How many atlases do you really need for accurate multi-atlas segmentation?

Jon Pipitone, Jason P. Lerch, Miriam Friedel, Aristotle N. Voineskos, M. Mallar Chakravarty

Abstract

Neuroimaging research often relies on automated anatomical segmentations of MR images of the brain. Current multi-atlas based approaches provide accurate segmentations of brain images by propagating manually derived segmentations of specific neuroanatomical structures to unlabelled data. These approaches often rely on a large number of such manually segmented atlases that take significant time and expertise to produce. We present an algorithm for the automatic segmentation of the hippocampus that minimizes the number of atlases needed while still achieving similar accuracy to multi-atlas approaches. We perform repeated random subsampling validation on the International Brain Segmentation Repository (IBSR) dataset to compare our approach to basic multi-atlas segmentation using the full IBSR dataset, and to single-atlas (model-based) segmentation. Our results show that with only 8 input atlases, MAGeT brain can achieve to within 2.0% segmentation accuracy of the basic multi-atlas approach using 17 input atlases (mean $\kappa = 0.775$ vs. $\kappa = 0.791$). These results compare favourably to those of other investigators who have used the IBSR data for validation.

keywords: hippocampus, segmentation, multi-atlas methods, label fusion, magnetic resonance imaging

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 magnetic resonance imaging (MRI) data accurate identification of the hippocampus and its subregions is a necessary first step to better understand the individual neuroanatomy of subjects.

Currently, the gold standard for neuroanatomical segmentation is manual delineation 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 (A. Hammers et al. 2003) 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 definition of hippocampal anatomy in MR images is disputed (Geuze, Vermetten, and Bremner 2004), as evidenced by efforts to create an unified segmentation protocol (Jack et al. 2011).

Compounding each of these problems is the significant neuroanatomical variability in the hippocampus throughout the course of aging, development, and neuropsychiatric disorders (Mouiha and Duchesne 2011). Additionally, it may be necessary to use several different hippocampal definitions or, in fact, make specific modifications in the course of research. For example, Poppenk et al. (Poppenk and Moscovitch 2011) found that subdividing the hippocampus into anterior and posterior regions resulted in a predictive relationship between volume difference of those regions and recollection memory performance. Thus, while manual segmentation of the hippocampus is a necessary technique, to researchers or clinicians who do not have access to the needed human expertise its use may be infeasible.

Automated segmentation techniques overcome the need for human expertise by performing segmentations computationally. A popular class of automated methods, *multi-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 (D. Collins et al. 1995; Klein et al. 2009). Atlas labels are then transformed by

this warping and a *label fusion* technique, such as voxel-wise voting, is used to merge the competing label definitions from each atlas into a final segmentation for a subject.

Many descriptions of multi-atlas-based segmentation algorithms report relying on an atlas library containing anywhere between 30 and 80 expertly labeled brains (R. A. Heckemann et al. 2011; D. Collins and Pruessner 2010; P. Aljabar et al. 2009; Leung et al. 2010; Lötjönen et al. 2010). 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 method to address the above issues of existing multi-atlas-based methods. Principally, our method aims to dramatically reduce the number of manually labelled atlases needed (under 10). This is achieved by using the small atlas library to boot-strap a much larger "template library", which is then used to segment the subjects in a similar fashion to basic multi-atlas segmentation. This approach has the additional advantage of using the unique subject population on hand to initialize the segmentation process and improve accuracy.

The essential insight of generating a template library is not new. Heckemann (R. Heckemann et al. 2006) 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 (R. Wolz et al. 2010) 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. The novelty of our method is to demonstrate the possibility of producing comparable segmentation accuracy to these and other multi-atlas-based methods while using significantly fewer manually created atlases.

In our previous work (Chakravarty et al. 2012), we applied MAGeT brain to segmentation of the human striatum, globus pallidus, and thalamus using a single histologically-derived atlas. The main contribution of this paper is to extend our approach to the human hippocampus and perform a thorough validation over a range of atlas and template library sizes, which was not done in our previous work. Due to the small number of atlases required, our method can 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 while providing sufficient accuracy for clinical and research applications.

Materials and Methods

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 basic multi-atlas segmentation (D. Collins and Pruessner 2010). In multi-atlas segmentation, an atlas library and unlabelled MR images are given as input. Every atlas image is nonlinearly registered to each unlabelled image, and then each atlas' labels are propagated via the resulting transformations. These labels are then fused to produce a single, definitive segmentation by some label fusion method (e.g. voxel-wise majority vote).

Our extension adds a preliminary stage in which a template library is constructed from input images, and used in place of an atlas library in the standard multi-atlas-based method. To create the template library, labels from each atlas image are propagated to each template library image via the transformation resulting from a non-linear registration between pair of images. As a result, each template library image has a label from each atlas. Basic multi-atlas segmentation is then used to produce segmentations for the entire set of unlabelled images (including those images used in the template library).

Label fusion is performed by cross-correlation weighted voting, a strategy weighted towards an optimal combination of subjects from the template library which has been previously shown to improve segmentation accuracy (P. Aljabar et al. 2009; D. Collins and Pruessner 2010). In this method, each template library image is ranked in similarity to each unlabelled image by the normalized cross-correlation of image intensities after linear registration in a region of interest (ROI) generously encompassing the hippocampus. Only the top ranked template library image labels are used in a voxel-wise majority vote. The ROI is heuristically defined as the extent of all atlas labels after linear registration to the template, dilated by three voxels (Chakravarty et al. 2012).

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/.

Image Processing and Registration Method

The N3 algorithm (Sled, Zijdenbos, and Evans 1998) is first used to minimize the intensity nonuniformity in each image. 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 (D. Collins et al. 1994). In the second phase, nonlinear registration is completed using the ANIMAL algorithm (D. Collins et al. 1995): 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. (Robbins et al. 2004). It should be noted that the MAGeT brain algorithm is not dependent on this, or any, particular choice of registration method (Chakravarty et al. 2012).

Experiments

We explored how varying the size of the atlas library and the template library effects labeling accuracy. For each parameter setting 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 images are used both as template library images and unlabeled subjects to be labeled using the MAGeT brain algorithm. We varied the size of the atlas library from 3 to 8, and used cross-correlation weighted label fusion to select the top n candidate templates from the remaining images. n was varied in the range $[3, 18 - a]$, where a is the size of the atlas library.

Evaluation

Goodness-of-fit

Automatically produced segmentations are evaluated against IBSR manual segmentations dataset using the Dice Kappa (κ) overlap metric, $\kappa = 2a/(2a + b + c)$, where a is the number of voxels common to both

segmentations and $b + c$ is the sum of the voxels uniquely identified in either segmentation.

Comparison Approaches

The resulting segmentations from each of our experiments are compared to those produced from two alternative segmentation approaches. The *single-atlas* approach uses one atlas to segment a unlabelled subject by directly propagating labels from the atlas by way of nonlinear registration. We computed a single-atlas segmentation for each image in the IBSR dataset from each of the other 17 labelled images. Similarly, we computed a segmentation for each image using the *basic multi-atlas* approach, described above, using the other 17 images as the atlas library. Additionally, we also varied the number of atlas images used in the label fusion step by employing cross-correlation weighted voting. In total we evaluated approximately $52,000$ segmentations for the work presented in this manuscript.

Results

Sample segmentations from a single IBSR subject compared with the gold-standard segmentation are in Fig. 1. Qualitatively, as the size of the template library is increased, the number of false negatives in the hippocampal tail region is reduced, and segmentation accuracy increases.

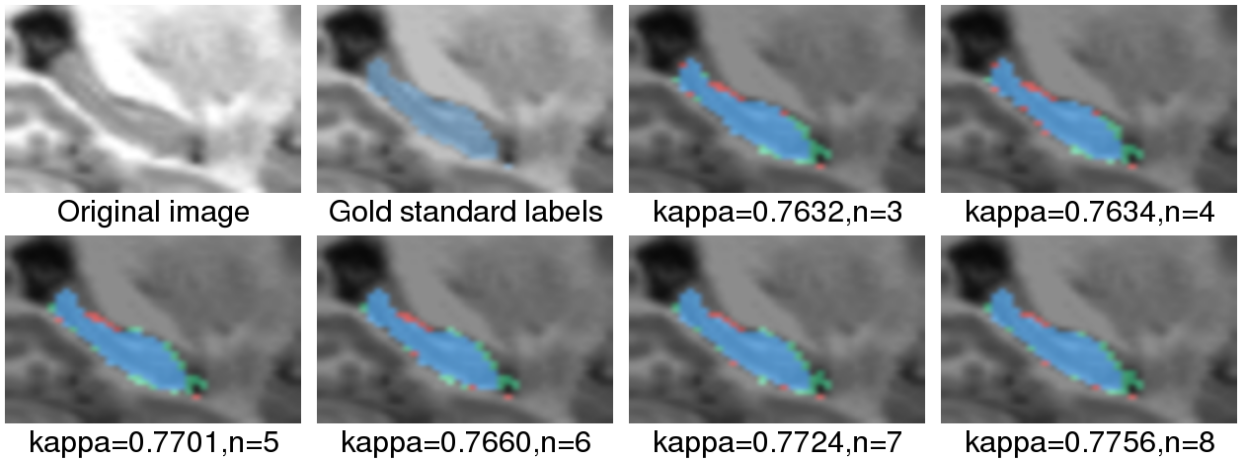

Figure 1: Segmentations of a single subject are shown when three atlases are used, with varying template library size n. Blue colouring represents agreement between the gold-standard and MAGeT brain. Green colouring indicates false positive voxels labelled by MAGeT brain (i.e. labelled voxels not appearing in the gold-standard labels), and red colouring indicates false negative voxels (i.e. voxels labelled in the gold-standard but not by MAGeT brain).

MAGeT brain achieves a level of segmentation accuracy that is to within 2.0% of the accuracy of the basic multi-atlas approach (Fig. *results*; mean $\kappa = 0.775$ vs mean $\kappa = 0.791$) in the best case. Importantly, to do this MAGeT brain only requires 8 manual segmented images whereas the basic multi-atlas requires 17 atlases. This represents a significant savings in manual effort, and supports our contention that it is possible with MAGeT brain to trade a small decrease in accuracy for a significant decrease in the number of manual segmentations needed.

Surprisingly, our analysis does not show any significant improvements in segmentation accuracy for either method when applying cross-correlation weighted voting to reduce the number of template labels being fused.

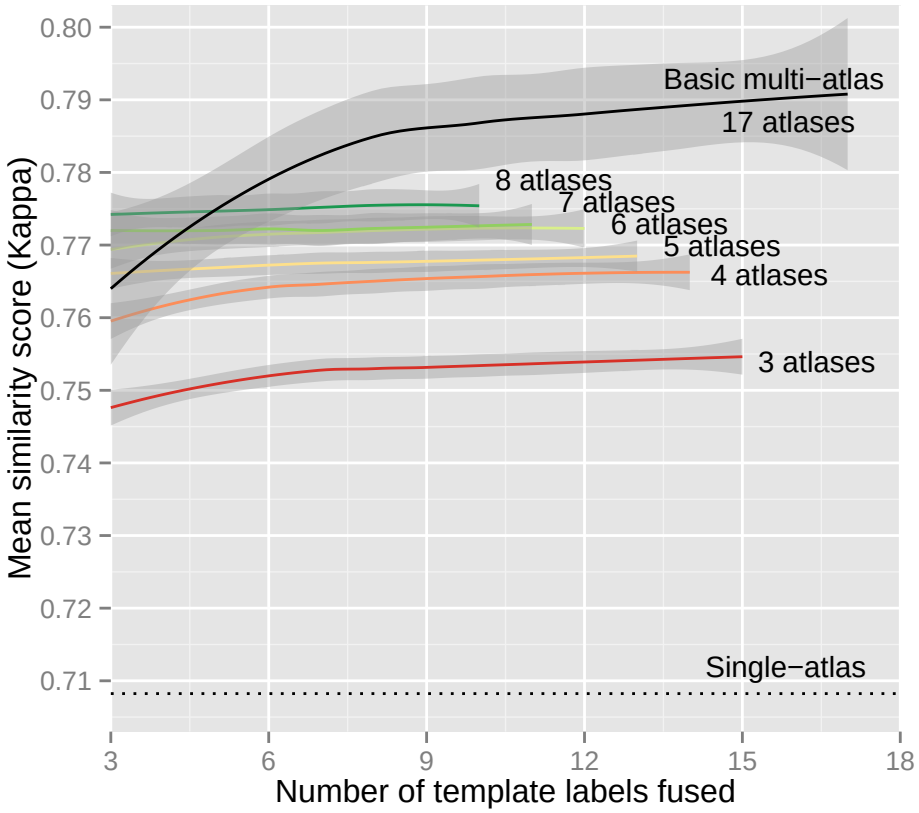

Figure 2: Mean performance of MAGeT brain with varying atlas library size and number of template labels fused. Also shown is the mean basic multi-atlas performance when using an atlas library of 17 images and varying the number of labels fused, as well as the mean performance of single-atlas segmentations. Data is fit with LOESS local regression smoothing. One standard deviation is shown in grey.

Discussion

In this paper, we have demonstrated that accurate segmentations can be produced by automatically deriving a template library from a small set of input atlases. MAGeT brain segmentations were compared to both single-atlas segmentations and multi-atlas segmentations with cross-correlation voting. For the IBSR dataset, on average MAGeT brain achieves within 2.0% of the segmentation accuracy of the multi-atlas approach but requires only 8 atlases as compared to using the entire IBSR manually segmented library of 17 atlases.

Lötjönen et al. (Lötjönen et al. 2010) 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 (Warfield, Zou, and Wells 2004) for label fusion. 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 (Klein et al. 2009)). Future work from our group will attempt to address some of these issues.

Acknowledgements

Computations were performed on the gpc supercomputer at the SciNet HPC Consortium (Loken et al. 2010). SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

This work was supported by the Canadian Institutes of Health Research (CIHR), National Alliance for Research on Schizophrenia and Depression (NARSAD), Ontario Mental Health Foundation (OMHF), and the CAMH Foundation (Koerner New Scientist Program and Paul Garfinkel New Investigator Catalyst Fund).

Aljabar, P., R Heckemann, a Hammers, J V Hajnal, and D Rueckert. 2009. "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy." *NeuroImage* 46 (3). Elsevier Inc.: 726–38. doi:10.1016/j.neuroimage.2009.02.018.

Chakravarty, M., Patrick Steadman, Matthijs van Eede, Rebecca Calcott, Victoria Gu, Phillip Shaw, Armin Raznahan, Louis Collins, and Jason P Lerch. 2012. "Performing label-fusion based segmentation using multiple automatically generated tempaltes." *Human Brain Mapping*.

Collins, D., and Jens C Pruessner. 2010. "Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion." *NeuroImage* 52 (4). Elsevier Inc.: 1355–66. doi:10.1016/j.neuroimage.2010.04.193.

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

Collins, D., P Neelin, T M Peters, and A C Evans. 1994. "Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space." *Journal of Computer Assisted Tomography* 18 (2): 192–205. http://www.ncbi.nlm.nih.gov/pubmed/8126267.

Geuze, E., E. Vermetten, and J D Bremner. 2004. "MR-based in vivo hippocampal volumetrics: 2. Findings in neuropsychiatric disorders." *Molecular Psychiatry* 10 (2): 160. doi:10.1038/sj.mp.4001579.

Hammers, A., Richard Allom, Matthias J Koepp, Samantha L Free, Ralph Myers, Louis Lemieux, Tejal N Mitchell, David J Brooks, and John S Duncan. 2003. "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe." *Human Brain Mapping* 19 (4): 224–47. doi:10.1002/hbm.10123.

Heckemann, R. A, Shiva Keihaninejad, Paul Aljabar, Katherine R Gray, Casper Nielsen, Daniel Rueckert, Joseph V Hajnal, and Alexander Hammers. 2011. "Automatic morphometry in Alzheimer's disease and mild cognitive impairment." *NeuroImage* 56 (4): 2024–37. doi:10.1016/j.neuroimage.2011.03.014.

Heckemann, R., J V Hajnal, P Aljabar, D Rueckert, and A Hammers. 2006. "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy." *NeuroImage* 46 (3): 726–38.

doi:10.1016/j.neuroimage.2009.02.018.

Jack, C., Frederik Barkhof, Matt A Bernstein, Marc Cantillon, Patricia E Cole, Charles Decarli, Bruno Dubois, et al. 2011. "Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease." *Alzheimer's & Dementia* 7 (4): 474–485.e4. doi:10.1016/j.jalz.2011.04.007.

Klein, A., Jesper Andersson, Babak A Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E Christensen, et al. 2009. "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration." *NeuroImage* 46 (3): 786–802. doi:10.1016/j.neuroimage.2008.12.037.

Leung, K., Josephine Barnes, Gerard R Ridgway, Jonathan W Bartlett, Matthew J Clarkson, Kate Macdonald, Norbert Schuff, Nick C Fox, and Sebastien Ourselin. 2010. "Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease." *NeuroImage* 51 (4). Elsevier Inc.: 1345–59. doi:10.1016/j.neuroimage.2010.03.018.

Loken, C., Daniel Gruner, Leslie Groer, Richard Peltier, Neil Bunn, Michael Craig, Teresa Henriques, et al. 2010. "SciNet: Lessons Learned from Building a Power-efficient Top-20 System and Data Centre." *Journal of Physics: Conference Series* 256 (November): 012026. doi:10.1088/1742-6596/256/1/012026.

Lötjönen, J., Robin Wolz, Juha R Koikkalainen, Lennart Thurfjell, Gunhild Waldemar, Hilkka Soininen, and Daniel Rueckert. 2010. "Fast and robust multi-atlas segmentation of brain magnetic resonance images." *NeuroImage* 49 (3): 2352–65. doi:10.1016/j.neuroimage.2009.10.026.

Mouiha, A., and S. Duchesne. 2011. "Multi-decade hippocampal and amygdala volume analysis: equal variability and limited age effect." *Neuroscience Letters* 499 (2). Elsevier Ireland Ltd: 93–98. doi:10.1016/j.neulet.2011.05.041.

Poppenk, J., and Morris Moscovitch. 2011. "A Hippocampal Marker of Recollection Memory Ability among Healthy Young Adults: Contributions of Posterior and Anterior Segments." *Neuron* 72 (6): 931–37. doi:10.1016/j.neuron.2011.10.014.

Robbins, S., Alan C Evans, D Louis Collins, and Sue Whitesides. 2004. "Tuning and comparing spatial normalization methods." *Medical Image Analysis* 8 (3): 311–23. doi:10.1016/j.media.2004.06.009.

Sled, J., a P Zijdenbos, and a C Evans. 1998. "A nonparametric method for automatic correction of intensity nonuniformity in MRI data." *IEEE Transactions on Medical Imaging* 17 (1): 87–97. doi:10.1109/42.668698.

Warfield, S., Kelly H Zou, and William M Wells. 2004. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." *IEEE Transactions on Medical Imaging* 23 (7): 903–21. doi:10.1109/TMI.2004.828354.

Wolz, R., Paul Aljabar, Joseph V Hajnal, Alexander Hammers, and Daniel Rueckert. 2010. "LEAP: learning embeddings for atlas propagation." *NeuroImage* 49 (2): 1316–25. doi:10.1016/j.neuroimage.2009.09.069.