Eurographics Workshop on Visual Computing for Biology and Medicine (2010)
D. Bartz, C. P. Botha, J. Hornegger, R. Machiraju, A. Wiebel, and B. Preim (Editors)

Evaluation of hippocampal segmentation methods for healthy and pathological subjects

C. A. Bishop[†1] and M. Jenkinson[1] and J. Declerck[2] and D. Merhof[3]

[1]FMRIB, University of Oxford, UK
[2]Siemens Molecular Imaging, Oxford, UK
[3]University of Konstanz, Germany

Abstract

Hippocampal atrophy is a clinical biomarker of Alzheimer's disease (AD) and is implicated in many other neurological and psychiatric diseases. For this reason, there is much interest in the accurate, reproducible delineation of this region of interest (ROI) in structural MR images. Here, both current and novel MR hippocampal segmentation methods are presented and evaluated: Two versions of FMRIB's Integrated Registration and Segmentation Tool (FIRST and FIRSTv2), Freesurfer's Aseg (FS), Classifier Fusion (CF) and a Fast Marching approach (FMClose). Segmentation performance on two clinical datasets is assessed according to three common measures: Dice coefficient, false positive rate (FPR) and false negative rate (FNR). The first clinical dataset contains 9 normal controls (NC) and 8 highly-atrophied AD patients, whilst the second is a collection of 16 NC and 16 bipolar (BP) patients. Results show that CF outperforms all other methods on the BPSA data, whilst FIRST and FIRSTv2 perform best on the CMA data, with average Dice coefficients of 0.81±0.01, 0.85±0.00 and 0.85±0.01, respectively. This work brings to light several strengths and weaknesses of the evaluated hippocampal segmentation methods, of utmost importance for robust and accurate segmentation in the presence of specific and substantial pathology.

Categories and Subject Descriptors (according to ACM CCS): I.4.6 [Image processing and computer vision]: Segmentation—Edge and feature detection, Region growing, partitioning

1. Introduction

The hippocampus belongs to the limbic system of the brain and is located inside the medial temporal lobe, bordering the lateral ventricles, the thalamus and the amygdala. Crucially involved in episodic and spatial memory processes, this structure has been implicated in the pathophysiology of many neurological and psychiatric diseases [KUN*09]. Hippocampal volume loss is a characteristic feature of both early Alzheimer's Disease (AD) and temporal lobe epilepsy, with more ambiguous findings in affective disorders such as major depression, bipolar disorder and schizophrenia. Of these aforementioned diseases, AD presents one of the largest socio-economic impacts, currently affecting around 26 million elderly people worldwide and costing the US economy over $100 billion per year [MWT*05]. Consequently, there is particular interest in the early diagnosis of AD and preservation of neurological function at the highest possible level, requiring accurate, reproducible delineation of this region of interest (ROI) in structural MR images. This task is complicated by the complex shape of the hippocampus, large inter-subject variability and poor contrast at the hippocampus-amygdala border. Conventionally used manual segmentation requires expert knowledge and can be extremely time consuming, thus impractical for large-scale clinical studies, fuelling the development of semi-automated and automated segmentation methods for this purpose.

Previous segmentation methods for the hippocampus include the label-fusion, atlas-based approach of Heckemann *et al.* [HHA*06] and extensions [AHH*07, LWK*10], competitive region-growing approaches [CMBH*07, CCC*08] and model-based methods such as the later-mentioned

† Corresponding author: courtney.bishop@merton.ox.ac.uk

FIRST and FS tools [WJP*09, FSB*02]. Whilst many segmentation tools perform well on healthy subject data, they often fail to capture the substantial and specific atrophy displayed in clinical data. Furthermore, with most studies focusing on development and evaluation of an individual segmentation method on a single dataset, reporting any number of performance measures, direct comparison of hippocampal segmentation methods is compromised.

This work evaluates five MR hippocampal segmentation methods: Two versions of FMRIB's Integrated Registration and Segmentation Tool, FIRST and FIRSTv2, Freesurfer's Aseg (FS), Classifier Fusion (CF) and an early development of a novel Fast Marching approach, FMClose. Segmentation performance on two clinical datasets (Section 2.1) is assessed according to three common measures: Dice coefficient, false positive rate (FPR) and false negative rate (FNR).

2. Materials and Methods

2.1. Image Data

Two clinical MR datasets with corresponding expert manual labels are used in this work: The first consists of 9 normal control (NC) subjects and 8 highly-atrophied AD patients, supplied by the Center for Morphometric Analysis (CMA), whilst the second is a collection of 16 NC and 16 bipolar (BP) patients from a collaborator in San Antonio (denoted BPSA). Although mainly motivated by hippocampal segmentation methods for AD, the underlying methodology is also applicable to other diseases in which hippocampal atrophy is implicated (e.g. BP disorder), hence incorporation of the BPSA data in this study.

Experts from the respective research sites used a semi-automated contouring tool to manually define the hippocampus in sequential coronal slices, based on intensity boundaries and well-established geometrical rules of neuroanatomy. All experts underwent a period of training (of up to three methods) until they had reached a defined reproducibility. Reported inter-rater reliability for the BPSA and CMA data are 0.90 and 0.80, respectively, with image resolution and demographics given in Table 1:

Data	Size	Age	Resolution (mm)	Subjects
CMA	17	65 - 83	0.94 x 1.50 x 0.94	NC, AD
BPSA	32	20 - 58	0.8 x 0.8 x 0.8	NC, BP

Table 1: *Group size, age (years) and resolution of the two clinical datasets.*

2.2. Segmentation Methods

FIRST

FIRST is a Bayesian statistical shape and appearance (intensity) model [WJP*09]. From a training dataset, shape is modelled as a multivariate Gaussian point distribution model (PDM), parameterized by a linear combination of the mean and eigenvectors. Intensity is also modelled by its mean and eigenvectors, with intensity profiles sampled along the surface normal at each vertex location. Following intensity normalization, each new image is spatially normalized to the MNI152 reference space using a two-stage linear registration process; whole-brain followed by a subcortical-weighted affine transformation, minimizing the correlation ratio similarity function. In the reference space, the average mesh is initialized and iteratively deformed (for a fixed number of iterations) to give the resultant structure mesh.

FIRSTv2

FIRSTv2 is an adaptation of the aforementioned FIRST tool, using a novel structure-specific spatial normalization process prior to model-fitting. The second stage of the two-stage registration process is replaced by a hippocampus-weighted affine transformation, using an eight-times dilated hippocampus mask in the reference space. FMRIB's Linear Image Registration Tool (FLIRT) is used for this affine registration of each image to the reference space.

FS

FS uses Bayesian parameter estimation theory [FSB*02], with voxel classification determined by the segmentation that maximizes the probability of input data given prior probabilities from the training set. Each new image is initially segmented, assigning voxels to the class of maximum probability determined by an initial probability map. Class probabilities are re-computed using a local neighbourhood function and the image is re-segmented, repeating these two steps until the segmentation does not change.

In this work, the first step of the program 'recon-all' is run for all subjects to transform images from their native space to FS Talairach space, with the 'subcortseg' option used to segment all subcortical structures. The affine transformation between each image's native space and Talairach space is computed using the 'tkregister2' function. In FS Talairach space, hippocampus labels are extracted from the FS output volume, binarized and transformed back into their native space (using the aforementioned affine transform). Here, the final FS hippocampal labels are obtained by threshold.

CF

The CF method presented here (Figure 1) is an extension of the label-fusion atlas-based approach of Aljabar *et al.* [AHH*07]. In this work, the term 'atlas' refers to the paired T1-weighted anatomical image and its corresponding manual labels. When the atlas anatomies are aligned (i.e. registered) to a new query image, the atlas labels can be considered as a 'classifier', providing a segmentation estimate for this query image.

Here, all atlas anatomies A_j are registered to the query image Q using a three-step affine transformation process;

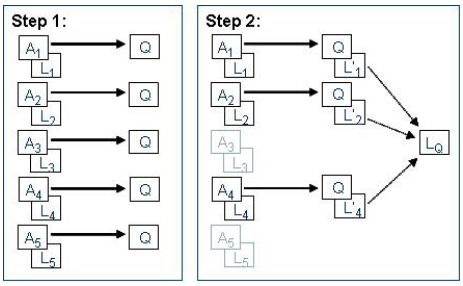

Figure 1: *Flow chart of the Classifier Fusion (CF) method. Step 1: All atlas anatomies A_i are affine-registered to the query image Q. Step 2: Top-ranked atlases are selected and non-linearly transformed to the query image space, using the three-step affine transformation (from Step 1) for initialization. Here, labels are fused.*

whole-head, subcortical- and hippocampus-weighted transformations. Top-ranked atlas anatomies are selected and non-linearly registered to the query image, using the pre-calculated three-step affine transformation for initialization. The resultant non-linear transforms propagate the corresponding atlas labels L_i to the query image, providing a series of segmentation estimates L_i' for this image. Finally, in the query image space, the labels L_i' are fused (using majority voting) to give a single segmentation estimate for the query image. All registrations are performed using FSL tools: FLIRT for affine transformations and FNIRT for non-linear warps. For affine registrations the normalized cross-correlation cost function is minimized, with the same function evaluated over a restricted hippocampus region for selection of top-ranked atlas anatomies.

As opposed to the approach in [AHH*07] which employs a reference space to avoid the computational burden of non-linearly registering all atlas anatomies to the query image space, we perform CF in the query image space to avoid degradation of the query image. This is in line with a recent study by Lotjonen *et al.* [LWK*10] suggesting that CF performs best in the query image space. To avoid the computational burden associated with non-linear registration of every atlas to each query image, and to maximize classifier separation, our approach selects the top-ranked classifiers after a three-step affine transformation.

FMClose

This novel semi-automated region growing method for MR data is inspired by the work of Maroy *et al.* [MBC*08] on segmentation of dynamic PET images. Implemented in MATLAB (http://www.mathworks.com/), a Sethian Fast Marching (FM) approach is used to construct the arrival time surface $U(i,j,k)$ of an infinitesimal front propagating outwards from an initial seed point p_0. The arrival time surface is so-called because it gives the arrival time of the propagating front at any given point i,j,k in 3D space, illustrated in Figure 2 for a 2D problem. The front progresses along the path of least resistance, adding voxels which constitute the lowest potential P, thus finding the minimum energy curve.

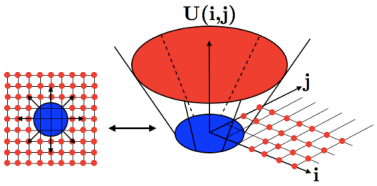

Figure 2: *Schematic illustration of the Sethian Fast Marching approach in 2D, showing the initial curve (left) and the resulting arrival time surface $U(i,j)$ (right).*

To initialize the algorithm, the user specifies a seed point in the centre of the hippocampus, as well as a hippocampus bounding ellipsoid. The FM front propagates outwards from the selected seed point to the defined ellipsoid boundary, generating a segmented volume for post-hoc manual thresholding. Window/level is used to visually assess the optimum voxel count threshold. The resultant hippocampal volume is morphologically closed using a 3x3 structuring element, generating the final FM volume. To correct for known spillover at the hippocampus-amygdala boundary, a partially dilated amygdala mask is used to remove any amygdala voxels incorrectly labelled as hippocampus by the FM algorithm. Whilst this is not ideal, causing some bias in favour of the FM algorithm, removal of amygdala voxels ensures that these findings do not dominate the hippocampal segmentation estimate. Ongoing development of the FM method aims to address this hippocampus-amygdala boundary problem, incorporating a spatial probability map to penalize propagation of the FM front across the boundary.

2.3. Performance Measures

Due to the vast array of published performance measures, with no universal standard agreed upon, it is often difficult to formulate direct comparisons of segmentation methods. Here, we report a combination of metrics used in a recent online resource for validation of brain segmentation methods [SPM*09] and the MICCAI 2008 competition workshop on MS lesion segmentation: Dice coefficient, FPR and FNR. If X is the set of all voxels in the image, we define the ground truth $T \subset X$ as the set of voxels labelled as hippocampus by the expert manual labels. Similarly, we define the set $S \subset X$ as the set of voxels labelled as hippocampus by the segmentation algorithm or method being tested.

2.3.1. Success and Error Rates

The true positive set is the set of voxels common to both T and S, defined as $TP = T \cap S$. The true negative set is the set of voxels that are labelled as non-hippocampus in

both sets, defined as $TN = \overline{T} \cap \overline{S}$. Similarly, the false positive set is defined as $FP = \overline{T} \cap S$ and the false negative set is $FN = T \cap \overline{S}$. From these four sets, we can compute various success and error rates for image segmentation:

$$\text{FPR} = \frac{|FP|}{|FP| + |TN|} = 1 - \text{ specificity} \quad (1)$$

$$\text{FNR} = \frac{|FN|}{|FN| + |TP|} = 1 - \text{ sensitivity} \quad (2)$$

Generation of standard-space FP and FN maps

Standard-space maps are generated to assess the spatial distribution of FP and FN voxels for each segmentation method. The output of the FMRIB tool 'first_flirt' is a linear transform mapping each image to the standard MNI152 space. For each segmentation method, the 'first_flirt' transforms are used to map all subjects' FP and FN voxels from their native space to MNI152 space. Here, the average FP and FN maps are computed (summing partial trilinear interpolation values), with voxel values corresponding to the fraction of subjects showing a FP or FN result at that voxel.

2.3.2. Similarity Metrics

Dice Coefficient

The Dice coefficient is defined as the size of the intersection of two sets divided by their average size:

$$D(T,S) = \frac{|T \cap S|}{\frac{1}{2}(|T| + |S|)} = \frac{|TP|}{\frac{1}{2}(2|TP| + |FN| + |FP|)} \quad (3)$$

2.4. Statistical Analysis

To test for overall difference across the segmentation methods and clinical group-by-method interactions, a 5*2 (method*group) mixed analysis of variance (ANOVA) design is employed for each performance measure (Dice, FPR and FNR), with method as the within-subjects factor and clinical group as the between-subjects factor. Finally, paired-sample t-tests with Bonferroni correction for multiple comparisons are used to identify between-method differences; p_d, p_{fp}, p_{fn} denote corrected p-values for Dice, FPR and FNR, respectively.

3. Results

Where segmentation estimates of both the left and right hippocampi are obtained, we do not observe any obvious left-right asymmetries in method performance. Consequently, results are given for the left hippocampus only. Box plots showing method performance on the CMA and BPSA data are given in Figures 3 and 5, respectively. Additionally, Figures 4 and 6 present the standard-space FP and FN maps for each method on the CMA and BPSA data, respectively. For each ANOVA, Mauchly's test indicates that the assumption of sphericity for method has been violated ($p<.0005$) so degrees of freedom are corrected using Greenhouse-Geisser estimates of sphericity.

3.1. CMA Data

The ANOVA designs reveal a significant effect of method for each performance measure ($p<.0005$), but no significant method*group interaction for Dice ($F(2.05,30.76)=1.27$, $p=.295$) or FNR ($F(1.58,23.69)=.64$, $p=.500$) and no significant group effect for Dice ($F(1,15)=1.63$, $p=.221$) or FPR ($F(1,15)=.07$, $p=.796$). The significant method*group interaction for FPR ($F(2.19,32.85)=3.67$, $p=.033$) is a result of the CF method having an opposite trend in FPR across the groups compared to other methods (not shown). For all methods FNR(AD)>FNR(NC) results in a significant group effect ($F(1,15)=7.01$, $p=.018$).

Paired t-test results give a strict relative ordering (1-5) of the segmentation methods. Based on a simple sum of these rankings, the overall ordering of the segmentation methods is as follows:

- FIRST > FIRSTv2 > FS > CF > FMClose

The differences in performance between FIRST and FIRSTv2, and between FS and CF, are minimal, with both FIRST and FIRSTv2 clearly outperforming all other segmentation methods on the CMA data.

FIRST gives significantly higher Dice coefficients and lower FPR compared to most other methods (FS: p_d <.0005, p_{fp} <.0005; CF: p_d <.0005, p_{fp}=.010; FMClose: p_d <.0005). FP findings occur mainly in the medial-inferior region of the hippocampus head, whilst medial boundaries and posterior regions of the hippocampus head tend to be under-estimated (Figure 4: Panels A and B, respectively, top row).

FIRSTv2 displays no significant difference in Dice coefficients, FPR or FNR compared to FIRST (p_d=p_{fn}=1, p_{fp}=.246) with a similar pattern of FP and FN findings (Figure 4: Panels A and B, respectively, second row).

FS has a respectable average Dice coefficient of 0.80 ± 0.01 and significantly reduced FNR compared to CF and FMClose (CF: p_{fn}=.021; FMClose: p_{fn} <.0005), although visual assessment of results indicates a 'greedy' labelling tendency, leading to significantly higher FPR compared to FIRST, FIRSTv2 and FMClose (FIRST: p_{fp} <.0005; FIRSTv2: p_{fp} <.0005; FMClose: p_{fp}=.009). The standard-space maps in Figure 4 (third row) show a general over-estimation at the hippocampus-amygdala border (i.e most anterior-superior direction), the body and superior tail sections, together with FN findings in the medial and posterior regions of the hippocampus head.

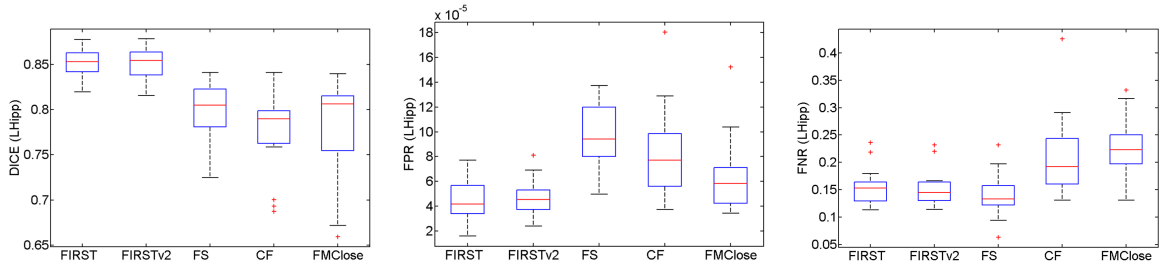

Figure 3: *Box plots showing Dice coefficients (left), FPR (middle) and FNR (right) for each segmentation method on the CMA data. Although all FPR are extremely low (of the order* 10^{-5}*), due to the high TN count in the calculation of FPR (Equation 1), statistically significant and important differences between the segmentation methods are observed.*

Whilst CF displays an average Dice coefficient comparable to that of human raters (0.77±0.01), its overall segmentation performance on the CMA data is hampered by that of a few outliers, contributing to significantly higher FPR or FNR compared to the model-based methods (FIRST: p_{fp}=.010; FIRSTv2: p_{fp}=.012; FS: p_{fn}=.021). Subjects have a lot of neck present in the T1-weighted images and some have noticeable bias field effects, giving poor brain extraction and subsequent registration errors. These registration errors contribute to both FP and FN findings at most hippocampal boundaries (Figure 4: fourth row), although the relatively low frequency and even distribution of FP and FN counts suggests that the registration errors are uncorrelated.

FMClose displays no significant difference in Dice coefficients compared to both FS and CF and significantly reduced FPR compared to FS (FS: p_d=1, p_{fp}=.009; CF: p_d=1). Still, the method ranks lowest overall, with under-estimation of hippocampal volume giving significantly higher FNR than FIRST, FIRSTv2 and FS (All: p_{fn} <.0005). It is important to note that spillover into the amygdala has been neglected in this analysis, so reported FPR is lower than actual FPR. In general, the current implementation of FMClose suffers from spillover in anterior and inferior regions, with under-estimation of medial boundaries, the tail and posterior regions of the hippocampus head, contributing to both FP and FN count (Figure 4: fifth row). Interestingly, sub-hippocampal regions of contrasting intensity, such as the dentate gyrus, are excluded from the FMClose segmentation estimate and contribute further to FNR.

3.2. BPSA Data

For each performance measure, the ANOVA design reveals a significant effect of method (Dice: F(1.23,37.00)=12.09, p=.001; FPR: F(2.54,76.13)=70.33, p<.0005; FNR: F(1.23,36.88)=19.24, p<.0005), but no significant method*group interaction (Dice: F(1.23,37.00)=.311, p=.627; FPR: F(2.54,76.13)=.482, p=.664; FNR: F(1.23,36.88)=.246, p=.672) and no significant group effect (Dice: F(1,30)=.591, p=.448; FPR: F(1,30)=.352, p=.558; FNR: F(1,30)=.395, p=.535).

The paired t-test results give a strict relative ordering (1-5) of the segmentation methods for each performance measure. Based on a simple sum of these rankings, the overall ordering of the segmentation methods is as follows:

- CF > FIRSTv2 > FS > FMClose > FIRST

CF clearly outperforms all other segmentation methods on the BPSA data, whilst FS and FMClose rank a close third and fourth.

FIRST performs relatively well for the majority of subjects, but its overall segmentation performance is hindered by that of a few subjects, where poor initial registration with the standard (MNI152) space results in poor segmentation estimates. Consequently, whilst FIRST displays the third highest median Dice coefficient and a median FNR comparable to that of CF (Figure 5: left and middle), the method has significantly reduced Dice coefficients, higher FPR and FNR compared to most other methods (FIRSTv2: p_d=.006, p_{fn}=.003; FS: p_{fn} <.0005; CF: p_d=.001, p_{fp} <.0005; FMClose: p_{fp}=.001). Interestingly, the standard-space maps in Figure 6 show over-estimation of all anterior and superior hippocampal borders by all model-based methods (FIRST: top row, FIRSTv2: second row, FS: third row) and a prevalence of FN counts at all inferior borders for FIRST, attributed to a superior-shift of the entire hippocampus.

The registration errors experienced with FIRST on the BPSA data are largely addressed by FIRSTv2, incorporating a hippocampus-specific spatial normalization step prior to model-fitting. Consequently, FIRSTv2 displays significantly higher Dice coefficients and lower FNR compared to most other methods (FIRST: p_d=.006, p_{fn}=.003; FS: p_d <.0005; CF: p_{fn} <.0005; FMClose: p_d <.0005, p_{fn} <.0005). FN findings are evident in the most anterior-superior part of the hippocampus head and tail (Figure 6: Panel B, first row).

The FS tool displays Dice coefficients comparable to that of human raters (0.75±0.01) and a significantly reduced FNR compared to all other methods (All: p_{fn} <.0005), only

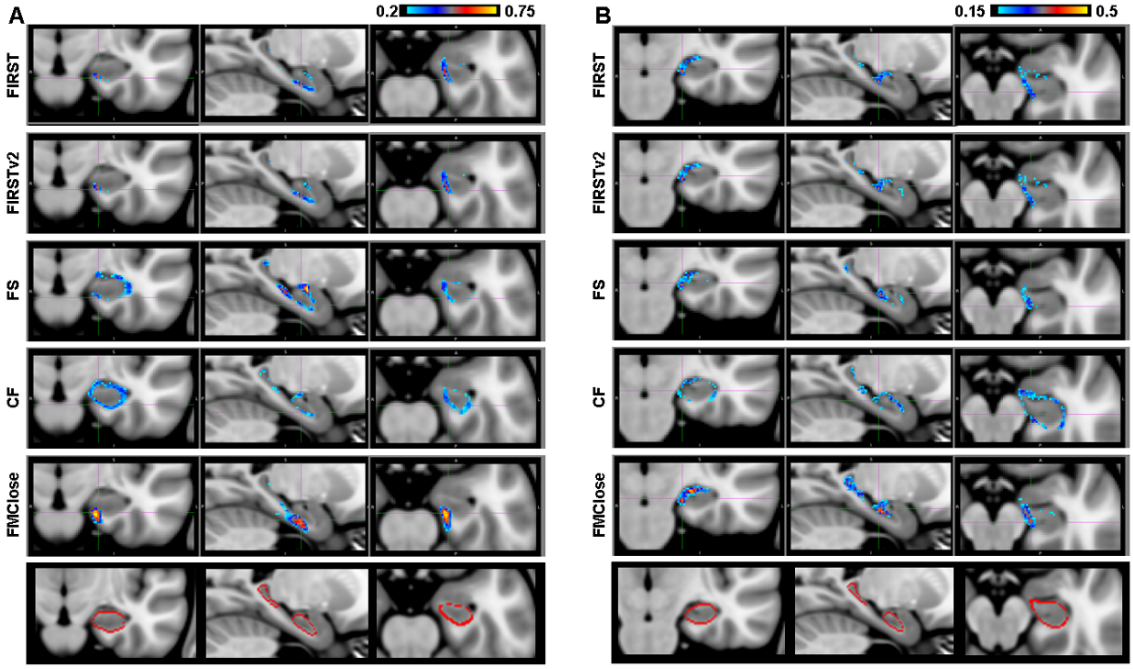

Figure 4: *Panels A and B show coronal- (left), sagittal- (middle) and axial- (right) view images of the standard-space FP and FN maps, respectively, for each segmentation method on the CMA data. A voxel-wise threshold of 20% FP finding across the dataset is applied to all FP maps, whilst the FN maps are threshold at 15%. To aid visual assessment, the hippocampal boundary defined by FIRST on the MNI152 standard image is shown (sixth row) and the cursor position is the same for all images in each panel.*

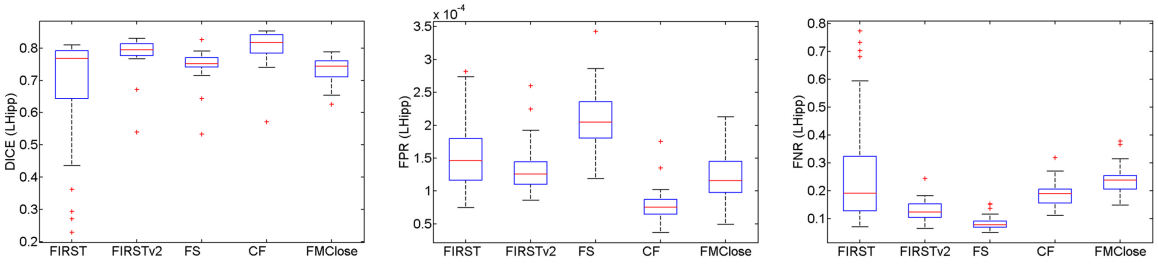

Figure 5: *Box plots showing Dice coefficients (left), FPR (middle) and FNR (right) for each segmentation method on the BPSA data. Although all FPR are extremely low (of the order 10^{-4}), due to the high TN count in the calculation of FPR (Equation 1), statistically significant and important differences between the segmentation methods are observed.*

underestimating the most superior sections of the hippocampus tail (Figure 6: Panel B, third row). However, visual assessment of segmentation estimates and the standard-space FP map (Figure 6: Panel A, third row) indicate a 'greedy' labelling tendency, giving significantly higher FPR compared to all other methods (All: p_{fp} <.0005).

CF has significantly higher Dice coefficients and lower FPR compared to all other segmentation methods (FIRST: p_d=.001, p_{fp} <.0005; FIRSTv2: p_d=.004, p_{fp} <.0005; FS: p_d <.0005, p_{fp} <.0005; FMClose: p_d <.0005, p_{fp} <.0005). The standard-space FP and FN maps (Figure 6: fourth row) show a relatively low frequency of FP findings at medial and ventral borders of the hippocampus head, with FN counts in the most anterior-superior part of the hippocampus head.

FMClose performs on a par with FIRST, having significantly lower Dice coefficients and higher FNR compared to other methods (FIRSTv2: p_d <.0005, p_{fn} <.0005; FS: p_{fn} <.0005; CF: p_d <.0005, p_{fn}=.001). Whilst FMClose has a significantly reduced FPR compared to both FIRST

and FS (FIRST: p_{fp}=.001; FS: p_{fp} <.0005), it should be noted that spillover into the amygdala has been neglected in this analysis, so reported FPR is lower than actual FPR. Generally, the current implementation of FMClose suffers from a combination of spillover into anterior regions and under-estimation of the hippocampus tail, contributing to both FP and FN findings (Figure 6: fifth row). In addition, sub-hippocampal regions of contrasting intensity, such as the dentate gyrus, are excluded from the FMClose segmentation estimate and contribute further to FNR.

4. Discussion

When comparing the performance of segmentation methods on different datasets, it is important to appreciate that a number of factors can affect segmentation accuracy, including image quality, manual labelling protocol, clinical status and demographics. In addition, model-based methods may be biased towards data used in their training. Indeed, decreased performance of FIRST, and to a lesser extent FIRSTv2 and FS, on the BPSA data suggests that these model-based methods are biased towards the CMA data previously used in their training. The registration errors experienced with FIRST on the BPSA data are largely overcome with the hippocampus-specific spatial normalization in FIRSTv2, giving significantly improved segmentation estimates compared to FIRST. We predict similar improvements in registration accuracy and segmentation performance with other subcortical structures, using an appropriate structure-specific normalization step prior to model-fitting with FIRST.

For the BPSA data, previously unseen by any of the segmentation methods, CF clearly outperforms all other methods, with Dice coefficients comparable to those previously reported [AHH*07, HHA*06, LWK*10]. In particular, the improved segmentation result over FIRST and FS suggests a potential advantage of using flexible subject-specific selection methods like CF over more constrained model-based methods trained from more diverse data. Care should be taken, however, as results for the CMA data suggest that CF performance is restricted by reduced image quality and registration error. CF registration errors are caused by poorly-formed brain masks, which are themselves a consequence of too much neck present in the T1-weighted images and bias field effects (i.e. image quality). Ongoing work using cropped images and better brain masks have indicated that CF results for the CMA data can be improved by further preprocessing of images. Likewise, the performance of CF on all datasets can be expected to improve with further advances in the applied registration algorithms. In addition, it is possible that a sub-optimal number of classifiers are chosen for CF of the CMA dataset, reducing overall segmentation accuracy.

The current implementation of the FM algorithm performs surprisingly well against the other segmentation methods and highlights several important strengths and weaknesses of the different approaches, despite some limitations (namely spillover across reduced-contrast hippocampal boundaries and under-estimation of the hippocampus tail). Firstly, user input is required, with the resultant segmentation dependent on seed point selection and post-hoc manual thresholding; favouring less interactive (i.e. automated) approaches with increased robustness for clinical application. Secondly, the FM algorithm classifies voxels based on local intensity differences, so it excludes sub-hippocampal regions of contrasting intensity, such as the dentate gyrus, from the segmentation estimate. Although not beneficial in the current work, where total hippocampal segmentation is required, results suggest that intensity-dominant methods may be useful for segmentation of hippocampal sub-structures. Finally, whilst intensity information alone cannot prevent spillover across poor-contrast hippocampal boundaries, current extension of the FM algorithm to include a spatial prior promises improved segmentation estimates in a relatively time-efficient manner.

5. Conclusions

In summary, the work presented here is an evaluation of both current and novel MR hippocampal segmentation methods, highlighting general issues hampering the current clinical utility of these methods (potential model-bias, image preprocessing and registration errors) and therefore possible mechanisms for making them more robust and accurate in the presence of specific and substantial pathology. This work is an investigatory stepping-stone towards developing a dedicated hippocampal segmentation method for pathological subjects. As practical guidance towards the design and construction of this method, we suggest exploration of combinatorial methods that incorporate a spatially-varying weighting of subject-specific intensities and model-based shape features. Where possible (i.e. at good contrasting boundaries) subject-specific image information should drive the segmentation, with models used to constrain the segmentation at poorer contrasting borders such as the hippocampus-amygdala border.

6. Acknowledgments

With thanks to the LSI DTC, EPSRC and BBSRC David Phillips Fellowship for funding this research, David Kennedy and David Glahn for providing MR data and expert manual labels from the Center of Morphometric Analysis, Massachusetts, USA and the Research Imaging Center, University of Texas Health Science Center at San Antonio, Texas, USA, respectively.

References

[AHH*07] ALJABAR P., HECKEMANN R., HAMMERS A., HAJNAL J., RUECKERT D.: Classifier selection strategies for la-

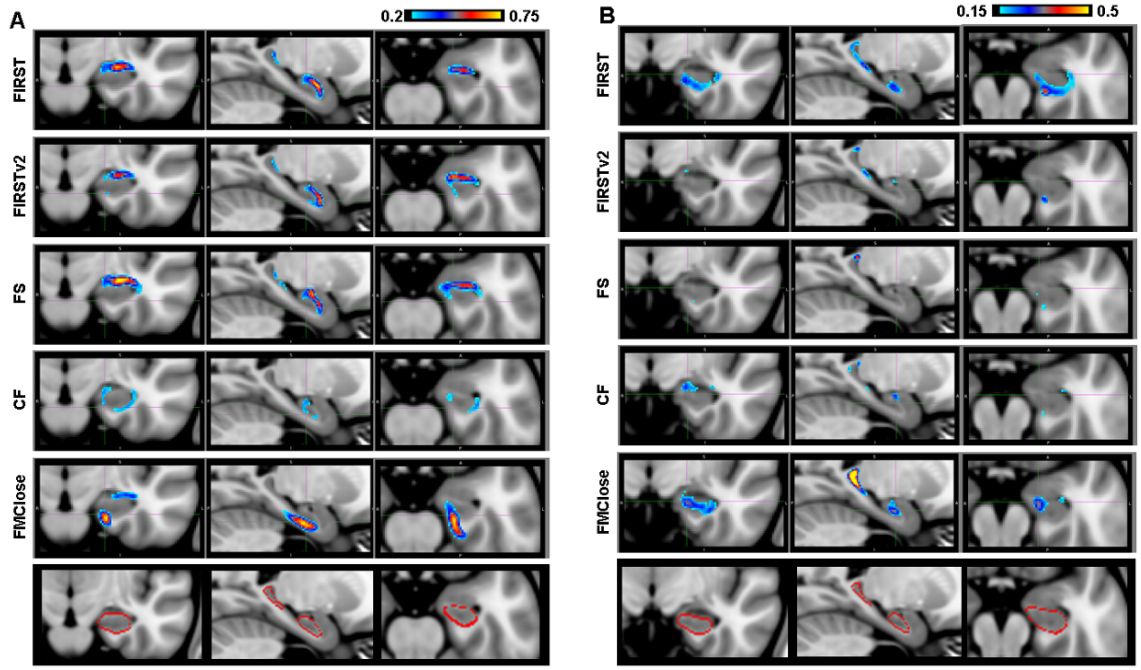

Figure 6: *Panels A and B show coronal- (left), sagittal- (middle) and axial- (right) view images of the standard-space FP and FN maps, respectively, for each segmentation method on the BPSA data. A voxel-wise threshold of 20% FP finding across the dataset is applied to all FP maps, whilst the FN maps are threshold at 15%. To aid visual assessment, the hippocampal boundary defined by FIRST on the MNI152 standard image is shown (sixth row) and the cursor position is the same for all images in each panel.*

bel fusion using large atlas databases. *In: MICCAI 4791* (2007), 523–531. 1, 2, 3, 7

[CCC*08] COLLIOT O., CHÉTELAT G., CHUPIN M., DESGRANGES B., MAGNIN B., BENALI H., DUBOIS B., GARNERO L., EUSTACHE F., LEHÉRICY S.: Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. *Radiology 248* (2008), 194–201. 1

[CMBH*07] CHUPIN M., MUKUNA-BANTUMBAKULU A., HASBOUN D., BARDINET E., BAILLET S., KINKINGNEHUN S., LEMIEUX L., DUBOIS B., GARNERO L.: Anatomically constrained region deformation for the automated segmentation of the hippocampus and amygdala: Method and validation on controls and patients with Alzheimer's disease. *NeuroImage 34* (2007), 996–1019. 1

[FSB*02] FISCHL B., SALAT 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 A.: Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron 33* (2002), 341–55. 2

[HHA*06] HECKEMANN R., HAJNAL J., ALJABAR P., RUECKERT D., HAMMERS A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage 33* (2006), 115–126. 1, 7

[KUN*09] KONRAD C., UKAS T., NEBEL C., AROLT V., TOGA A., NARR K.: Defining the human hippocampus in cerebral magnetic resonance images - An overview of current segmentation protocols. *NeuroImage 47* (2009), 1185–1195. 1

[LWK*10] LOTJONEN J., WOLZ R., KOIKKALAINEN J., THURFJELL L., WALDEMAR G., SOININEN H., RUECKERT D.: Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage 49*, 3 (2010), 2352–2365. 1, 3, 7

[MBC*08] MAROY R., BOISGARD R., COMTAT C., FROUIN V., CATHIER P., DUCHESNAY E., DOLLE F., NIELSEN P., TREBOSSEN R., TAVITIAN B.: Segmentation of rodent whole-body dynamic PET images: An unsupervised method based on voxel dynamics. *IEEE 27* (2008), 342–354. 3

[MWT*05] MUELLER S., WEINER M., THAL L., PETERSEN R., JACK C., JAGUST W., TROJANOWSKI J., TOGA A., BECKETT L.: The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N. Am. 15* (2005), 869–877. 1

[SPM*09] SHATTUCK D., PRASAD G., MIRZA M., NARR K., TOGA A.: Online resource for validation of brain segmentation methods. *NeuroImage 45* (2009), 431–439. 3

[WJP*09] WOOLRICH M., JBABDI S., PATENAUDE B., CHAPPELL M., MAKNI S., BEHRENS T., BECKMANN C., JENKINSON M., SMITH S.: Bayesian analysis of neuroimaging data in FSL. *NeuroImage 45* (2009), 173–186. 2