Novel Fast Marching for Automated Segmentation of the Hippocampus (FMASH): Method and Validation on Clinical Data

Courtney A. Bishop[a], Mark Jenkinson[a], Jesper Andersson[a], Jerome Declerck[b], Dorit Merhof[c]

[a]*FMRIB Centre, University of Oxford, UK*
[b]*Siemens Molecular Imaging, Oxford, UK*
[c]*Visual Computing, University of Konstanz, Germany*

Abstract

With hippocampal atrophy both a clinical biomarker for early Alzheimer's Disease (AD) and implicated in many other neurological and psychiatric diseases, there is much interest in the accurate, reproducible delineation of this region of interest (ROI) in structural MR images. Here we present *Fast Marching for Automated Segmentation of the Hippocampus* (FMASH): a novel approach using the Sethian Fast Marching (FM) technique to grow a hippocampal ROI from an automatically-defined seed point. Segmentation performance is assessed on two separate clinical datasets, utilising expert manual labels as gold standard to quantify Dice coefficients, false positive rates (FPR) and false negative rates (FNR). The first clinical dataset (denoted CMA) contains normal controls (NC) and atrophied AD patients, whilst the second is a collection of NC and bipolar (BP) patients (denoted BPSA). An optimal and robust stopping criterion is established for the prop-

Email address: `courtney.bishop@merton.ox.ac.uk` (Courtney A. Bishop)

agating FM front and the final FMASH segmentation estimates compared to two commonly-used methods: FIRST/FSL and Freesurfer (FS). Results show that FMASH outperforms both FIRST and FS on the BPSA data, with significantly higher Dice coefficients (0.80±0.01) and lower FPR. Despite some intrinsic bias for FIRST and FS on the CMA data, due to their training, FMASH performs comparably well on the CMA data, with an average bilateral Dice coefficient of 0.82±0.01. Furthermore, FMASH most accurately captures the hippocampal volume difference between NC and AD, and provides a more accurate estimation of the problematic hippocampus-amygdala border on both clinical datasets. The consistency in performance across the two datasets suggests that FMASH is applicable to a range of clinical data with differing image quality and demographics.
Keywords: Structural MRI, Hippocampus, Automated, Segmentation, Region-growing

1. Introduction

Crucially involved in episodic and spatial memory processes, the hippocampus has been implicated in the pathophysiology of many neurological and psychiatric diseases [Konrad et al., 2009]. Hippocampal atrophy is a clinical biomarker of both early Alzheimer's disease (AD) and temporal lobe epilepsy (TLE), with additional findings in post traumatic stress disorder and schizophrenia, and affective disorders such as major depression and bipolar disorder (BP). Consequently, there is much interest in the delineation and quantification of this region of interest (ROI) on structural magnetic resonance (MR) images, requiring accurate and reproducible segmentation methods. These methods face a complicated task; arising from the complex shape and large inter-subject variability of the hippocampus as well as the poor contrast 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 hippocampal segmentation methods can be broadly classified into three groups: region-growing, atlas-based, and parameterized-modelling approaches. Whilst some methods may still be identified as belonging to one particular classification [Taylor and Barrett, 1994; Carmichael et al., 2005; Heckemann et al., 2006; Aljabar et al., 2007; Patenaude, 2007], there is a progressive trend towards 'multi-approach' segmentation methods that combine aspects of two or more groups to achieve increased robustness and segmentation accuracy [Fischl et al., 2002; Chupin et al., 2007; Colliot et al., 2008; Morra et al., 2008; van der Lijn et al., 2008; Chupin et al., 2009b; Lotjonen et al., 2010; Wolz et al., 2010]. This work will focus on region-growing methods and parameterized-models, but will utilise aspects of atlas-based approaches. In turn, the resultant segmentation performance will be compared to two commonly-used methods: the parameterized model FIRST/FSL [Patenaude, 2007; Woolrich et al., 2009], and the combined atlas-based and parameterized-model Freesurfer (FS) [Fischl et al., 2002].

As the name suggests, region-growing methods typically grow a ROI from an initial start or seed point manually defined on the target image. The ROI grows and deforms according to various constraints (e.g. anatomical landmarks, surface curvature, energy constraints) imposed by the respective algorithms. Competitive region-growing from one or more seeds may be performed by comparison of border voxels to the first-order statistics of voxels that have already been absorbed (or added) to the ROI [Taylor and Barrett, 1994]. Combining this concept with a deformable model, Ashton et al. [1997] elastically deform a region from a line of seeds in the presence of constraining forces (elastic surface tension, deviation from the expected surface normal and resistance from surrounding tissue). Both the initial line of seeds and the deformation constraint are derived from manually-segmented contours in three orthogonal slices. In turn, a statistical model of hippocampal geometry is combined with elastic deformations, utilising displacement forces derived from matching local intensity profiles [Kelemen et al., 1999]. More recently, several methods have focused on the simultaneous extraction of both the hippocampus and the amygdala [Yang and Duncan, 2004; Yang et al., 2004; Chupin et al., 2007; Colliot et al., 2008; Chupin et al., 2009b]. Specifically, a competitive region-growing approach based on homotopically deforming regions is used to segment both the hippocampus and the amygdala in Chupin et al. [2007]. This initial development of the algorithm requires manual definition of two seed points and a bounding box, with the deformation constraint based on prior knowledge of anatomical features in the vicinity of automatically-detected landmarks. Applied to clinical data [Colliot et al., 2008], the algorithm detects significant hippocampal volume reductions for both mild cognitively impaired (MCI) and AD patients with respect to normal controls, facilitating accurate group classification. Subsequent development has given rise to an automated region-growing method [Chupin et al., 2009b], imposing additional energy constraints and automated detection/correction of atlas mismatch for improved segmentation performance, and similar group classification accuracy [Chupin et al., 2009a]. Instead of initializing the algorithm with a manually-defined seed point, Chupin et al. [2009b] utilise the probabilistic atlas priors for automated intialization, defining much larger hippocampal and amygdala objects from the maximal probability zone of the respective atlases. As such, the method is essentially a voxel re-classification approach on these initial objects.

Whilst the aforementioned region-growing methods can provide relatively efficient segmentation accuracy and group classification, most require time-consuming and knowledge-based manual intervention (for definition of a specific bounding box and seed points), with the resultant segmentation dependent on good initialisation. Furthermore, the majority of studies focus on development and evaluation of an individual segmentation method on a single dataset, reporting a non-standardized selection of performance measures, such that direct comparison of hippocampal segmentation methods is compromised.

Here we present *Fast Marching for Automated Segmentation of the Hippocampus* (FMASH): a novel region-growing method for automated hippocampal segmentation on structural MR data. The algorithm uses a 3D Sethian Fast Marching (FM) technique (Section 2.5) to propagate a hippocampal ROI from an automatically-defined seed point p_0 (Section 2.4). The propagating ROI is driven by a potential function (Section 2.7) comprised of both subject-specific intensity features and a model-based shape prior. In contrast to Chupin et al. [2009b], the FMASH algorithm extracts a subject-specific, single-voxel hippocampal seed from the structural MR data (based on local means analysis) and does not rely on competitive growth of the neighbouring amygdala structure. The FMASH algorithm is applied to two clinical datasets (Section 2.1) with segmentation performance assessed according to three common measures (Section 2.9), providing both valuable assessment of method performance on multiple datasets and ease of comparison with previous reported methods. An optimal and robust stopping criterion is established for the propagating FM front (Section 3.3) and the resultant FMASH segmentation compared to FIRST and FS (Section 3.4). Additionally, Section 4 provides context based on the most recent literature.

2. Materials and methods

2.1. Clinical MR data

This work utilises two clinical datasets, consisting of T1-weighted MR images and their corresponding expert manual labels. The first dataset contains 9 normal control (NC) subjects and 8 atrophied AD patients (Figure 1), supplied by the Center for Morphometric Analysis (CMA), whilst the second is a collection of 16 NC and 16 BP patients from the University of Texas Health Science Center in San Antonio (denoted BPSA). Evaluation of the FMASH algorithm on both of these datasets is motivated by the desire to construct a general hippocampal segmentation method that works well for a range of diseases.

The research sites that provided the data used their own expert raters and semi-automated contouring tools to perform a similar manual definition of the hippocampus in sequential coronal slices, based on intensity boundaries and well-established geometrical rules of neuroanatomy [Kennedy et al., 1989].

All experts underwent a period of training (of up to three months) until they had reached a defined reproducibility. Inter-rater reliability (i.e. Dice coefficient) 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	9 NC, 8 AD
BPSA	32	20 - 58	0.8 x 0.8 x 0.8	16 NC, 16 BP

Table 1: Image resolution and demographics. Group size, age (years) and resolution of the two clinical datasets, containing normal control (NC) subjects, Alzheimer's Disease (AD) and bipolar (BP) patients.

2.2. Software

Tools from the FMRIB Software Library (FSL, www.fmrib.ox.ac.uk/fsl) are used to generate the hippocampal spatial priors (Section 2.3) and standard-space performance maps (Section 2.9), with the remainder of the algorithm currently implemented in MATLAB (www.mathworks.com).

2.3. Generation of hippocampal spatial priors

Atlas-based, label-propagation and decision-fusion approaches have previously provided segmentation estimates for the hippocampus, with label-fusion performed either in the native (subject) space [Heckemann et al., 2006; Aljabar et al., 2009; Bishop et al., 2010] or a standard reference space [Aljabar et al., 2007]. Here, a similar approach is used to construct the hippocampal spatial priors, propagating all subject data (MR images and corresponding manual labels) to the standard MNI152 space via a multi-step registration scheme: an initial three-step affine registration with the FSL tool *FLIRT*, followed by a multi-resolution non-linear warp with *FNIRT*. From this repository of spatially-normalized data, the target (to-be-segmented) image is excluded, manual labels are fused (summing trilinearly-interpolated values) and averaged to generate a standard-space hippocampal prior for each individual subject. Each inverted non-linear warp subsequently transfers the standard-space hippocampal prior to the corresponding subject (native) space for use in the FMASH algorithm. For all non-linear registrations with FNIRT, a control point spacing of 10mm was specified, with a sub-sampling scheme of 4,4,2,2,1,1 and a bending energy model for regularisation of the warp-field.

2.4. Image preprocessing and seed extraction

For each subject, the minimum and maximum index of non-zero voxels along each dimension of the hippocampal spatial prior defines a rectangular box (i.e. cuboid) surrounding the hippocampus. Maroy's Local Means Analysis (LMA) method for dynamic PET images [Maroy et al., 2008] is adapted to extract a series of points with minimal local intensity variance within this ROI. For all voxels n in the hippocampal ROI, the amplitude of local signal variations are estimated by

$$\Gamma_n = \frac{1}{\#V_n - 1} \times \sum_{j \in V_n} \frac{(I_j - \mu_n)^2}{\mu_n} \tag{1}$$

where $\#V_n$ is the number of voxels contained within the local neighbourhood V_n of voxel n, μ_n is the mean signal within V_n, and I_j is the signal at a neighbouring voxel j. Spatial neighbourhoods of 5x5x5 and 5x3x5 voxels are considered for the BPSA and CMA data, respectively, maintaining a similar spatial extent (in mm) of V_n for differing voxel dimensions.

We define $\Lambda = \{n \,|\, \forall j \in V_n, \Gamma_n < \Gamma_j\}$ the set of locations of all local minima of the resulting "Γ map" (i.e. the set of voxels that are local minima with respect to the local intensity variance measure). These local minima are points of stability in the hippocampal ROI, usually the least affected by noise and image artefacts. Defining the most central local minimum as the single hippocampal seed point $m = p_0$ thus provides stable and robust initialization of the FM algorithm.

2.5. Active contours: a global minimum

Classical deformable models are typically driven by an energy term composed of both internal and external constraints, requiring precise initialization for extracting the path of minimum energy between two fixed extremities. Applied to aerial road images, angiographic images of the eye and cardiac MR images, Cohen and Kimmel [1997] simplify this model to external forces only and solve the minimal energy path problem in 2D, using the Sethian FM algorithm for propagation of the initial front [Sethian, 1996]. Later, the FM method is extended to 3D images [Deschamps and Cohen, 2000] with application to path tracking in virtual endoscopy, computed tomography (CT) colonoscopy and extraction of brain vessels in a magnetic resonance angiography (MRA) scan. These minimal energy path active contour approaches aim to find the curve of minimum energy amongst all curves joining a seed point p_0 to any other point p. A path C(s) in the image is found that minimises the energy

$$E(C) = \int_{\Omega} \tilde{P}(C(s))ds \tag{2}$$

where Ω is the curve domain $[0, L]$, L is the length of the curve and $\tilde{P}$ is the input potential. The arrival time surface or minimal action U, is then defined as the minimal energy integrated along the path

$$U(p) = \inf_{\mathcal{A}_{p_0,p}} E(C) = \inf_{\mathcal{A}_{p_0,p}} \left\{ \int_{\Omega} \tilde{P}(C(s))ds \right\} \tag{3}$$

with $\mathcal{A}_{p_0,p}$ denoting the set of all possible paths between p_0 and p. The arrival time surface is so-called because it gives the arrival time of the propagating front at any given point in 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 $\tilde{P}$, thus finding the minimum energy curve for the structure of interest (in this case, the hippocampus).

2.6. FM resolution in 3D

Noticing that the arrival time surface U satisfies the Eikonal equation

$$\|\nabla U\| = \tilde{P} \tag{4}$$

Sethian [1996] presents a fast and efficient method for its construction in 2D, based on the fact that information is propagating *outwards* from the initial seed point (p_0), needing only one pass on the image. Extended to 3D [Deschamps and Cohen, 2000], the details of the FM algorithm are as follows

1. *Initialization*

    - For each grid point p, let $U(p) = \infty$ (large positive value). Label all points as *far*.
    - Set the start point p_0 to be zero: $U(p_0) = 0$ and label it *trial*.

2. *Marching Loop*

- Let $(i_{min}, j_{min}, k_{min})$ be the trial point with the smallest U value.
- Label the point $(i_{min}, j_{min}, k_{min})$ as *alive*, and remove it from the trial list.
- For each of the 6 neighbouring grid points (i, j, k) of $(i_{min}, j_{min}, k_{min})$
  - if (i, j, k) is labelled as far, then label it trial;
  - if (i, j, k) is *not* alive, then compute $U_{i,j,k}$ according to the following equation

$$\begin{aligned}
(max\left\{u - min\left\{U_{i-1,j,k}, U_{i+1,j,k}\right\}, 0\right\})^2 &\quad + \\
(max\left\{u - min\left\{U_{i,j-1,k}, U_{i,j+1,k}\right\}, 0\right\})^2 &\quad + \\
(max\left\{u - min\left\{U_{i,j,k-1}, U_{i,j,k+1}\right\}, 0\right\})^2 &\quad = \quad \tilde{P}_{i,j,k} \qquad (5)
\end{aligned}$$

    and let $U_{i,j,k} = u$.

With each iteration through the *Marching Loop*, the seed point p_0 propagates outwards along the path of least resistance, adding a voxel which constitutes the lowest potential $\tilde{P}$, thus finding the (minimal action) arrival time surface U. Here, U is computed for all voxels in the hippocampal ROI, before thresholding to give the resultant segmentation estimate. Computation of an accurate and robust stopping criterion is detailed in Section 2.10. Novelty arises from automated initialization of the propagating front using a single-voxel hippocampal seed derived from the structural MR data (Section 2.4), and the potential function presented below.

2.7. A novel potential function

We define a novel input potential $\tilde{P}_m$ at voxel n, associated with the curve joining the seed point m to any other voxel n:

$$\tilde{P}_m(n) = \frac{(I_n - \mu_m)^2}{{\mu_m}^2} - \lambda \log p\,(Hipp\,(n)) \tag{6}$$

This potential utilises both image intensity features (first term) and prior probabilities established from training data (second term) to compute the path of lowest potential and least resistance for the propagating FM front. The probability of voxel n being labelled as hippocampus in the training data is $p(Hipp(n))$, with a weighting parameter λ for the prior term.

Initial work, utilising only an intensity-based potential function, demonstrated the need for the hippocampal spatial prior; providing further constraint of the propagating FM front at poor contrast intensity boundaries, such as the hippocampus-amygdala border. The form of the potential function is based on the observation that the first term takes the form of a log Gaussian likelihood probability, and thus the natural Bayesian formulation for a log posterior probability involves the addition of a log prior probability term. The weighting parameter λ provides a relative normalisation of the likelihood and prior terms, as is common in such posterior probabilities. In addition, as each term is dimensionless, λ is independent of the scale of the image intensities and should therefore remain constant over a large range of datasets.

2.8. The weighting parameter λ

Where possible, we wish the propagating FM front to be driven by subject-specific intensity features, only incorporating model-based shape features when intensity contrast is lacking. Consequently, we search for the minimum value of the parameter λ that constrains the FM front at these poor-contrast boundaries. For a given λ and seed point m, we can compute the potential at every voxel n in the hippocampal ROI (Equation 6) and use this information to generate a 3D potential map. By empirically tuning the value of λ (searching over the range 0 to 100 using increments of 10, then 1, then 0.01), a series of 3D potential maps are generated for a small subset of test subjects (2 NC and 2 patients) in each dataset. Section 3.1 describes the impact of λ on these 3D potential maps and its determined value. The value of λ is subsequently fixed for all subjects.

2.9. 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 [Shattuck et al., 2009] and the MICCAI 2008 competition workshop on MS lesion segmentation: Dice coefficient, false positive rate (FPR) and false negative rate (FNR). Additionally, these three metrics are combined to give an indication of overall segmentation performance, as described below. The performance measures serve two purposes: firstly, during development of the FMASH algorithm, they are used to compute the FM stopping criterion (Section 2.10), and secondly, they evaluate the resultant FMASH segmentation (Section 2.11).

If X is the set of all voxels in the image, we define the gold standard (i.e. truth set) $T \subset X$ as the set of voxels labelled as hippocampus by the expert manual rater. Similarly, we define the set $S \subset X$ as the set of voxels labelled as hippocampus by the segmentation algorithm or method being tested.

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:

$$\text{FPR} = \frac{\mid FP \mid}{\mid FP \mid + \mid TN \mid} = 1 - \text{ specificity} \tag{7}$$

$$\text{FNR} = \frac{\mid FN \mid}{\mid FN \mid + \mid TP \mid} = 1 - \text{ sensitivity} \tag{8}$$

It is worth noting that the size of the image field of view (FOV) affects the FPR but not the FNR, such that the FPR are very low when considered over the whole image as opposed to a hippocampal ROI. Since we define the hippocampal ROI as the minimum and maximum index of non-zero voxels along each dimension of the hippocampal spatial prior, the size of the ROI (and FOV) varies across individuals. To maintain a constant FOV, we consider the whole MR image in our calculations of the TN set and therefore the resultant FPR are extremely low.

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 three-step affine transformation, computed during generation of the hippocampal spatial priors, propagates all subjects' FP and FN voxels from their native space to MNI152 space. Here, the average FP and FN maps are computed (summing trilinearly-interpolated values), with voxel values corresponding to the fraction of subjects showing a FP or FN result at that voxel.

Similarity metric: Dice coefficient

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

$$D(T,S) \quad = \quad \frac{2 \mid T \cap S \mid}{(\mid T \mid + \mid S \mid)} \tag{9}$$

Combined Performance Measure (CPM)

Segmentation estimates of the left and right hippocampi are obtained for each segmentation method, and the corresponding performance measures computed. These measures are normalized to 0-1 (inverting FPR and FNR such that a high normalized measure corresponds to better segmentation performance) and summed to give the *combined performance measure* (CPM). The CPM can take any value between 0 and 3, corresponding to the minimum and maximum obtainable segmentation performance, respectively.

2.10. FM stopping criterion

During development of the FMASH algorithm, we compute the stopping time, U, for the propagating FM front using the following approach. Firstly, we threshold each subjects' left and right hippocampal arrival time surface at a range of different arrival times ($0.12 \leq U \leq 0.30$ in increments of 0.02), generating a series of left and right segmentation estimates for each subject, from which the corresponding performance measures are computed. The arrival time corresponding to the segmentation with the *maximum* combined performance measure (CPM: Section 2.9) defines the optimal stopping criterion for the propagating FM front. In this way, we compute the optimal stopping time for the left and right hippocampus of each subject.

Within each clinical dataset, we then adopt a leave-one-out approach to compute an independent (i.e. unbiased) stopping time for each subject. This involves excluding the target (to-be-segmented) subject and averaging the remaining optimal stopping times for the other subjects in that dataset. These independent stopping times are extremely stable across all subjects and both clinical datasets (Section 3.3). Consequently, we compute the dataset-average of these independent stopping times, which is identical for both clinical datasets. This single stopping time is subsequently used to threshold every subjects' arrival time surface, with the resulting FMASH segmentations used for validation of the algorithm, as described below.

2.11. Segmentation evaluation

Following development of the FMASH algorithm and computation of the FM stopping criterion, FMASH segmentation performance is compared to

FIRST and FS using the aforementioned performance measures and a statistical analysis technique: *analysis of variance* (ANOVA). The left and right hippocampal performance measures (Dice, FPR, FNR and CPM) are computed for each method on both clinical datasets. For each of these measures, we test for overall difference across the segmentation methods and clinical group-by-method interactions using a 3*2 (method*group) mixed ANOVA design, with method as the within-subjects factor and clinical group as the between-subjects factor. Where Mauchly's test indicates that the assumption of sphericity for method has been violated (p<.05), degrees of freedom are corrected using Greenhouse-Geisser estimates of sphericity. Paired-sample t-tests with Bonferroni correction for multiple comparisons are used to identify between-method differences; Lp_d, Rp_d, Lp_{fp}, Rp_{fp}, Lp_{fn}, Rp_{fn}, Lp_{cpm}, Rp_{cpm} denote bilateral corrected p-values for Dice coefficient, FPR, FNR and CPM, respectively.

Finally, to investigate the dependence of the FMASH algorithm on the hippocampal spatial prior, segmentation estimates are obtained for both clinical datasets using a series of different priors, and performance measures computed. For the results presented herein (Figures 3 - 7), a leave-one-out approach is adopted for generation of a general, mixed-dataset prior (representative of differing labelling protocols and expert raters) with no bias towards the target (to-be-segmented) image anatomies. In addition, we explore the cross-dataset performance; segmenting the CMA data using a single-dataset prior generated from the BPSA labels only, and vice versa. For both clinical datasets, the left and right hippocampal performance measures are computed, and the aforementioned ANOVA designs repeated to evaluate

FMASH performance using the single-dataset priors.

3. Results

The results may be considered in two parts: Sections 3.1 - 3.3 refer to the development of the FMASH algorithm, whilst Sections 3.4 - 3.5 detail the segmentation evaluation results. For both datasets, we present box plots of FMASH segmentation performance at the ten arrival time thresholds and 3D scatter plots of the normalized performance measures for the left and right hippocampus (Figures 3 and 4, respectively), computing the FM stopping criterion. In turn, Figures 5 - 7 compare the resultant FMASH segmentation with those obtained from FIRST and FS.

3.1. Determination of the weighting parameter λ

The weighting parameter λ controls the contribution of the hippocampal spatial prior towards the FM input potential. At $\lambda = 0$, the potential function depends only on image intensities (first term), with features on the potential map corresponding to intensity features in the MR image. Consequently, at good contrast boundaries in the MR image, the potential map is similarly well-defined, showing the intricate details and complexity of the hippocampus boundary and providing sufficient resistance for the propagating FM front. However, at poor contrast intensity boundaries, such as the hippocampus-amygdala border, there is inconsistent contrast on the potential map and therefore insufficient resistance to stop region growth; regions of lower contrast can be thought of as *holes* in the hypothetical barrier provided by high contrast regions on potential map, such that the FM front spills-out across the true hippocampal boundary. As the value of λ increases from zero, previously well-defined features on the potential map are smoothed by incorporation of probabilistic shape features (second term), whilst poor contrast boundaries become more defined. We therefore search for a value of λ that provides sufficient resistance (i.e contrast) on the potential map at ill-defined intensity boundaries in the MR image, whilst maintaining subject-specific features at good contrast hippocampal borders. Visual assessment of the 3D potential maps for a small subset of test subjects (Section 2.8), and the contrast at the hippocampal borders, estimates the value of λ at 0.05 for both datasets.

3.2. FMASH performance at varying arrival times

As shown in Figures 3 and 4, FMASH segmentation performance depends on the arrival time threshold applied to the corresponding arrival time surface. In the lower range, Dice coefficients increase with U, peaking at 0.16-0.18 on both datasets (Figures 3 and 4 (A): first row) before subsequently decreasing. As U increases, the FM front propagates further and obtains progressively larger estimates of hippocampal volume, with a resultant increase in FPR (Figures 3 and 4 (A): second row). On the other hand, a progressively larger hippocampal volume reduces the tendency for FN findings, with FNR showing an inverse dependency on U (Figures 3 and 4 (A): third row). Clearly, no single arrival time is optimal for all three performance measures; there is always a trade-off or compromise to be made. In Section 3.3 below, we present the results of combining these performance measures and computing the FM stopping criterion.

3.3. An accurate and robust stopping criterion

3D plots of the left and right hippocampal normalized performance measures are given in Figures 3 and 4 (B) for the CMA and BPSA data (left and right columns, respectively). A good arrival time threshold corresponds to high Dice coefficients, low FPR and low FNR, and will therefore cluster in the top, right-back corner of the plots. As previously described (Section 2.10), we define the arrival time corresponding to the *maximum* combined performance measure (CPM) as the optimal stopping criterion for the propagating FM front, before utilising a leave-one-out approach to compute an independent (i.e. unbiased) stopping time for each subject. The (left, right) mean$\pm$SD of these independent stopping times are 0.18$\pm$0.00, 0.18$\pm$0.00 and 0.18$\pm$0.01, 0.17$\pm$0.01 for the CMA and BPSA data, respectively, giving the same bilateral-average stopping time of 0.18 for each dataset. Consequently, the stability of the stopping criterion, across all subjects and both clinical datasets, justifies thresholding every subjects' arrival time surface at this single stopping time of 0.18. We compare the resultant segmentation performance of FMASH with that of FIRST and FS below.

3.4. Comparison to FIRST and FS

Using the statistical analysis technique, ANOVA, FMASH segmentation performance is compared to FIRST and FS. In general, results of the ANOVA designs reveal a significant effect of method for each performance measure, but no clinical group-by-method interaction or group effect for either clinical dataset. The only exceptions to this trend are a slightly significant group-by-method interaction ($F(1.95,29.3)=4.65$, $p=.02$) and group effect ($F(1,15)=7.35$, $p=.02$) for the right hippocampal FPR on the CMA data, and no significant method effect for the right hippocampal FNR on the CMA data. Whilst FMASH FPR shows negligible bilateral group effect on the CMA data, both FIRST and FS have considerably higher FPR for NC compared to AD patients: a bilateral increase of 23% is observed for FIRST, but the effect is even more pronounced for FS, with a 63% increase for the left hippocampus and a two-fold increase for the right hippocampus.

On the CMA data (Figure 5: top row), FMASH obtains significantly higher Dice coefficients (left: 0.82±0.01; right: 0.82±0.01) and lower FPR (left: 7.38±0.00 x10^{-5}; right: 6.33±0.00 x10^{-5}) compared to FS (Lp_d <0.001, Rp_d=0.009, Lp_{fp} <0.005, Rp_{fp} <0.001), with FNR comparable to FIRST. Furthermore, FMASH outperforms both FIRST and FS on the BPSA data (Figure 5: bottom row), with significantly higher Dice coefficients (left: 0.79±0.01; right: 0.80±0.01) and lower FPR (left: 1.03±0.25 x10^{-4}; right: 1.11±0.24 x10^{-4}; FIRST: Lp_d=0.01, Rp_d, Lp_{fp}, Rp_{fp} all <0.005; FS: Lp_d, Rp_d, Lp_{fp}, Rp_{fp} all <0.005).

Just as we used the CPM at different arrival times to find the FM stopping criterion, the CPM for each segmentation method reveals a strict overall ranking of the methods on each dataset. Considering both the left and right hippocampus of all subjects, we have 34 segmentation estimates with corresponding CPM for each method on the CMA data, and an additional 64 measurements for the BPSA data. FMASH has significantly higher CPM on the BPSA data compared to both FIRST and FS (FIRST: Lp_{cpm}=0.06, Rp_{cpm} <0.005; FS: Lp_{cpm}, Rp_{cpm} both <0.005), obtaining the highest CPM for 86% of segmentations (left: 2.54±0.17; right: 2.62±0.14), with FIRST performing best for the remaining 14%. The ordering of FIRST and FMASH is reversed on the CMA data: FMASH performs best for 15% of segmentations, whilst FIRST obtains the highest CPM for 85% (left: 2.08±0.28; right: 2.28±0.36). Consequently, FIRST has significantly higher CPM than both FMASH and FS on the CMA data (FMASH: Lp_{cpm}, Rp_{cpm} both <0.005; FS: Lp_{cpm}, Rp_{cpm} both <0.005). We discuss this heightened performance of FIRST on the CMA data in Section 4. Of final note here is the failure of FS to obtain a maximum CPM for any subject in either dataset, with a reduced average CPM of 33% and 16% relative to the best performing method on the CMA and BPSA data, respectively.

The standard-space FP maps for the CMA data (Figure 6: left panel) show a low frequency of FP voxels for FIRST, with only slight over-estimation (or spillover) at the most medial-inferior boundary of the hippocampus head; a pattern similarly displayed by the FMASH algorithm. FS, on the other hand, displays FP voxels in the body of the hippocampus and more severe spillover in the most anterior-superior regions of the hippocampus head (i.e. hippocampus-amygdala border). On the BPSA data (Figure 6: right panel), both FIRST and FS show a marked increase in FP findings, with severe spillover at the hippocampus-amygdala border. In contrast, FMASH displays a low frequency of FP voxels at this boundary, and reduced spillover at the most medial-inferior boundary compared to both itself and FIRST on the CMA data.

Additionally, the standard-space FN maps for the CMA data (Figure 7: left panel) show that medial boundaries and posterior regions of the hippocampus head are underestimated by FIRST (top row) and to a lesser extent FS (second row), whilst FMASH displays FN voxels at the most anterior tip of the hippocampus head (third row). FS maintains a low frequency of FN voxels on the BPSA data (Figure 7: right panel, second row). However, FIRST severely underestimates medial-inferior borders (top row) and FMASH has a propensity for FN voxels in the most posterior-superior regions of the hippocampus tail (third row).

3.5. Evaluation of the hippocampal spatial prior

Investigating the dependence of the FMASH algorithm on the hippocampal spatial prior reveals that, on average, the single-dataset prior results in a 1% drop in Dice coefficients for the CMA data and a 4% drop for the BPSA data, compared to the results presented in Figure 5 (left column) for the mixed-data prior. Despite this reduction in Dice coefficients using the single-dataset prior, FMASH still obtains significantly lower FPR and higher CPM than FS on the CMA data (FS: Lp_{fp}, Rp_{fp}, Lp_{cpm}, Rp_{cpm} all <0.005), with FNR comparable to FIRST. For the BPSA data, FMASH retains the highest mean Dice coefficient and lowest FPR of all methods, with significance achieved against the FPR of FS (Lp_{fp}, Rp_{fp} both <0.005). Furthermore, FMASH still has the highest CPM of all methods on the BPSA data (left: 2.33±0.18; right: 2.46±0.15), with significantly improved segmentation performance compared to FS (Lp_{cpm}, Rp_{cpm} both <0.005).

4. Discussion and conclusions

We find that FMASH segmentation performance depends on the arrival time threshold applied to the corresponding arrival time surface. No single arrival time is optimal for all three performance measures; there is always a trade-off or compromise to be made. We presented the results of combining these performance measures and computing the stopping time for the propagating FM front (Section 3.3). With the same value of λ (0.05) and the same bilateral-average stopping time (0.18) computed for both datasets, results suggest that we have a robust stopping criterion that is applicable to a range of clinical data with differing image quality, disease-status and demographics.

As previously reported [Bishop et al., 2010], both model-based methods FIRST and FS have an inherent bias towards their training data, of which this CMA data forms a small subset. Excluding this CMA subset from a re-trained FIRST model did not alter its segmentation performance (results not shown), suggesting that FIRST results are dataset dependent, rather than biased explicitly by inclusion of this particular CMA subset. Nevertheless, we expect FIRST and FS to dominate on this CMA data, but results presented herein demonstrate that this is not the case. FMASH still obtains significantly higher Dice coefficients and lower FPR compared to FS, lower FNR than FIRST, and a maximum CPM for 15% of segmentations on the CMA dataset. In contrast, FS fails to obtain a maximum CPM for any subject in either dataset, most likely due to a slight "greedy labelling" tendency yielding a higher FPR.

Comparing FMASH performance with that of FIRST and FS also reveals a surprising group-wise discrepancy in FPR for the two parameterized models: both FIRST and FS have considerably higher FPR for NC compared to AD patients, with this effect even more pronounced for FS, and in particular, the right hippocampus. This finding suggests that these commonly-used methods exaggerate the true hippocampal volume difference between NC and AD patients, by as much as two-fold, and that FS has a tendency to exaggerate these differences more for the right hippocampus; providing a strong argument for using the proposed FMASH method instead.

Previous studies comparing FIRST and FS [Morey et al., 2009; Pardoe et al., 2009], report Dice coefficients in the range of 0.71-0.80 for FIRST and 0.73-0.82 for FS on NC data. The variability in Dice coefficients for FIRST and FS on the CMA and BPSA data (0.71-0.85 and 0.74-0.80, respectively) are therefore in-line with previous studies, with the exception of the high Dice coefficients for FIRST on the CMA data (discussed above).

Here we provide context for the FMASH algorithm based on the most current literature, comparing FMASH Dice coefficients with those of recently-published methods. In the literature, mean Dice coefficients are either separately reported for NC and diseased patients, or as reported herein, for mixed-cohorts (containing both NC and diseased patients). Recent atlas-based approaches report Dice coefficients in the range 0.75-0.83 for NC [Carmichael et al., 2005; Heckemann et al., 2006; Aljabar et al., 2007, 2009], 0.74-0.76 for TLE patients (diseased side) [Hammers et al., 2007; Avants et al., 2010] and 0.80 for mixed-cohorts [Leung et al., 2010]. However, values as high as 0.86 for NC [van der Lijn et al., 2008] and 0.82-0.88 for mixed-cohorts [Lotjonen et al., 2010; Wolz et al., 2010] can be achieved with graph-cuts, embedded learning, and intensity modelling, respectively. Lotjonen et al. [2010] suggest differences in image quality, manual protocol, clinical status and demographics as possible causes of discrepancy between their two clinical datasets, which may also be contributing factors herein. The aforementioned semi-automated region-growing method of Chupin et al. [2007] finds an average Dice coefficient of 0.84 for NC data, whilst their more recent automated method utilises a probabilistic prior, with automatic detection and correction of atlas mismatch to achieve Dice coefficients of 0.87$\pm$0.03 for NC and 0.84$\pm$0.05 for a mixed-cohort of NC and TLE patients [Chupin et al., 2009b]. The improved segmentation performance of Chupin et al. [2009b] on the NC data is not surprising given that the probabilistic prior is generated from NC subjects, segmented by the same investigator and using the same manual labelling protocol. The FMASH bilateral-average Dice coefficients of 0.82$\pm$0.01 and 0.80$\pm$0.01 for the CMA and BPSA dataset, respectively, are nevertheless comparable with the current literature for mixed-cohorts. This said, there are possible improvements and extensions to be made to the FMASH method. For example, automated detection and correction of atlas mismatch, as implemented in Chupin et al. [2009b], and generation of hippocampal priors using alternative registration algorithms, could provide improved segmentation performance. FMASH group classification accuracy could be investigated with application of the algorithm to publicly available data, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI, www.loni.ucla.edu/ADNI). Additionally, with an appropriate spatial prior, the FMASH algorithm has the ability to segment other brain structures, such as the neighbouring amygdala.

Finally, even with a small weighting parameter λ, we appreciate that there is a dependence of the FMASH algorithm on the hippocampal spatial priors, generated from a set of manually-labelled images. The leave-one-out approach, used to generate the mixed-dataset priors and the results in

Figures 3 - 7, ensures no bias towards the target (to-be-segmented) image anatomies, whilst incorporation of labels from both datasets aims to generate more general priors that are representative of differing labelling protocols and expert raters. We do, however, also explore the cross-dataset performance; segmenting the CMA data using a spatial prior generated from the BPSA labels only, and vice versa. These single-dataset priors favour a specific labelling protocol and expert raters, so we expect differences in segmentation performance compared to using the mixed-dataset hippocampal priors. Dice coefficients are reduced using the single-dataset priors, but FMASH cross-dataset performance is still superior to FS on the CMA data and has the highest overall performance of all methods on the BPSA data. Furthermore, the observed difference in performance (between the mixed- and single-dataset priors) is comfortably within the range of inter-rater variability, differences in labelling protocol and registration error, although it does not rule-out bias of the mixed-dataset priors towards the two clinical datasets. Future work will look to explore this effect further. Nevertheless, for the aforementioned reasons, we recommend use of a mixed-data hippocampal prior for the FMASH algorithm and we are confident that this will result in a more robust and accurate segmentation method.

Overall, this novel hippocampal segmentation method shows extremely good, consistent performance on the two clinical datasets compared to the most widely-used alternatives, with more accurate estimation of the problematic hippocampus-amygdala border. This fully-automated approach also performs comparably well against the most recently-published methods, as described above. Primarily driven by subject-specific intensities in the MR image, this method is capable of capturing both subject- and disease-specific features, with results suggesting no inherent bias towards either dataset. We envisage accurate and robust hippocampal segmentation estimates using this fully-automated method on a range of clinical datasets, with differing image quality, disease-status and demographics.

5. Acknowledgments

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

6. References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D., 2007. Classifier selection strategies for label fusion using large atlas databases. In: MICCAI 4791, 523–531.

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage 46, 726–738.

Ashton, E., Parker, K., Berg, M., Chen, C., 1997. A novel volumetric feature extraction technique with applications to MR images. IEEE Trans. Med. Imag. 16 (4), 365–371.

Avants, B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J., 2010. The optimal template effect in hippocampus studies of diseased populations. NeuroImage 49 (3), 2457–2466.

Bishop, C., Jenkinson, M., Declerck, J., Merhof, D., 2010. Evaluation of hippocampal segmentation methods for healthy and pathological subjects. EG VCBM, 17–24.

Carmichael, O., Aizenstein, H., Davis, S., Becker, J., Thompson, P., Meltzer, C., Liu, Y., 2005. Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment. NeuroImage 27, 979–990.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., the Alzheimer's Disease Neuroimaging Initiative, 2009a. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19, 579–587.

Chupin, M., Hammers, A., Liu, R., Colliot, O., Burdett, J., Bardinet, E., Duncan, J., Garnero, L., Lemieux, L., 2009b. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: Method and validation. NeuroImage 46, 749–761.

Chupin, M., Mukuna-Bantumbakulu, A., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnehun, S., Lemieux, L., Dubois, B., Garnero, L., 2007. 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, 996–1019.

Cohen, L., Kimmel, R., 1997. Global minimum for active contour models: A minimal path approach. Int. J. Comp. Vis. 24, 57–78.

Colliot, O., Chételat, G., Chupin, M., Desgranges, B., Magnin, B., Benali, H., Dubois, B., Garnero, L., Eustache, F., Lehéricy, S., 2008. Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 248, 194–201.

Deschamps, T., Cohen, L., 2000. Minimal paths in 3D images and application to virtual endoscopy. Lecture Notes in Computer Science 1843, 543–557.

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., 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–55.

Hammers, A., Heckemann, R., Koepp, M., Duncan, J., Hajnal, J., Rueckert, D., Aljabar, P., 2007. Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: A proof-of-principle study. NeuroImage 36 (1), 38–47.

Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115–126.

Kelemen, A., Szekely, G., Gerig, G., 1999. Elastic model-based segmentation of 3-D neuroradiological data sets. IEEE Trans. Med. Imag. 18 (10), 828–839.

Kennedy, D., Filipek, P., Caviness, V., 1989. Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging. IEEE Trans. Med. Imag. 8, 1–7.

Konrad, C., Ukas, T., Nebel, C., Arolt, V., Toga, A., Narr, K., 2009. Defining the human hippocampus in cerebral magnetic resonance images - An overview of current segmentation protocols. NeuroImage 47, 1185–1195.

Leung, K., Barne, J., Ridgway, G., Bartlett, J., Clarkson, M., Macdonald, K., Schuff, N., Fox, N., Ourselin, S., Initiative, A. D. N., 2010. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. NeuroImage 51, 1345–1359.

Lotjonen, J., Wolz, R., Koikkalainen, J., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 49, 2352–2365.

Maroy, R., Boisgard, R., Comtat, C., Frouin, V., Cathier, P., Duchesnay, E., Dolle, F., Nielsen, P., Trebossen, R., Tavitian, B., 2008. Segmentation of rodent whole-body dynamic PET images: An unsupervised method based on voxel dynamics. IEEE 27, 342–354.

Morey, R., Petty, C., Xu, Y., Hayes, J., Wagner II, H., Lewis, D., LaBar, K., Styner, M., McCarthy, G., 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. NeuroImage 45, 855–866.

Morra, J., Zhuowen, T., Apostolova, L., Green, A., Avedissian, C., Madsen, S., Parikshak, N., Hua, X., Toga, A., Jack Jr., C., Weiner, M., Thompson, P., The Alzheimer's Disease Neuroimaging Initiative, 2008. Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. NeuroImage 43, 59–68.

Pardoe, H., Pell, G., Abbott, D., Jackson, G., 2009. Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? Epilepsia 50, 2586–2592.

Patenaude, B., 2007. Bayesian shape and appearance models. In: FMRIB Technical Report TR07BP1 Oxford.

Sethian, J., 1996. A fast marching level set method for monotonically advancing fronts. Proc. Natl. Acad. Sci. USA 93, 1591–1595.

Shattuck, D., Prasad, G., Mirza, M., Narr, K., Toga, A., 2009. Online resource for validation of brain segmentation methods. NeuroImage 45, 431–439.

Taylor, D., Barrett, W., 1994. Image segmentation using globally optimal growth in three dimensions with an adaptive feature set. Visualization in Biomedical Computing, 98–107.

van der Lijn, F., den Heijer, T., Breteler, M., Niessen, W., 2008. Hippocampus segmentation in MR images using atlas registration, voxel classification and graph cuts. NeuroImage 43, 708–720.

Wolz, R., Aljabar, P., Hajnal, J., Hammers, A., Rueckert, D., the Alzheimer's Disease Neuroimaging Initiative, 2010. LEAP: Learning embeddings for atlas propagation. NeuroImage 49, 1316–1325.

Woolrich, M., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., Smith, S., 2009. Bayesian analysis of neuroimaging data in FSL. NeuroImage 45, 173–186.

Yang, J., Duncan, J., 2004. 3D image segmentation of deformable objects with joint shape-intensity prior models using level sets. Med. Image Anal. 8, 285–294.

Yang, J., Staib, L., Duncan, J., 2004. Neighbor-constrained segmentation with 3D deformable models. IEEE Trans. Med. Imag. 23 (8), 940–948.

7. Figure captions

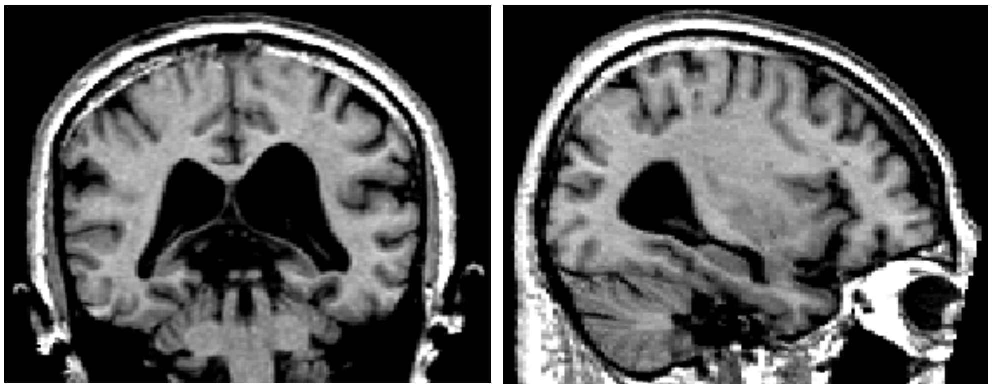

Figure 1: Clinical MR data. Coronal- (left) and sagittal- (right) view images of an Alzheimer's Disease (AD) patient from the CMA dataset, showing ventricular enlargement and hippocampal atrophy.

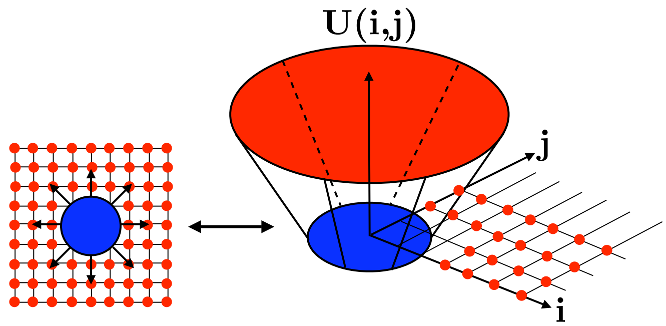

Figure 2: The Fast Marching (FM) arrival time surface. Schematic illustration of the minimal energy path active contour approach in 2D, with the initial curve shown in blue and pixels represented by red dots. On the left, we present a hypothetical isotropic propagation of the FM front across a homogeneous image field, with a later arrival time surface $U(i,j)$ shown on the right. The arrival time surface is so-called because it gives the arrival time of the propagating front at any given point in space. Conceptually similar to level-set methods, a slice through the arrival time surface at any given height (or time) gives the spatial extent of the propagating front at that time.

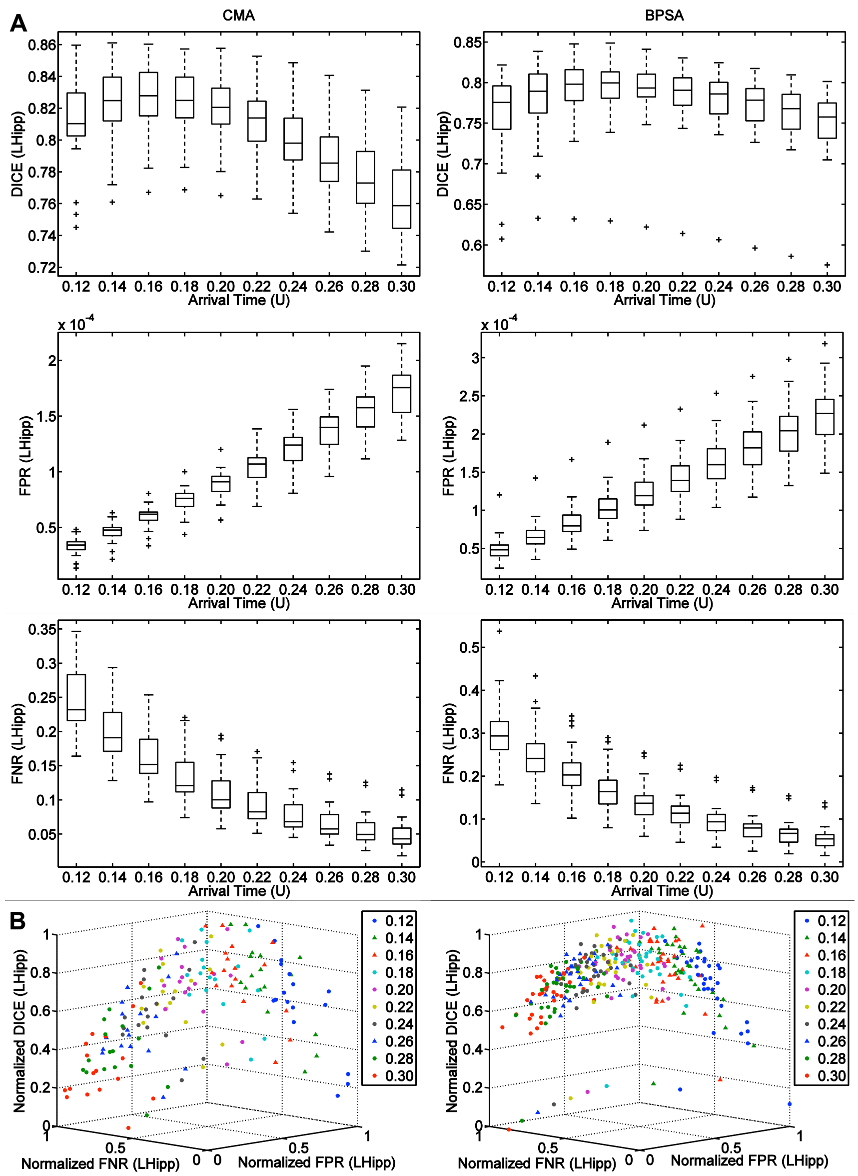

Figure 3: FMASH performance: left hippocampus (LHipp). (A) Plots of Dice coefficient (top row), FPR (second row) and FNR (third row) for the left hippocampus showing FMASH performance on the CMA data (left column) and BPSA data (right column) at varying arrival time thresholds. Boxes have lines at the lower quartile, median, and upper quartile values, with whiskers extending to 1.5 times the inter-quartile range. Outliers are indicated by a plus sign. (B) 3D scatter plot of the normalized performance measures, with the optimal arrival time clustering in the top, right-back corner of the plot.

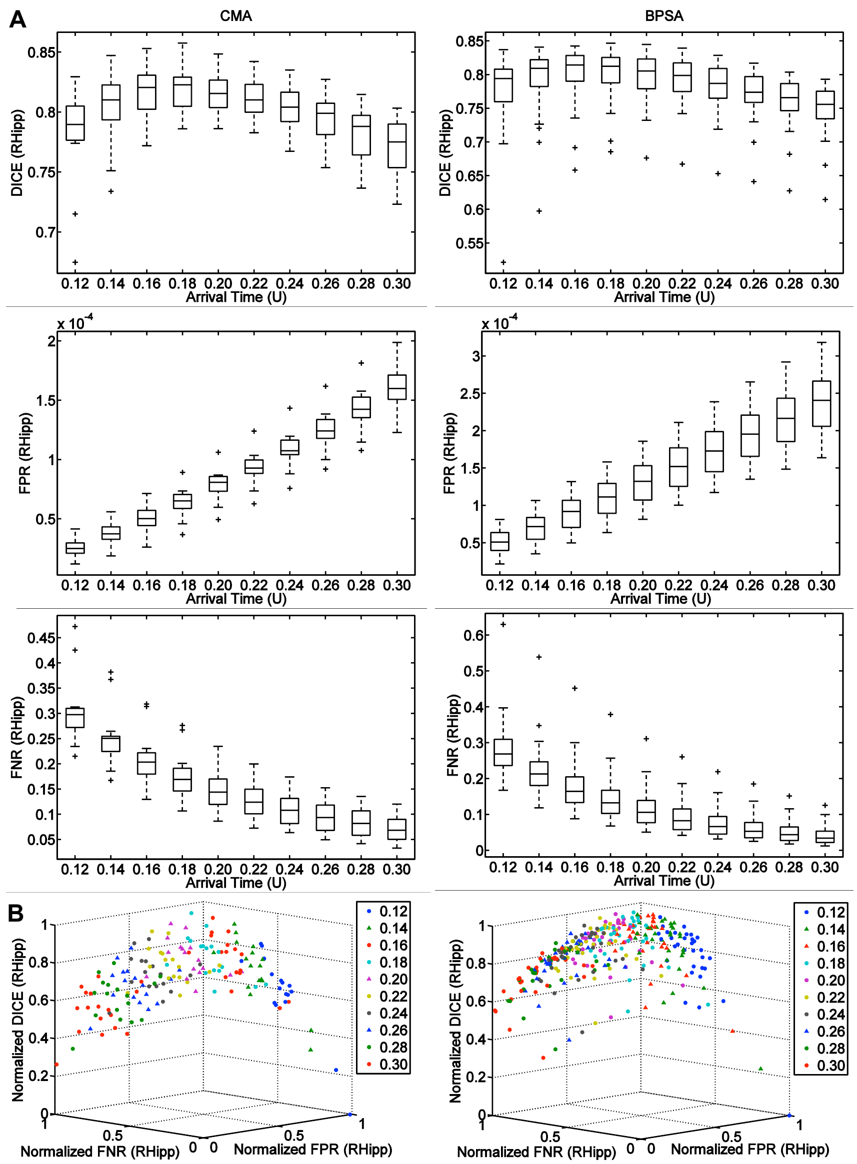

Figure 4: FMASH performance: right hippocampus (RHipp). Plots of right hippocampal FMASH performance on the CMA data (left column) and BPSA data (right column) at varying arrival time thresholds. Please refer to Figure 3 for more details.

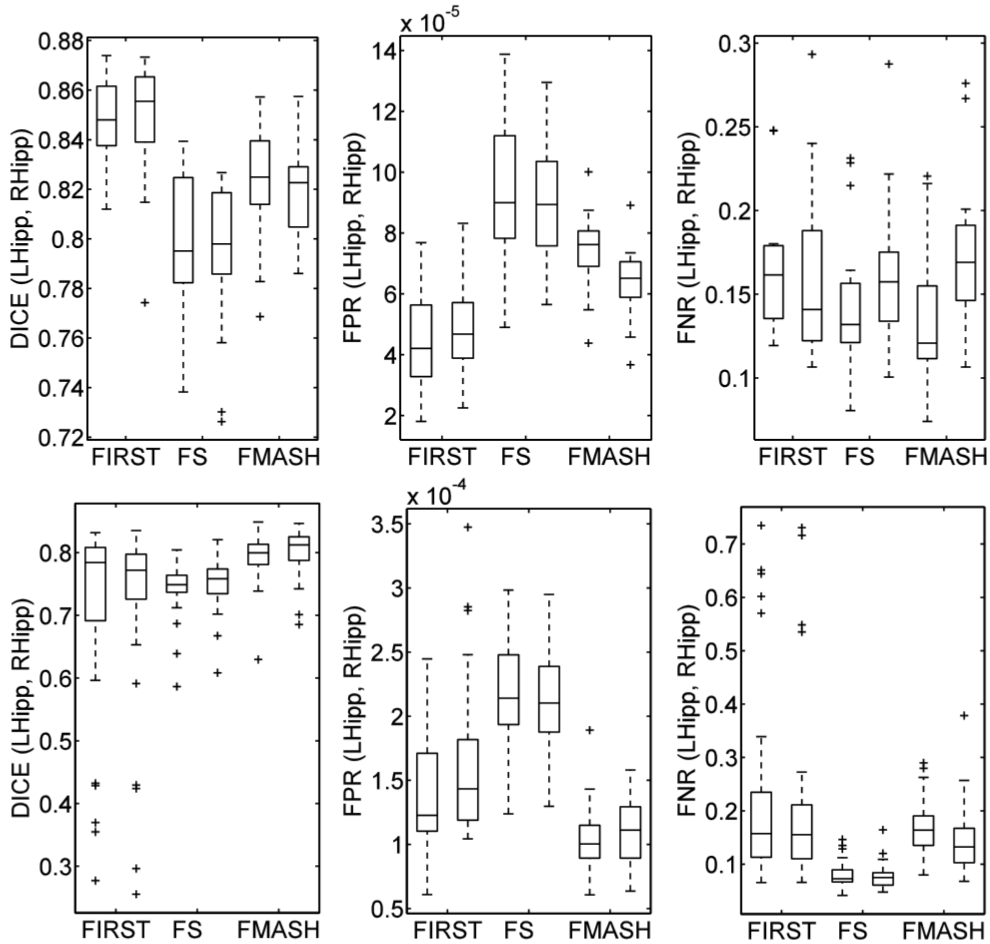

Figure 5: Method comparison. Box plots comparing FMASH segmentation performance with that of FIRST and FS on the CMA data (top row) and BPSA data (bottom row), showing Dice coefficients (left column), FPR (middle column) and FNR (right column) for both the left hippocampus (left) and right hippocampus (right). Although all FPR are extremely low (of the order $10^{-5} - 10^{-4}$), due to the high TN count in the calculation of FPR (Equation 7), statistically significant and important differences between the segmentation methods are observed.

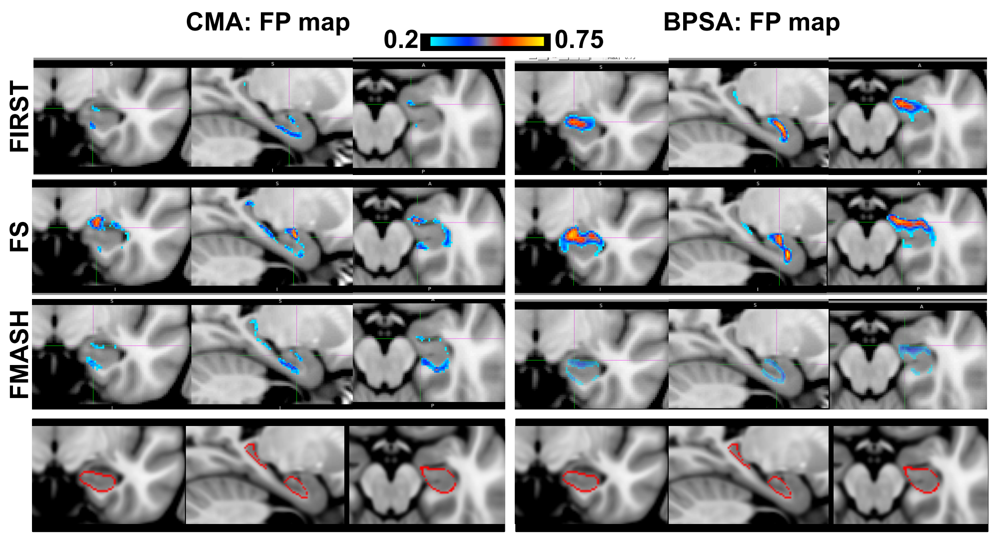

Figure 6: False positive (FP) maps. Each panel shows coronal- (left), sagittal- (middle) and axial- (right) view images of the left hippocampus (LHipp) standard-space FP maps for FIRST (top row), FS (second row) and FMASH (third row) on the CMA and BPSA data (left and right panel, respectively). A voxel-wise threshold of 20% FP finding across the dataset is applied to all FP maps. To aid visual comparison, the hippocampal boundary defined by FIRST on the MNI152 standard image is shown (fourth row) and the cursor position is the same for all images.

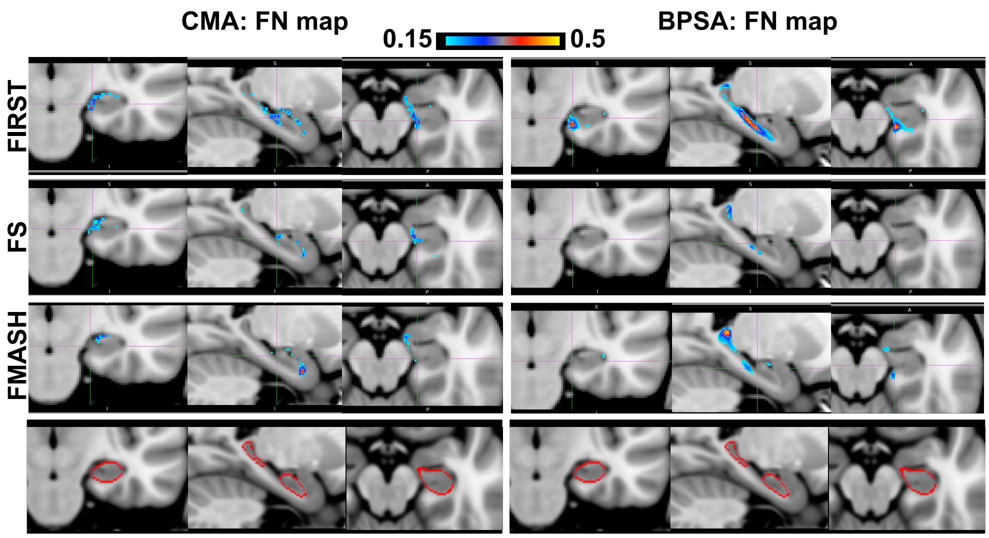

Figure 7: False negative (FN) maps. In a similar layout to Figure 6, each panel shows coronal- (left), sagittal- (middle) and axial- (right) view images of the left hippocampus (LHipp) standard-space FN maps for FIRST (top row), FS (second row) and FMASH (third row) on the CMA and BPSA data (left and right panel, respectively). A voxel-wise threshold of 15% FN finding across the dataset is applied to all FN maps. The hippocampal boundary defined by FIRST on the MNI152 standard image is shown (fourth row) and the cursor position is the same for all images.