

Spatial Warping Network for 3D Segmentation of the Hippocampus in MR Images

Nicola K Dinsdale¹, Mark Jenkinson¹, and Ana IL Namburete²

¹ Wellcome centre for Integrative Neuroimaging, FMRIB, University of Oxford

² Institute of Biomedical Engineering, University of Oxford

`nicola.dinsdale@dtc.ox.ac.uk`

Abstract. Accurate segmentations of neuroanatomical structures are essential for volumetric and morphological assessment but manual segmentation is time-consuming and error-prone. We propose a convolutional neural network for structural segmentation based on deformation of an example mask that is disease-state agnostic, which we apply to the hippocampus. The hippocampus is one of the first subcortical structures affected by Alzheimer's disease, atrophying as the disease progresses. As the disease state is not always known, and due to the varying degrees of atrophy, an accurate shape prior is not always available. The network is based on an adapted spatial transformer network that learns a deformation field to resample an initial binary mask, to create an output segmentation. This segmentation is learnt by the network from the input T1-weighted MRI in an end-to-end manner. Experiments on the HarP dataset show that the network outperforms other segmentation methods and is consistent across disease states, independent of the degree of disease-related atrophy. We also explore the effect of the initial binary mask on the segmentation and show that the segmentation is insensitive to the initialisation of this mask.

Keywords: Segmentation · Spatial Transformer · Hippocampus

1 Introduction

The hippocampus is a grey matter structure located within the medial temporal lobe memory circuit [1]. Hippocampal atrophy, observed through MRI, is one of the most validated biomarkers of Alzheimer's disease [2]. Manual segmentation of the hippocampus is time-consuming and error-prone, while existing automated methods provide relatively low-accuracy segmentations.

The task of hippocampal segmentation is commonly approached via model-based methods [3][4] in which an atlas is first non-rigidly registered with a target image and then the labels are propagated from an atlas to create the output segmentations. Segmentation quality is strongly influenced by the regularisation applied, and the accuracy of the registration. FreeSurfer [3] and FSL FIRST [4] are two commonly used model-based segmentation packages. FIRST uses active shape and appearance models but places them within a Bayesian framework;

FreeSurfer uses a Bayesian framework with deformable registration to determine the subcortical labels. These methods are constrained by the regularisation imposed upon the deformation to solve the ill-conditioned problem. This also limits the inter-subject registration, especially when pathological changes, such as hippocampal atrophy, have occurred.

Recently, segmentation has been approached using deep convolutional neural networks, specifically encoder-decoder networks. Several encoder-decoder networks have been developed specifically for subcortical brain segmentation: QuickNAT [6] uses a deep network for segmentation, aggregating three 2D networks formed of dense connections to produce a 3D segmentation, whereas DeepNAT [7] uses a hierarchical 3D network, first learning the foreground from the background segmentation and then segmenting the different subcortical regions. Methods such as VoxelMorph [8] use deep learning to approach segmentation via pairwise registration. This uses a deformable framework to learn a registration field and then applies this to the reference labels to produce the output segmentation; however, this method produces a bias in the template selection that is not ideal for segmentation of structures with varying disease states, and so varying levels of atrophy.

Spatial Transformer Networks [9] were originally proposed to counteract the inability of neural networks to be spatially invariant to their input; they form a block that allows deformations to be learnt and applied through a network that can be trained end-to-end (for instance in VoxelMorph [8]). We adapt the spatial transformer so that it takes an initial binary mask as an additional input and resamples it to match the target segmentation. We also remove the constraint for the learnt deformation to be an affine transformation so that the network is able to fully transform the initial binary mask to the required segmentation.

We propose the Spatial WArping Network for Segmentation (SWANS), a method that marries neural networks with their ability to learn rich feature representations and deformation methods to create a network for segmentation of the hippocampus using a deep neural network and an adapted spatial transformer. The network is capable of learning accurate segmentations given an abstract initial spatial mask in the form of a sphere, with no constraints placed on the deformation and no fixed template image. We show that spatial transformers can be used directly for segmentation and create a network that is end-to-end trainable. We also show that the network is robust to atrophy, successfully segmenting hippocampi from three different disease stages.

2 Method

The aim of the proposed SWANS network shown in Fig. 1 - is to model a spatial transformation $\mathbf{T}_{\Theta}(\mathbf{B}_0, \mathbf{B}, \mathbf{B}') = \Phi$ using a CNN, where Φ is a transformation field and Θ is the set of trainable parameters of the network. SWANS maps from an initial binary mask, \mathbf{B}_0 , to an output segmentation, \mathbf{B}' , by deforming \mathbf{B}_0 to match the manually labelled binary mask \mathbf{B} , such that $\Phi(\mathbf{B}_0) = \mathbf{B}'$.

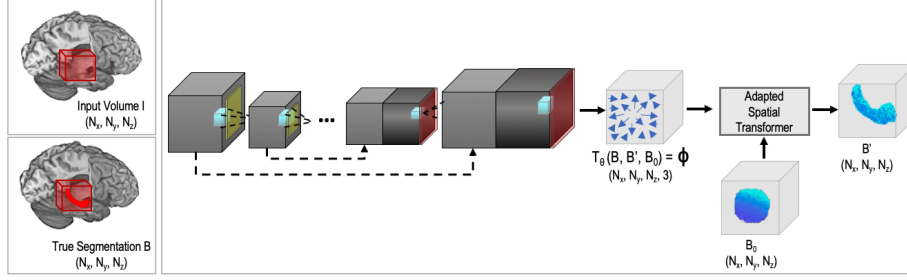


Fig. 1: Overall network architecture. \mathbf{I} is a 3D ROI input MRI volume, taken around the hippocampus. Φ is the transformation field learnt by the neural network. \mathbf{B}_0 is the initial binary mask, chosen arbitrarily to be a sphere. \mathbf{B} is the manually produced label and \mathbf{B}' is the segmentation produced by the network. The loss function aims to minimise the difference between the manual label \mathbf{B} and the output segmentation \mathbf{B}' .

The proposed network comprises of two main parts. The first is a neural network learning a deformation field. In this work, it was chosen to be a shallow (3 maxpooling layers) encoder-decoder network similar to the U-Net [10]. The second is the spatial transformer, which has been adapted for segmentation. The transformation field is applied to an initial binary mask, \mathbf{B}_0 , such that, $\Phi \in \mathbb{R}^{W \times H \times D \times 3}$ for the 3D case, such that $(\Phi)_{ijk}$ is the transformation learnt by the network at position (i, j, k) . The initial binary mask, \mathbf{B}_0 , is an arbitrary sphere which acts as an uninformative initialisation for the segmentation and is deformed by Φ to form the output segmentation.

This takes place in two stages: firstly the transformation $\Phi_{ijk} = (a, b, c)$ is applied to a coordinate (x, y, z) to map it to a new coordinate (x', y', z') where $(x', y', z') = (x, y, z) \circ (a, b, c) = (xa, yb, zc)$. This transformation is applied at each grid point G_{ijk} , to create a deformed grid, G'_{ijk} . The coordinates in the deformed grid are used to resample the initial binary mask, \mathbf{B}_0 , using trilinear interpolation.

The loss function is chosen to be a combination of a global loss term (Dice loss) and a local loss term (binary cross entropy) such that: $Loss(\mathbf{B}', \mathbf{B}) = L_{global}(\mathbf{B}', \mathbf{B}) + \alpha L_{local}(\mathbf{B}', \mathbf{B})$ where α determines the relative weighting of the local loss term.

Experimental Setup: The HarP dataset [11] was used for this study. It consists of 100 T1-weighted MRI scans for training and a further 31 for testing, labelled according to the HarP protocol in MNI space all with dimensions $(197 \times 233 \times 189)$ with voxel resolution of $1mm$. These consist of 45 Alzheimer’s disease subjects, 42 cognitively normal controls (CN), and 44 mildly cognitively impaired subjects (MCI). The dataset was designed to span the representative physiological and pathological variability (age, dementia severity), scanner field strength (1.5T and

3T) and scanner manufacturer (GE, Philips and Siemens) [11] and is preregistered to MNI space.

The MRI scans were then demeaned and divided by the 99th percentile value of the batch. 3D Regions Of Interest (ROIs) were extracted around the hippocampi with dimensions $(32 \times 64 \times 64)$ which is trivial due to the subjects being registered to the MNI template. The two hippocampi were taken separately, doubling the size of the dataset for testing and training. Small augmentations were applied to the dataset (namely, rotations, translations, flips and addition of Gaussian noise) increasing the size of the training set by a factor of 3.

Two main experiments were completed for this work: (1) comparing SWANS to existing methods; (2) exploring the effect of the initial binary mask on the output segmentations. The methods chosen to compare to were: FIRST³ [4], FreeSurfer⁴ [3], 3D U-Net [10] and QuickNAT [6]. QuickNAT and a full U-Net were trained on the augmented dataset until convergence. It was decided not to implement any pairwise registration-based methods because the requirement to choose a reference would create an implicit bias in the output segmentation, especially given the varying disease states of the subjects.

Implementation Details: The network was implemented in Python 3.6 using Keras (2.2.4) with a Tensorflow (1.11.0) backend. It was trained using the Adam optimiser with a learning rate of 1×10^{-4} on a P100 GPU taking 150 epochs at 83 seconds per epoch with a train/validation split of 90%/10%. A batch size of 2 volumes was used for training and α was set to 0.6 empirically.

3 Results

The segmentations produced by SWANS were thresholded at 0.5; however, it should be noted that the values all fell either below 0.01 or above 0.99 and so thresholding was required simply to binarise the output. Table 1 shows the 3D Dice results for the different methods tested, split according to disease state, and separately reporting Dice for left and right hippocampi. It can be seen that SWANS outperformed the other methods across the disease states, achieving an average Dice score of 0.865 ± 0.03 , and that the quality of the segmentation was fairly consistent across the disease states.

In Fig. 2, the comparison of the different methods for three metrics can be seen: 3D Dice score, Hausdorff Distance, and Volume Difference. These metrics were chosen because they demonstrate both the degree of overlap between the segmentation and the true label, and also the magnitude of any outliers. SWANS consistently outperformed the other methods for each metric, showing that although the network was trained to maximise Dice score, the segmentations produced have generalised well across metrics.

An example segmentation produced by SWANS can be seen in Fig. 3. The true label is shown by the light blue segmentation and the segmentation produced

³ Software download: fsl.fmrib.ox.ac.uk/fsldownloads_registration

⁴ Software download: surfer.nmr.mgh.harvard.edu/fswiki/DownloadAndInstall

Method	CN		MCI		AD	
	L	R	L	R	L	R
FIRST	0.782 \pm 0.03	0.779 \pm 0.03	0.747 \pm 0.03	0.740 \pm 0.04	0.778 \pm 0.04	0.780 \pm 0.04
Freesurfer	0.716 \pm 0.04	0.709 \pm 0.04	0.713 \pm 0.06	0.699 \pm 0.04	0.723 \pm 0.05	0.720 \pm 0.03
U-Net	0.787 \pm 0.03	0.799 \pm 0.04	0.775 \pm 0.08	0.825 \pm 0.04	0.778 \pm 0.06	0.826 \pm 0.04
QuickNAT	0.840 \pm 0.02	0.831 \pm 0.04	0.816 \pm 0.06	0.834 \pm 0.02	0.843 \pm 0.03	0.853 \pm 0.02
SWANS	0.842\pm0.02	0.870\pm0.02*	0.867\pm0.02*	0.874\pm0.02*	0.863\pm0.03*	0.886\pm0.01*

Table 1: 3D Dice Scores for each method on the testing data: mean \pm standard deviation - $n = 31$: $n_{CN}=13$, $n_{MCI}=12$, $n_{AD}=6$. * indicates significant improvement.

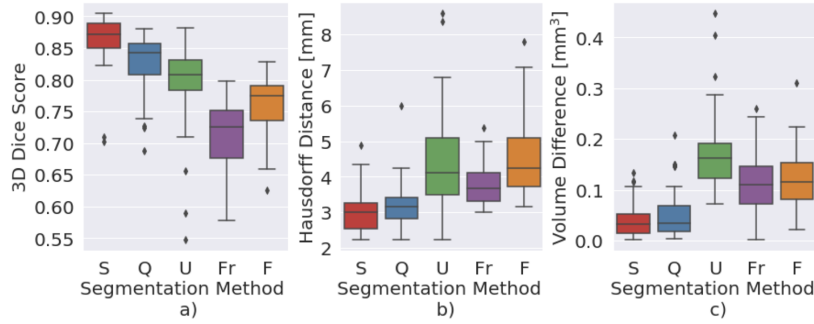


Fig. 2: Comparing different metrics on the testing dataset with each hippocampus (L/R) being treated as a separate datapoint a) 3D Dice scores, b) Hausdorff Distance c) Volume Difference. Segmentation Methods: S = SWANS, Q = QuickNAT, U = U-Net, Fr = FreeSurfer, F = FIRST.

by SWANS is shown by the overlaid red outline. The lowest Dice score output segmentation produced by SWANS can also be seen in Fig. 3: the segmentation is noisy and failed to segment some regions inside the hippocampus. It should be noted, however, that all three of the deep networks performed poorly on this subject: it represents the lowest 3D Dice score for all three CNN methods (SWANS: 0.70, QuickNAT: 0.69, U-Net: 0.55).

Finally, we investigated the false positives and false negatives to understand if our network was systematically failing in specific areas of the segmentation. Figure 3 shows heat maps of regional false negatives and false positives for SWANS, QuickNAT and FreeSurfer. Overall SWANS had the lowest number of both false negatives and false positives and also has similar numbers of both false positives and false negatives, showing that there is no systematic over- or under-segmentation. QuickNAT also had low false negative and false positive rates but also incorrectly identified a small region near the hippocampus as part of the hippocampus. FreeSurfer had the highest levels of both false positives and false negatives, consistent with the lower Dice values for the FreeSurfer segmentations.

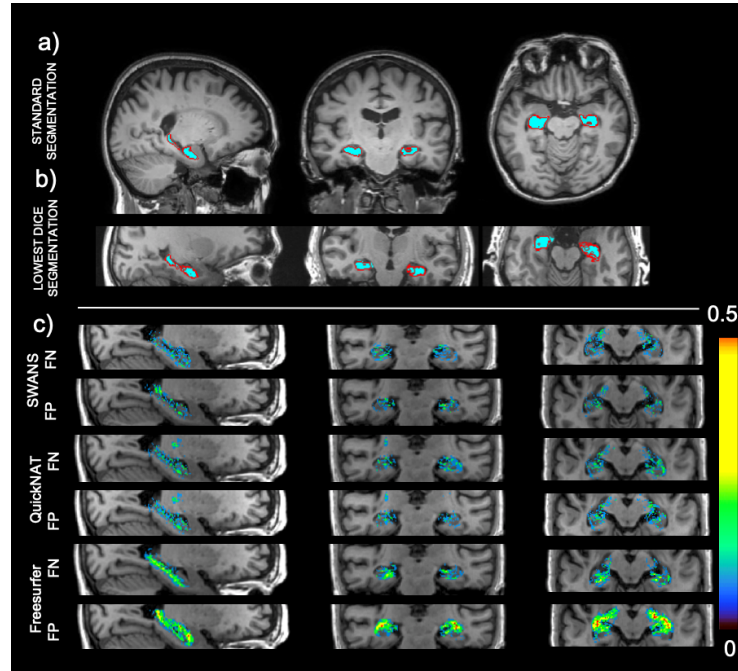


Fig. 3: a) Standard SWANS segmentation. Blue is the manual label and red is the outline of the segmentation produced by SWANS. b) The lowest Dice output segmentation produced by SWANS. Similarly blue is the manual label and red is the outline of the segmentation produced by SWANS. c) False negative (FN) and false positive (FP) heat maps for SWANS, QuickNAT and FreeSurfer, projected on a template image. Scale = proportion of times a voxel in that location was mislabelled.

Addition of Auxiliary Task: Motivated by the dataset containing subjects from three different disease states, an auxiliary task was added to predict this state. It was hypothesised that the addition of the auxiliary task would bias the network towards the different representations and thus make the network better able to segment the hippocampus accordingly for the different disease states. Therefore, a simple, fully-connected classifier was connected to the smallest representation in the encoder. Two different experiments were run in addition to the standard network: a two-class and a three-class classifier. The classes in the former were healthy (CN) or pathological (MCI and AD) and, in the latter, the disease state (CN, MCI or AD).

The Dice score with the three-class auxiliary task was 0.868 ± 0.04 , and with the two-classes 0.866 ± 0.04 , compared to 0.866 ± 0.04 without any auxiliary task. Therefore, the addition of the auxiliary tasks did not significantly improve the segmentation (2 classes: $t=0.535$, $p=0.957$; 3 classes: $t=0.709$, $p=0.481$), indicating that the network alone is sufficiently powerful to represent the variation

in the dataset.

Effect of the Initial Binary Mask: The initial binary mask was chosen arbitrarily and so it is important to understand its effect on the final segmentation. Figure 4a shows the effect of varying the radius of the sphere used as the initial binary mask. Once the radius is 4 voxels, the segmentation is robust to the size of the initial binary mask. It was found that the network is unable to warp initial binary masks with a radius smaller than 4 to represent the hippocampus.

Sensitivity to the initial mask size was tested for two initial positions of binary mask: the centre of the ROI and a corner. Figure 4a also shows that the segmentation was robust to the initial location of initial binary mask: the same pattern in size variation applies and the same 3D Dice score was achieved in both locations.

Finally, Fig. 4b visualises the deformation, with each channel shown separately. Points maintain their colour as they are deformed, indicating which regions of the binary mask have been resampled to produce the segmentations. We can clearly see that the initial binary mask is deformed in a methodical manner and also that there is considerable redundancy, as confirmed by the fact that the initial binary mask can produce the segmentation down to a radius of 4 voxels.

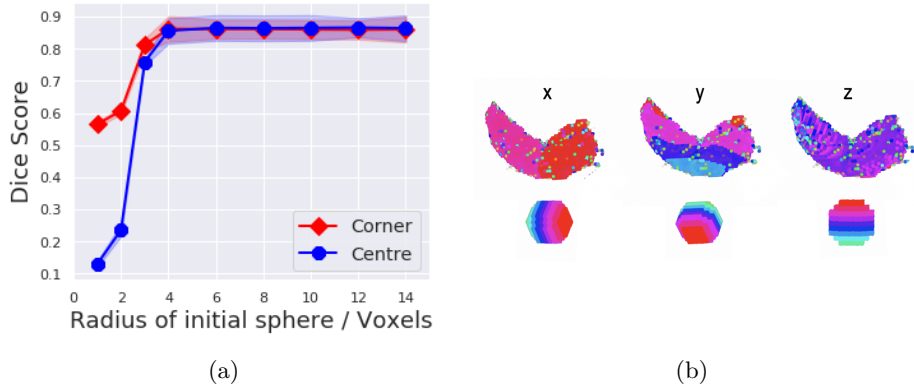


Fig. 4: a) The effect of the size of the initial binary mask on the average Dice score of the testing data. It can be seen that once the radius is four voxels or greater, the Dice score is stable, regardless of starting position or size. b) Visualisation of the deformation to the initial binary mask (radius of 6 voxels) to produce the segmentation. Voxels in the segmentation are sampled from the corresponding colour in the original binary mask. The specular values on the surface are caused by the interpolation used in the resampling.

4 Discussion

In this work we proposed the use of the Spatial Warping Network for Segmentation (SWANS) of the hippocampus. We have shown that SWANS performs well across several metrics compared to other methods and that the network was not affected by the initialisation of the mask. Critically for the network’s application to hippocampal segmentation, the learnt deformations successfully represent the pathological variation seen across healthy and Alzheimer’s disease patients, performing consistently across disease groups.

5 Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) [grant number EP/L016052/1]. A. Namburete is grateful for support from the UK Royal Academy of Engineering under the Engineering for Development Research Fellowships scheme. M. Jenkinson is supported by the National Institute for Health Research (NIHR) and the Oxford Biomedical Research Centre (BRC). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Mufson EJ, Mahady L, Waters D, et al.: Hippocampal plasticity during the progression of Alzheimer’s Disease. *Neuroscience* **309** 51 – 67 (2015)
2. Flores R, Joie R, Chetelat G: Structural imaging of the hippocampal subfields in healthy aging and Alzheimer’s disease. *Neuroscience* **309** 29 – 50 (2015)
3. Fischl B, Salat DH, et al.: Whole brain segmentation: automated labelling of neuroanatomical structures in the human brain. *Neuron* **33** 341-355 (2002)
4. Patenaude B, Smith SM, et al.: A Bayesian Model of Shape and Appearance for Subcortical Brain Segmentation. *Neuroimage* **56**(3) 907-922 (2011)
5. Ronneberger O, Fischer P, Brox T: U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234241 (2015)
6. Roy A, Conjeti S, Navab N, Wachinger C: QuickNAT: Segmenting MRI Neuroanatomy in 20 seconds. *NeuroImage* **186** 713-727 (2019)
7. Wachinger C, Reuter M, Klein T: DeepNAT: Deep Convolutional Neural Network for Segmenting Neuroanatomy. *NeuroImage* **170** (2017)
8. Balakrishnan G, Zhao A, Sabuncu M, Guttag J, Dalca A: VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *arXiv* (2018)
9. Jaderberg M, Simonyan K, Zisserman A, Kuvukcuoglu K: Spatial Transformer Networks. *Neural Information Processing Systems* (2015)

10. Çiçek O, Abdulkadir A, Lienkamp S, Brox T, Ronnenberger O: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention (2016)
11. Boccardi M, Bocchetta M, Morency FC et al.: Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimer's Dementia* **11**(2) 175-83 (2015)