logarithm odds maps for shape representation. In: Larsen, R., Nielsen, M., Sporring, J. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006. Springer Berlin Heidelberg, volume 4191 of Lecture Notes in Computer Science, pp. 955–963.
Pohl, K.M., Bouix, S., Nakamura, M., Rohlfing, T., McCarley, R.W., Kikinis, R., Grimson, W.E., Shenton, M.E., Wells, W.M., 2007. A hierarchical algorithm for MR brain image parcellation. IEEE Trans. Med. Imaging 26, 1201–1212.
Pohl, M.K., Fisher, J., Grimson, L.E.W., Kikinis, R., Wells, M.W., 2006b. A Bayesian model for joint segmentation and registration. NeuroImage 31, 228–239.
Prastawa, M., Gerig, G., Lin, W., Gilmore, J.H., 2005. Automatic segmentation of MR images of the developing newborn brain. Med. Image Anal. 9, 457–466.
Puonti, O., Iglesias, J.E., Van Leemput, K., 2013. Fast, sequence adaptive parcellation of brain MR using parametric models. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, pp. 727–734.
Puonti, O., Van Leemput, K., 2016. Simultaneous whole-brain segmentation and white matter lesion detection using contrast-adaptive probabilistic models. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, volume 9556 of Lecture Notes in Computer Science, pp. 9–20.
Reuter, M., Schmansky, J.N., Rosas, D.J., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage 61, 1402–1418.
Roche, A., Malandain, G., Ayache, N., 2000. Unifying maximum likelihood approaches in medical image registration. Int. J. Imaging Syst. Technol. 11, 71–80.
Rohfling, T., Brandt, R., Menzel, R., Maurer Jr., R.C., 2004a. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21, 1428–1442.
Rohfling, T., Russakoff, B.D., Maurer Jr., R.C., 2004b. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imaging 23, 983–994.
Rousseau, F., Habas, A.P., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. IEEE Trans. Med. Imaging 30, 1852–1862.
Roy, S., Carass, A., Prince, J., 2013. Magnetic resonance image example-based contrast synthesis. IEEE Trans. Med. Imaging 32, 2348–2363.
Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations application to breast MR images. IEEE Trans. Med. Imaging 18, 712–721.
Sabuncu, M.R., Yeo, T.T.B., Van Leemput, K., Fischl, B., Golland, P., 2010. A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imaging 29, 1714–1729.
Shewchuk, J.R., 1994. An Introduction to the Conjugate Gradient Method without the Agonizing Pain (Technical Report). School of Computer Science, Carnegie Mellon University.
Studholme, C., Hill, D.L., Hawkes, D.J., 1999. An overlap invariant entropy measure of 3d medical image alignment. Pattern Recognit. 32, 71–86.
Ta, V.T., Giraud, R., Collins, D.L., Coupé, P., 2014. Optimized patchmatch for near real time and accurate label fusion. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, pp. 105–112.
Tang, X., Oishi, K., Faria, A.V., Hillis, A.E., Albert, M.S., Mori, S., Miller, M.I., 2013. Bayesian parameter estimation and segmentation in the multi-atlas random orbit model. PLoS One 8, e65591.
Thirion, J.P., 1998. Image matching as a diffusion process an analogy with Maxwell's demons. Med. Image Anal. 2, 243–260.
Tong, T., Wolz, R., Coupé, P., Hajnal, V.J., Rueckert, D., 2013. Segmentation of MR images via discriminative dictionary learning and sparse coding application to hippocampus labeling. NeuroImage 76, 11–23.
Van Leemput, K., 2009. Encoding probabilistic brain atlases using Bayesian inference. IEEE Trans. Med. Imaging 28, 822–837.
Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999a. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging 18, 897–908.
Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999b. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18, 885–896.
Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging 20, 677–688.
Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.A., 2011. A learning-based wrapper method to correct systematic errors in automatic image segmentation consistently improved performance in hippocampus, cortex and brain segmentation. NeuroImage 55, 968–985.
Wang, H., Suh, W.J., Das, R.S., Pluta, J., Craige, C., Yushkevich, A.P., 2013. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35, 611–623.
Wells III, M.W., Grimson, L.E.W., Kikinis, R., Jolesz, A.F., 1996a. Adaptive segmentation of MRI data. IEEE Trans. Med. Imaging 15, 429–442.
Wells, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996b. Multi-modal volume registration by maximization of mutual information. Med. Image Anal. 1, 35–51.
Wu, G., Wang, Q., Zhang, D., Nie, F., Huang, H., Shen, D., 2014. A generative probability model of joint label fusion for multi-atlas based brain segmentation. Med. Image Anal. 18, 881–890.
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.
Zijdenbos, A.P., Forghani, R., Evanc, A.C., 2002. Automatic "pipeline" analysis of 3-DMRI data for clinical trials application to multiple sclerosis. IEEE Trans. Med. Imaging 21, 1280–1291.
Zikic, D., Glocker, B., Criminisi, A., 2014. Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. Med. Image Anal. 18, 1262–1273.