HIPPOCAMPAL SEGMENTATION

Jon Pipitone[⋆] *Jason P. Lerch*[†] *Miriam Friedel*[†] *Aristotle Voineskos*[⋆] *Mallar Chakravarty*[⋆]

[⋆] Centre for Additions and Mental Health, Toronto ON, Canada
[†] Mouse Imaging Centre, ..., Toronto ON, Canada

ABSTRACT

Neuroimaging research often relies on automated anatomical segmentations of MR images of the brain. Current multi-atlased based approaches provide accurate segmentations of brain images by propagating region labels from manual delineations to unlabeled images. Unfortunately, these approaches often rely on a large number of such manual segmentations which take time and significant expertise to produce, neither of which may be readily available. We present an algorithm for the automatic segmentation of the hippocampus that minimizes the number of atlases needed whilst still achieving nearly the same accuracy as other multi-atlas approaches. We perform repeated random sub-sampling validation on the IBSR dataset to validate our approach and find an average best case percent difference to the multi-atlas approach of 1.9%.

Index Terms— hippocampus, segmentation, multi-atlas, automated

1. INTRODUCTION

The hippocampus is of particular interest to many researchers because it is implicated in forms of brain dysfunction such as Alzheimer's disease and schizophrenia, and has functional significance in cognitive processes such as learning and memory. For many research questions involving neuroimaging data, accurate identification of the hippocampus and its subregions in participant MR images is a necessary first step in order to analyse neuroanatomical differences and changes in sample populations.

Currently, the gold standard for neuroanatomical segmentation is manual labelling by an expert human rater. This is problematic for segmentation of the hippocampus for several reasons. First, manual segmentation takes a significant investment of time and expertise [**?**] which may not be readily available to researchers or clinicians. Second, the amount of data produced in neuroimaging experiments increasingly exceeds the capacity for identification of specific neuroanatomical structures by an expert manual rater. Third, the true delination of hippocampal anatomy in MR images is disputed[**?**], as evidenced by efforts to create an unified segmentation protocol[**?**]. Compounding each of these problems is that significant neuroanatomical variability in the hippocampus throughout the development and aging process and across the time course of different diseases is expected. As well, it may be necessary to use several different hippocampal definitions or, in fact, make research-specific modifications in order to test hypotheses (for example, [**?**] found that overall hippocampal volume difference did not predict recollection memory performance but by dividing the hippocampus into anterior and posterior regions, a predictive volume difference was found). Thus, whilst manual segmentation of the hippocampus is an important technique, it may be a bottleneck for researchers or clinicians who do not have access to the needed human expertise.

Automated segmentation techniques attempt to address the need for human expertise by performing segmentations computationally. A popular class of automated methods, *atlas-based segmentation*, rely on a set of expertly labeled neuroanatomical atlases. Each atlas is warped to fit a subject's neuroanatomy using nonlinear registration techniques[**?**, **?**]. Atlas labels are then transformed by this warping and a *label fusion* technique, such as voxel-wise voting, is used to merge the labellings from each atlas into a final segmentation for a subject.

Many descriptions of atlas-based segmentation algorithms report relying on an atlas library containing between 30 and 80 expertly labeled brains[**?**, **?**, **?**, **?**, **?**]. As noted, the production of an atlas library requires significant manual effort, and is limited since the choice of atlases or segmentation protocol may not reflect the underlying neuroanatomical variability of the population under study or be suited to answer the research questions at hand.

In this paper we propose an automated segmentation technique to address the issues found in existing atlas-based methods. Principly, our method aims to dramatically reduce the number of manually labelled atlases necessary. It does this by using a small atlas set to *generate* a much larger "template library", which is then used to segment the subjects in the same fashion as other

atlas-based methods. The essential insight of generating a template library is not new. Heckemann [**?**] compared generating a template library from a single atlas to standard multi-atlas segmentation and found poor performance and so deemed the approach as inviable. The LEAP algorithm [**?**] proceeds by iteratively segmenting the unlabelled image most similar to the atlas library images and then incorporating the now-labelled image into the atlas library, but requires 30 starting atlases. Our contribution is provide a method that provides comparable segmentation accuracy to these and other atlas-based maethods whilst using significantly fewer manually created atlases.

In previous work from our group [**?**], we explored this same approach to the segmentation of certain subcortical structures (striatum, globus pallidus, and thalamus) using a single histologically-derived atlas. In this work we extend our approach to the hippocampus, and validate it by varying the atlas and template library sizes. Due to the small number of atlases required, our method could easily accommodate different hippocampal definitions. Our aim is not to improve on segmentation accuracy beyond existing methods, but instead to provide a method that trades off manual segmentation expertise for computational processing time whilst providing sufficient accuracy for clinical and research applications.

2. MATERIALS AND METHODS

2.1. The Multiple Automatically Generated Templates (MAGeT) Algorithm

In this paper, we use the term *atlas* to mean any manually segmented MR image, and the term *atlas library* to mean a set of such images. We use the term *template* to refer to any MR image, and associated labelling, used to segment another image, and the term *template library* to refer to a set of such images. An atlas library may be used as a template library but, as we will discuss, a template library may also be composed of images with computer generated labellings.

The segmentation approach we propose is best understood as an extension of traditional multi-atlas segmentation. In multi-atlas segmentation, the inputs are an atlas library and a set of unlabelled MR images. An unlabelled image is segmented in the following way. Each atlas image is non-linearly registered to the unlabelled image, and then each atlas' labels are propagated via the resulting transformation. These candidate labels are then fused to produce a definitive segmentation. The label fusion method used may determine which templates are used to segment a particular subject (for instance, voxelwise majority vote takes labels from all atlases, whereas a weighted majority vote selects only those templates most similar to the subject).

Our extension to this approach is to add a preliminary stage in which we construct a template library to be used in the standard atlas-based approach as if it were the atlas library. To do this we select unlabelled images from the input images to form the template library. Next, labels from each atlas image are propagated to each image in the template library (via the transformation resulting from a non-linear registration between each pair of images). Thus, each template library image has a labelling from each atlas. Traditional multi-atlas segmentation is then used to produce segmentations for the entire set of unlabelled images (including those images used in the template library).

We use the *cross-correlation weighted voting* method to fuse the candidate labels of each unlabeled image. In this method, a voxel-wise majority vote is carried out the labels from the top n ranked template library images. The scores are given by the normalized cross-correlation between the unlabeled image and each template library image, after the unlabeled image is linearly registered to the template image. The cross-correlation score itself is calculated over a region of interest (ROI) defined on each template. A customized ROI is defined by first linearly registering each unlabeled subject to all possible templates in the template library, transforming all template hippocampus labels, defining the possible extent of all hippocampal labels, and finally dilating this extent by three voxels. redjp: it isn't clear that only the top ranked images are being used. perhaps some rationale for CC voting would help here. i.e. discuss improvement in overall registration quality – throwing out atlases liable to introduce unwanted variance

2.2. Image Processing and Registration Method

The N3 algorithm [**?**] is first used to minimize the intensity nonuniformity in each of the atlases and unlabeled subject images. Image registration is carried out in two phases. In the first, a 12-parameter linear transformation (3 translations, rotations, scales, shears) is estimated between images using an algorithm that maximizes the correlation between blurred MR intensities and gradient magnitude over the whole brain [**?**]. In the second phase, nonlinear registration is completed using the ANIMAL algorithm [**?**]: an iterative procedure that estimates a 3D deformation field between two MR images. At first, large deformations are estimated using blurred version of the input data. These larger deformations are then input to subsequent steps where the fit is refined by estimating smaller deformations on data blurred with a Gaussian kernel with a smaller FWHM. The final transformation is a set of local translations defined on a bed of equally spaced nodes that were estimated through the

optimization of the correlation coefficient. For the purposes of this work we used the regularization parameters optimized in Robbins et al.[**?**].

2.3. MRI dataset evaluated

For evaluation purposes we used the publicly available IBSR dataset. This dataset consists of T1 weighted MR image volumes from 18 subjects (4 females, 14 males) with ages between 7 and 71 years. Image dimensions for all MR volumes are normalized to $256 \times 256 \times 128$ voxels, with the voxel size ranging from $0.8 \times 0.8 \times 1.5mm$ to $1.0 \times 1.0 \times 1.5mm$. The images come 'positionally normalized' into the Talairach orientation (rotation only), and processed by the CMA 'autoseg' biasfield correction routines. The MR brain data sets and their manual segmentations are publicly available and were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/.

2.4. Experiments

In our experiments we verified how varying two different parameters effects labeling accuracy.

2.4.1. Varying Number of Atlases.

First we varied the number of input atlases that could be used from IBSR dataset from 3 to 8. In each case, we conducted 30 rounds of random subsampling cross-validation using the 18 manual segmented templates from the IBSR dataset as input. In each round, atlases were randomly chosen from the IBSR dataset and the remaining templates are used both as template library images and unlabeled subjects to be labeled using the MAGeT Brain algorithm.

2.4.2. Varying Number of Templates and Cross-Correlation Voting.

The second parameter we varied was the number of possible candidate labels to fuse for the final segmentation. For each round of validation in Experiment 1, we carried out cross-correlation weighted label fusion amongst only the labels from the n candidate templates from the library (i.e. selecting only those labels derived from templates with highest cross-correlation within the hippocampal ROI). The number of templates (n) was varied in the range $[3, r]$ (where r is the number of possible templates in the library).

2.5. Evaluation

2.5.1. Goodness-of-fit

Each segmentation was evaluated against the manual segmentation from the IBSR dataset using the Dice Kappa (κ) overlap metric:

$$\kappa = \frac{2a}{2a + b + c}$$

where a is the number of voxels common to the candidate segmentation and the gold standard and $b + c$ is the sum of the voxels uniquely identified by either the automatically generated candidate segmentation or the gold-standard.

2.5.2. Comparison Approaches

The resulting segmentations from each of our experiments is compared to two alternative segmentation approaches. The *naive* approach is that in which a single atlas is used to segment an unlabeled subject by directly propagating labels. The *multi-atlas* approach, as described above, in which we used cross-correlation to select the top most similar atlases to each subject a atlas labels are propagated to an unlabeled subject and the labels are fused using the cross-correlation voting described above. In total we evaluated approximately $52,000$ segmentations for the work presented in this manuscript.

3. RESULTS

Sample segmentations from one of the IBSR subjects is shown in Fig. **??**. The segmentations shown are varied across the template selection for a case where 3 atlases were used. The figure demonstrates a reduction in false negatives (voxels labeled by the 'gold-standard' only) in the hippocampal tail, a constant number of false positives (voxels labeled by the MAGeT Brain method only) in the hippocampal head, and an increase in the segmentation accuracy as the number of templates used from the template library are increased.

In Fig. **??** we compare the performance of the MAGeT Brain method across all validations (i.e. a combination of all atlases and template selections) against the multi-atlas method and the naive approach. Regardless of the number of atlases and templates used in a particular segmentation we see a marked increase in segmentation accuracy over the average naive segmentation (the average Kappa of all segmentations derived from all pairwise registrations in the IBSR dataset). Clearly, including more input atlases increases segmentation accuracy. Surprisingly, increasing the number of labels in the label fusion step through cross-correlation

Fig. 1. *Sample MAGeT Brain segmentations.* Segmentations of a single subject are shown when 3 atlases are used, with cross-correlation label fusion. Blue represents agreement between the gold-standard and MAGeT Brain. Red indicates voxels labeled only by the gold-standard, and green indicates voxels labeled only by MAGeT Brain. Note improved accuracy as the number of templates is increased.

based voting, does little to improve segmentation accuracy (except for in the cases where 3 and 4 atlases are used). Finally, the best-case MAGeT Brain performance scenario is when 8 atlases are used. While this does not reach the accuracy of the multi-atlas segmentation, there is only a 1.9% difference in the average Kappa values (0.775 for MAGeT Brain with 8 atlases and 0.790 for multi-atlas after selecting the top 14 templates).

4. DISCUSSION

In this paper, we have demonstrated that accurate segmentations can be achieved by simply automatically deriving a template library from a small set of input atlases. MAGeT Brain segmentations were compared to both naive segmentations and multi-atlas segmentations with cross-correlation voting. In [**?**], the authors report a Kappa of 0.814 on hippocampal segmentation in the IBSR dataset using a multi-atlas approach where atlas selection is based on the similarity between atlas and unlabeled subject after nonlinear registration, and using STAPLE [**?**] for label fusion. We demonstrated that the trade off in accuracy over a multi-atlas segmentation us-

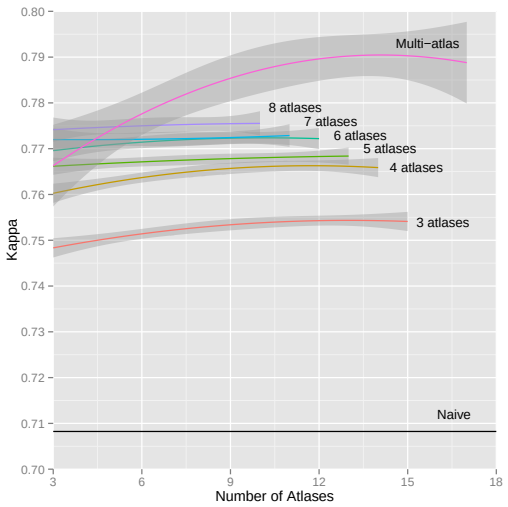

Fig. 2. *Performance of MAGeT Brain across all validations compared to mean naive segmentation, and mean multi-atlas with cross-correlation voting.* Multi-atlas and naive segmentations procedures provide upper and lower bounds (respectively) for accuracy. Note that improved segmentation accuracy is achieved through increasing the number of input atlases.

ing the entire IBSR manually segmented library is only on average 1.9%, using only 8 randomly chosen atlases. Discrepancies between our results and the above results may be due to choice of registration algorithm, regularization parameters, or similarity metric for label fusion. Performance may also be affected by the variability in the IBSR data set (as previously noted by [**?**]). Future work from our group will attempt to address some of these issues.

reddiscuss CC voting and why accuracy is monotonically increasing... wouldn't we expect an increase and then decrease?

5. REFERENCES

[1] Alexander Hammers, et al., "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe.," *Human brain mapping*, vol. 19, no. 4, pp. 224–47, Aug. 2003.

[2] E. Geuze, et al., "MR-based in vivo hippocampal volumetrics: 2. Findings in neuropsychiatric disorders," *Molecular Psychiatry*, vol. 10, no. 2, pp. 160, Sept. 2004.

[3] Clifford R Jack, et al., "Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion

for Alzheimer's disease.," *Alzheimer's & dementia*, vol. 7, no. 4, pp. 474–485.e4, July 2011.

[4] Jordan Poppenk et al., "A Hippocampal Marker of Recollection Memory Ability among Healthy Young Adults: Contributions of Posterior and Anterior Segments," *Neuron*, vol. 72, no. 6, pp. 931–937, Dec. 2011.

[5] D. Louis Collins, et al., "Automatic 3-D model-based neuroanatomical segmentation," *Human Brain Mapping*, vol. 3, no. 3, pp. 190–208, Oct. 1995.

[6] Arno Klein, et al., "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration.," *NeuroImage*, vol. 46, no. 3, pp. 786–802, July 2009.

[7] Rolf A Heckemann, et al., "Automatic morphometry in Alzheimer's disease and mild cognitive impairment.," *NeuroImage*, vol. 56, no. 4, pp. 2024–37, July 2011.

[8] D Louis Collins et al., "Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion.," *NeuroImage*, vol. 52, no. 4, pp. 1355–66, Oct. 2010.

[9] P Aljabar, et al., "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy.," *NeuroImage*, vol. 46, no. 3, pp. 726–38, July 2009.

[10] Kelvin K Leung, et al., "Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease.," *NeuroImage*, vol. 51, no. 4, pp. 1345–59, July 2010.

[11] Jyrki Mp Lötjönen, et al., "Fast and robust multi-atlas segmentation of brain magnetic resonance images.," *NeuroImage*, vol. 49, no. 3, pp. 2352–65, Mar. 2010.

[12] R A Heckemann, et al., "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy.," *NeuroImage*, vol. 46, no. 3, pp. 726–38, July 2006.

[13] Robin Wolz, et al., "LEAP: learning embeddings for atlas propagation.," *NeuroImage*, vol. 49, no. 2, pp. 1316–25, Jan. 2010.

[14] Mallar Chakravarty, et al., "Performing label-fusion based segmentation using multiple automatically generated tempaltes," *Human Brain Mapping*, 2012.

[15] J G Sled, et al., "A nonparametric method for automatic correction of intensity nonuniformity in MRI data.," *IEEE transactions on medical imaging*, vol. 17, no. 1, pp. 87–97, Feb. 1998.

[16] D L Collins, et al., "Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space.," *Journal of computer assisted tomography*, vol. 18, no. 2, pp. 192–205.

[17] Steven Robbins, et al., "Tuning and comparing spatial normalization methods.," *Medical image analysis*, vol. 8, no. 3, pp. 311–23, Sept. 2004.

[18] Simon K Warfield, et al., "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.," *IEEE transactions on medical imaging*, vol. 23, no. 7, pp. 903–21, July 2004.