Hippocampus Subfield Segmentation Using a Patch-Based Boosted Ensemble of Autocontext Neural Networks

José Manjón, Pierrick Coupé

Hippocampus subfield segmentation using a patch-based boosted ensemble of autocontext neural networks

José V. Manjón[1] and Pierrick Coupe[2,3]

[1]Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, España.
[2] Univ. Bordeaux, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.
[3] CNRS, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.

Abstract. The hippocampus is a brain structure that is involved in several cognitive functions such as memory and learning. It is a structure of great interest in the study of the healthy and diseased brain due to its relationship to several neurodegenerative pathologies. In this work, we propose a novel patch-based method that uses an ensemble of boosted neural networks to perform the hippocampus subfield segmentation on multimodal MRI. This new method minimizes both random and systematic errors using an overcomplete autocontext patch-based labeling using a novel boosting strategy. The proposed method works well on high resolution MRI but also on standard resolution images after superresolution. Finally, the proposed method was compared with a similar state-of-the-art methods showing better results in terms of both accuracy and efficiency.

1 Introduction

The hippocampus (HC) is a complex gray matter structure located under the surface of each temporal lobe. It is involved in many cognitive functions such as memory and spatial reasoning [1]. It has been largely studied in the last years to understand its healthy evolution across the lifespan in normal aging [2] but also due to its key role in several dysfunctions such as epilepsy [3], schizophrenia [4] or Alzheimer's disease [5].

The hippocampus is composed of multiple subfields that can be divided into sections called the dentate gyrus, the cornu ammonis (CA) and the subiculum. The CA is also subdivided in sub-sections CA1, CA2, CA3, CA4, layers alveus, stratum oriens, stratum pyramidale, stratum radiatum, stratum lancosum and stratum moleculare. These layers present a high neuron density and are very compact so high resolution imaging is required to identify them.

Due to this morphological complexity and MR related image resolution limitations, mainly whole hippocampus volume analysis has been performed in the past by segmenting it as a single object [6]. Even with this limitations whole HC volume has been shown to be a good biomarker for Alzheimer's disease [7]. However, hippocampus subfields have shown to be affected differently by AD and normal aging in ex-vivo studies [5] which makes them excellent candidates for early diagnosis.

Although high resolution MRI is becoming more accessible in research scenarios, manual segmentation, which is the most accurate analysis method, is not a feasible option since it is a highly time consuming procedure which requires expert trained raters taking many hours per case.

To overcome this problem some automated segmentation solutions have been developed in the last years. One of the first methods was proposed by Van Leemput et al. using a statistical model of MR image formation around the hippocampus to produce automatic segmentation [7]. Recently, Iglesias et al. pursued this work and replaced the model by a more accurate atlas generated using ultra-high resolution ex-vivo MR images [8]. Chakravarty et al. proposed a multiatlas method based on the estimation of several non-linear deformations and a label fusion step [9]. Also using a multiatlas approach, Yushkevich proposed a method where a multiatlas approach is combined with a similarity-weighted voting and a learning-based label bias correction [10] and Romero et al also proposed a multiatlas multispectral method[21].

In this work, we propose a fast and accurate patch-based method to segment the hippocampus subfields using an ensemble of boosted neural networks. In the next sections, we will describe the method details as well as some experiments to demonstrate the accuracy and efficiency of the proposed approach.

2 Material and Methods

2.1 Image data

In this paper, we used a High Resolution (HR) dataset composed of 25 cases with T1-weighted and T2-weighted images to construct a library of manually labeled cases. This dataset includes 25 subjects from a public repository (http://www.nitrc.org/projects/mni-hisub25) (31 ± 7 yrs, 12 males, 13 females) with manually-drawn labels dividing the HC in three parts (CA1-3, DG-CA4 and Subiculum). MRI data from each subject consist of an isotropic 3D-MPRAGE T1-weighted (0.6x0.6x0.6 mm^3) and anisotropic 2D T2-weighted TSE images (0.4×0.4×2 mm^3). Images underwent automated correction for intensity non-uniformity, intensity standardization and were linearly registered to the MNI152 space. T1w and T2w images were resampled to a resolution of 0.4 mm^3 (Figure 1). To reduce interpolation artifacts, the T2w data was upsampled using a non-local superresolution method [19]. For more details about the labeling protocol see the original paper [11].

2.2 Preprocessing

All the images (T1 and T2) were first filtered with a spatially adaptive non-local means filter [15] and inhomogeneity corrected using the N4 method [16]. Later, they were linearly registered to the Montreal Neurological Institute space (MNI) using the ANTS package [17] and the MNI152 template. Next, we left-right flipped the images and cropped them to the right hippocampus area to produce 50 right hippocampus crops. After that, we non-linearly registered the cropped images to the cropped MNI152

template to better match the hippocampus anatomy. Finally, we normalized the image intensities using Nyúl and Udupa [18] method. Hippocampus labels were spatially registered to the same space.

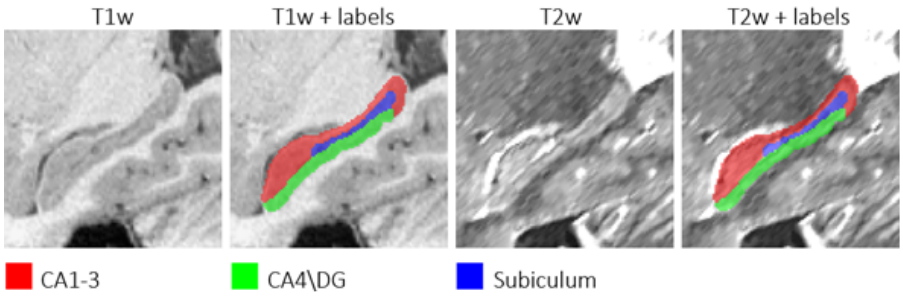

Figure 1: Example of an HR MRI case. Figure shows T1w and T2w images and its corresponding manual segmentation.

2.3 Proposed method

After the described preprocessing, a region of interest (ROI) is computed fusing all manual segmentations of the library and dilating the resulting region with a 5x5x5 voxels kernel to create a HC candidate region. For each voxel of this candidate region a feature set is created to be used to train a classifier. Several classifiers can be used to relate the image features and the corresponding labels. Lately high performance classifiers such as random forest [12] have been used. In our proposed method, we have used a neural network-based classifier [13].

- ***Features*:** The features used to train the network were three 3D patches per image modality of different size around the voxel/s to be classified, the x, y and z voxel coordinates of the center voxel of the patches and a value representing the *a priori* label probability. This apriori label probability map was obtained computing the average of all training label masks (convolved with a 5 mm^3 Gaussian kernel). In our experiments, we used a P_1 of size 3x3x3, a P_2 of 7x7x7 and P_3 of 9x9x9 voxels (however, for efficiency, we subsampled the patches P_2 and P_3 so we took only 27 samples uniformly spaced in all three dimensions). This leads to a feature vector X of 166 elements (i.e. 27 for P_1, 27 for P_2 and 27 for P_3 on T1, the same for T2, x, y and z coordinates and the prior probability).

- ***Network topology*:** A feedforward multilayer perceptron with two hidden layers was used. The network that we used had 166x85x55x27 weights. The network output is a patch of the same size of P_1. Note that an overcomplete approach was used so each voxel has contributions from several adjacent patches. This improves segmentation accuracy (more votes per voxel) and enforces the final label regularity. To further improve classification results a second autocontext network is trained using an expanded feature vector constructed concatenating the original feature

vector X with the output of the first network. This leads to a feature vector X_a of 193 elements (166 + 27). Final classification is obtained from the output of network 2. Note that both networks are independently trained (Figure 2) .

Ensemble-based classification

A common approach for improving classification results is the use of the so-called ensemble learning. Ensemble methods (i.e. combination of several classifiers) allow in general to improve classification results by minimizing random and systematic errors.

In our proposed method, we have used a boosting strategy to leverage classification accuracy. Boosting [14] is an algorithm that combines the output of several classifiers to minimize the variance and bias of the final classification. In boosting, each classifier is trained using the information of the previous one to minimize the errors of the current prediction. This is done giving more weight to the samples wrongly classified by the previous classifier or performing a non-random selection on the training dataset selecting with higher probability samples wrongly classified previously. While typically each network uses random initial weights (network reset) we decided to use the weights of the previous network as done in transfer learning which improves ensemble classifier accuracy while minimizing training time due to faster convergence. Finally, the different classifier outputs are combined according to their accuracy.

We trained four ensembles of M autocontext modules (figure 2) (one ensemble per subfield plus the background) over the whole hippocampus region and each voxel was labeled with the class of higher network output.

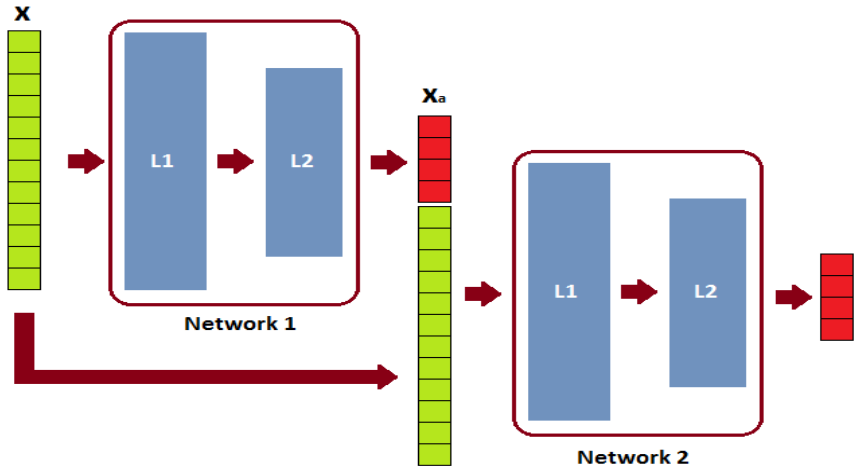

Figure 2: Autocontext neural network. Original feature vector x used to train network 1 is expanded using the output of the network (posterior probabilities) to train network 2.

3 Experiments and results

In this section, a set of experiments are presented to show the performance of the proposed method. All the experiments have been done by processing the cases from the described library splitting the 50 cases first into a 30 training cases set and 20 test cases and later switching training and test datasets to evaluate the whole dataset.

3.1 Ensemble training

We explored two variants of ensemble training, the classical one with network reset and one without reset. For these experiments, we trained M=10 autocontext networks using only 10000 samples randomly selected from the candidate regions of the training dataset. All resulting networks outputs were averaged according to the accuracy to produce the final output.

We evaluated the impact of the two boosting variants (with and without reset) and estimated the optimal number of neural networks. In figure 3 (left), the evolution of the DICE coefficient (during training without reset) as a function of the number of individual and averaged trained networks is shown. In figure 3 (right), the same results with reset option are also shown.

As can be noticed, both boosting variants improved the classification results reaching a plateau at around 10 networks. However, no-reset boosting produced a more pronounced improvement compared to classical reset approach (0.9091 versus 0.9052). To understand the improved results we can look at the accuracy of each individual network of the ensemble. As can be noticed, reseted networks show a pseudo stable behaviour while non-reseted networks show maintained improvements as long as the number of ensemble networks increases. In fact, non-reseted last individual networks almost reach the accuracy of the whole ensemble.

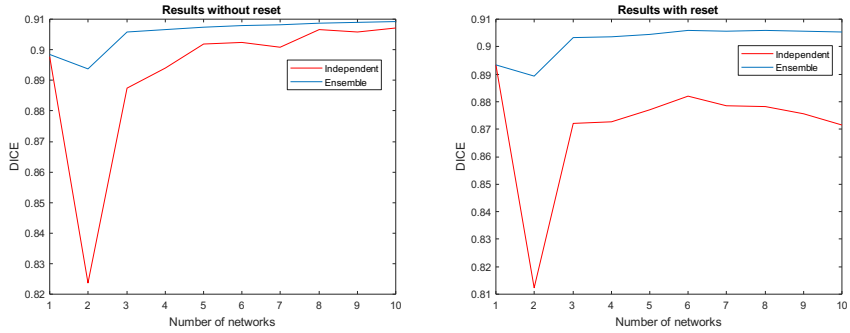

Figure 3. Left: Dice coefficient in function of the number of networks for each network in nthe ensemble and for the ensemble prediction with the proposed boosting. Right: Same results with classical boosting. Note that using previous network in the embeding not only improves overall ensemble accuarcy but also produces more accurate individual networks.

With this settings, we trained the final network ensemble (M=10) using randomly selected sets of 1000000 samples from the total population of 2600000 patches. To train the 4 ensembles took around 8 hours while the time to segment a new case is around 10 seconds. To evaluate all 50 cases we trained two ensemble sets (one using the 30 training cases and applied to the remaining 20 cases and another trained on the 20 cases and applied to the 30 cases). We could do a leave-one-out approach to further improving the results but this would result on a large training time of around two weeks. Table 1 shows the dice coefficient of the different subfields for the 50 cases. We have also included the results when using only the best network instead of the ensemble (thus requiring only 1 second to perform the segmentation).

Table 1: Mean DICE and standard deviation for each structure segmentation using two variants of the proposed method. Best results in bold.

Structure	***Proposed (best network)***	***Proposed (ensemble)***
Average	0.8681	**0.8695**
CA1-3	0.8992	**0.9001**
CA4\DG	0.8384	**0.8404**
Subiculum	0.8667	**0.8678**
Hippocampus	0.9518	**0.9523**

3.2 Standard resolution vs High resolution

High resolution MR images are not widely available, especially in clinical environments. For this reason, we analyzed the effectiveness of the proposed method on upsampled standard resolution images. For this purpose, we reduced the resolution of the library HR images by a factor 2 by convolving the HR images with a 2x2x2 boxcar kernel and then decimating the resulting image. The down-sampled images were upsampled by a factor 2 using BSpline interpolation and a superresolution method called Local Adaptive SR (LASR) [19]. Results are shown in Table 2. As can be noticed, segmentations performed on images upsampled with SR were better than using BSpline interpolation. Moreover, this experiment shows that the proposed method is able to produce competitive results even when using standard resolution images.

Table 2: Mean DICE for each structure segmentation using the high resolution library and applying BSpline interpolation and LASR to the previously downsampled image to be segmented. Best results in bold.

Structure	***BSpline***	***LASR***	***HR***
Average	0.8595	0.8662	**0.8695**
CA1-3	0.8930	0.8981	**0.9001**
CA4\DG	0.8250	0.8349	**0.8404**
Subiculum	0.8605	0.8655	**0.8678**
Hippocampus	0.9480	0.9513	**0.9523**

3.3 Comparison

We compared our method with other recent methods applied to hippocampus segmentation using the same number of structures and labeling protocol. The compared methods are called ASHS[10] and Surfpatch [20]. Table 3 shows that the proposed method obtained higher DICE coefficients for all the structures. In terms of computation efficiency, our method requires only few seconds while ASHS and Surfpatch have an execution time of several hours per case.

Table 3: Mean DICE in the native space for each structure. Segmentation performed by ASHS, SurfPatch, proposed method and human rater (intra-rater and inter-rater). Best results (for automatic segmentation) in bold.

Structure	***ASHS***	***SurfPatch***	***Proposed***	***Inte-rater***	***Intra-rater***
Average	0.8513	0.8503	**0.8584**	0.8833	0.9113
CA1-3	0.8736	0.8743	**0.8903**	0.8760	0.9290
CA4\DG	0.8254	0.8271	**0.8283**	0.9030	0.9000
Subicu-lum	0.8548	0.8495	**0.8565**	0.8710	0.9050

4 Discussion

In this paper, we present a new hippocampus subfield segmentation method based on a boosted ensemble of autocontext neural networks. The proposed method achieves state-of-the-art accuracy very efficiently. Furthermore, the proposed method has been shown to perform well on standard resolution images, obtaining competitive results on typical clinical data. This fact is very important because it will allow analyzing large amounts of legacy data. Finally, it has been also shown that the proposed method compares well to another related state-of-art method obtaining better results in terms of both accuracy and reduced execution time.

5 Acknowledgements

This research was supported by the Spanish UPV2016-0099 grant from Universitat Politécnica de Valencia. This study has been also carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX-03-02, HL-MRI Project) and Cluster of excellence CPU and TRAIL (HR-DTI ANR-10-LABX-57).

6 References

1. Milner, B, Psychological defects produced by temporal lobe excision Res. Publ. Assoc. Res. Nerv. Ment. Dis., 36, pp. 244–257, (1958).
2. Petersen, R et al. Memory and MRI-based hippocampal volumes in aging and AD Neurology, 54 (3), pp. 581–587, (2000).
3. Cendes, F et al., MRI volumetric measurement of amygdala and hippocampus in temporal lobe epilepsy Neurology, 43 (4), pp. 719–725, (1993).
4. Altshuler, LL et al., Amygdala enlargement *in bipo*lar disorder and hippocampal reduction in schizophrenia: an MRI study demonstrating neuroanatomic specificity Arch. Gen. Psychiatry, 55 (7), p. 663, (1998).
5. Braak H and Braak E, Neuropathological stageing of Alzheimer-related changes, A*cta Neu*ropathol., 82 (4), pp. 239–259, (1991)
6. Chupin, M et al., Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI Hippocampus, 19 (6), pp. 579–587, (2009).
7. Van Leemput, K el al., Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI Hippocampus, 19 (6), pp. 549–557, (2009).
8. Iglesias JE et al., A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI, NeuroImage, 115 (15), pp. 117–137, (2015).
9. Chakravart*y,M.* et al*.,* Performing label-fusion-based segmentation using multiple automatically generated templates, Human brain mapping,10(34), pp. 2635 - 2654, (2013).
10. Yushkevich, PA et al., Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment Hum. Brain Mapp., 36 (1), pp. 258–287, (2015).
11. Kulaga-Yoskovitz, J., Bernhardt, B.C., Hong, S., Mansi, T., Liang, K.E., van der Kouwe, A.J.W., Smallwood, J., Bernasconi, A., Bernasconi, N., 2015 Multi-contrast submillimetric 3Tesla hippocampal subfield segmentation protocol and dataset. Sci Data. 2, 150059.
12. Serag et al. SEGMA: An Automatic SEGMentation Approach for Human Brain MRI Using Sliding Window and Random Forests. Front Neuroinform. 2017; 11: 2.
13. Manjón JV et al.. HIST: HyperIntensity Segmentation Tool. PatchMI workshop, MICCAI2016, Athens, 2016.
14. Schapire R. E. (1990). The strength of weak learnability. Machine Learning, 5(2):197:227.
15. Manjón, J.V. et al. Adaptive non-local means denoising of MR images with spatially varying noise levels. J Magn Re-son Imaging. 31, 192–203 (2010).
16. Tustison, N.J.et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29(6): 1310 - 1320. (2010)
17. Avants, BB et al., Advanced normalization tools (ANTS), Insight Journal, (2009).
18. Nyúl, L.G., Udupa, J.K., 1999. On standardizing the MR image intensity scale. Magn Reson Med. 42(6), 1072 - 81.
19. Coupé P et al., Collaborative patch-based super-resolution for diffusion-weighted images, NeuroImage, 83, pp. 245-261, (2013).
20. Caldairou, B., Bernhardt, B.C., Kulaga-Yoskovitz, J., Kim, H., Bernasconi, N., Bernasconi, A., 2016. A Surface Patch-Based Segmentation Method for Hippocampal Subfields. MICCAI, Part II, LNCS 9901, 379 - 387.
    Romero JE, Coupe P, Manjón JV. High Resolution Hippocampus Subfield Segmentation Using Multispectral Multiatlas Patch-Based Label Fusion. International Workshop on Patch-based Techniques in Medical Imaging, 117-124. 2016.