Medical Image Analysis 48 (2018) 177–186

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features

Kaisar Kushibar[1,*], Sergi Valverde[1], Sandra González-Villà, Jose Bernal, Mariano Cabezas, Arnau Oliver, Xavier Lladó

Institute of Computer Vision and Robotics, University of Girona, Ed. P-IV, Campus Montilivi, Girona, 17003, Spain

ARTICLE INFO

Article history:
Received 26 September 2017
Revised 1 March 2018
Accepted 9 June 2018
Available online 15 June 2018

Keywords:
Brain
MRI
Sub-cortical structures
Segmentation
Convolutional neural networks

ABSTRACT

Sub-cortical brain structure segmentation in Magnetic Resonance Images (MRI) has attracted the interest of the research community for a long time as morphological changes in these structures are related to different neurodegenerative disorders. However, manual segmentation of these structures can be tedious and prone to variability, highlighting the need for robust automated segmentation methods. In this paper, we present a novel convolutional neural network based approach for accurate segmentation of the sub-cortical brain structures that combines both convolutional and prior spatial features for improving the segmentation accuracy. In order to increase the accuracy of the automated segmentation, we propose to train the network using a restricted sample selection to force the network to learn the most difficult parts of the structures. We evaluate the accuracy of the proposed method on the public MICCAI 2012 challenge and IBSR 18 datasets, comparing it with different traditional and deep learning state-of-the-art methods. On the MICCAI 2012 dataset, our method shows an excellent performance comparable to the best participant strategy on the challenge, while performing significantly better than state-of-the-art techniques such as FreeSurfer and FIRST. On the IBSR 18 dataset, our method also exhibits a significant increase in the performance with respect to not only FreeSurfer and FIRST, but also comparable or better results than other recent deep learning approaches. Moreover, our experiments show that both the addition of the spatial priors and the restricted sampling strategy have a significant effect on the accuracy of the proposed method. In order to encourage the reproducibility and the use of the proposed method, a public version of our approach is available to download for the neuroimaging community.

1. Introduction

Brain structure segmentation in Magnetic Resonance Images (MRI) is one of the major interests in medical practice due to its various applications, including pre-operative evaluation and surgical planning, radiotherapy treatment planning, longitudinal monitoring for disease progression or remission (Kikinis et al., 1996; Phillips et al., 2015; Pitiot et al., 2004), etc. The sub-cortical structures (i.e. thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens) have attracted the interest of the research community for a long time, since their morphological changes are frequently associated with psychiatric and neurodegenerative disorders and could be used as biomarkers of some diseases (Debernard et al., 2015; Mak et al., 2014). Therefore, segmentation of sub-cortical brain structures in MRI for quantitative analysis has a major clinical application. However, manual segmentation of MRI is extremely time consuming and hardly reproducible due to inter- and intra- variability among operators, highlighting the need for automated accurate segmentation methods.

Recently, González-Villà et al. (2016), reviewed different approaches for brain structure segmentation in MRI. One of the commonly used automatic brain structure segmentation tools in medical practice is FreeSurfer,[2] which uses non-linear registration and an atlas-based segmentation approach (Fischl et al., 2002). Another classical approach, also popular in the medical community,

* Corresponding author.
E-mail addresses: kaisar.kushibar@udg.edu (K. Kushibar), sergio.valverde@udg.edu (S. Valverde), sgonzalez@eia.udg.edu (S. González-Villà), jose.bernal@udg.edu (J. Bernal), mariano.cabezas@udg.edu (M. Cabezas), aoliver@eia.udg.edu (A. Oliver), xavier.llado@udg.edu (X. Lladó).

[1] These authors contributed equally to this work.

[2] https://surfer.nmr.mgh.harvard.edu/.

https://doi.org/10.1016/j.media.2018.06.006

is the method proposed by Patenaude et al. (2011) – FIRST, which is included into the publicly available software FSL.[3] This method uses the principles of Active Shape (Cootes et al., 1995) and Active Appearance Models (Cootes et al., 2001) that are put within a Bayesian framework, allowing to use the probabilistic relationship between shape and intensity to its full extent.

In recent years, deep learning methods, in particular, Convolutional Neural Networks (CNN), have demonstrated a state-of-the-art performance in many computer vision tasks such as visual object detection, classification and segmentation (Krizhevsky et al., 2012; He et al., 2016; Szegedy et al., 2015; Girshick et al., 2014). Unlike handcrafted features, CNN methods learn from observed data (LeCun et al., 1998) making relevant features to a specific task. Naturally, CNNs are also becoming a popular technique applied in medical image analysis. There have been many advances in the application of deep learning in medical imaging such as expert-level performance in skin cancer classification (Esteva et al., 2017), high rate detecting cancer metastases (Liu et al., 2017), Alzheimer's disease classification (Sarraf and Tofighi, 2016), and spotting early signs of autism (Hazlett et al., 2017).

Some CNN methods have also been proposed for brain structure segmentation. One of the common ways used in the literature is patch-based segmentation, where patches of a certain size are extracted around each voxel and classified using a CNN. Application of 2D, 3D, 2.5D patches (patches from the three orthogonal views of an MRI volume) and their combinations including multi-scale patches can be found in the literature for brain structure segmentation (Brébisson and Montana, 2015; Bao and Chung, 2016; Milletari, 2017; Mehta et al., 2017). Combining patches of different views and dimensions is done in a multi-path manner, where CNNs consist of different branches corresponding to each patch type, i.e. parallel interconnected processing modules analyze each of the inputs. In contrast to patch-based CNNs, fully convolutional neural networks (FCNN) produce segmentation for a neighborhood of an input patch (Long et al., 2015). Shakeri et al. (2016) adapted the work of Chen et al. (2016) for semantic segmentation of natural images using FCNN. Moreover, 3D FCNNs, which segment a 3D neighborhood of an input patch at once, have been investigated by Dolz et al. (2018) and Wachinger et al. (2018). Although FCNNs show improvement in segmentation speed due to parallel segmentation of several voxels, they suffer from a high number of parameters in the network in comparison with patch-based CNNs.

It is common to apply post-processing methods to refine the final segmentation output. Inference of CNN-priors and statistical models such as Markov Random Fields and Conditional Random Fields (Lafferty et al., 2001) were used in the experiments of Brébisson and Montana (2015), Shakeri et al. (2016), and Wachinger et al. (2018). A modified Random Walker based segmentation refinement has been also proposed by Bao and Chung (2016).

Apart from implicit information that is provided by the extracted patches from MRI volumes, explicit characteristics distinguishing spatial consistency have been studied. Brébisson and Montana (2015) included distances to centroids to their networks. Wachinger et al. (2018) used the Euclidean and spectral coordinates computed from eigenfunctions of a Laplace-Beltrami operator of a solid 3D brain mask, to provide a distinctive perception of spatial location for every voxel. These kinds of features provide additional spatial information, however, extracting these explicit features from an unannotated MRI volume requires some preliminary operations to be attended (e.g. repetitive training of the network to compute initial segmentation mask).

From the reviewed literature, we have observed that most of the current deep learning approaches for sub-cortical brain structure segmentation focus on segmenting only the large sub-cortical structures (thalamus, caudate, putamen, pallidum). However, other important small structures (i.e. hippocampus, amygdala, accumbens), which are used for examining neurological disorders such as schizophrenia (Altshuler et al., 1998; Lawrie et al., 2003), anxiety disorder (Milham et al., 2005), bipolar disorder (Altshuler et al., 1998), Alzheimer (Fox et al., 1996), etc., are not considered. These small structures have smaller volume – hence, lower number of samples – compared to the other larger structures, which hinders training deep learning strategies and makes the segmentation task more challenging. In this paper, we present our approach for segmenting the sub-cortical structures: a new 2.5D CNN architecture – i.e., the three orthogonal views of a 3D volume – that incorporates probabilistic atlases as spatial features. Although probabilistic atlases have been used before in deep learning methods (Ghafoorian et al., 2017), they have never been applied for segmenting the sub-cortical brain structures. Within our research, unlike most of the existing deep learning approaches, we address segmenting all the sub-cortical structures, including the smallest ones. To the best of our knowledge, this is the first deep learning method incorporating atlas probabilities into a CNN for sub-cortical brain structure segmentation. Moreover, we propose a particular sample selection technique, which allows the neural network to learn to segment the most difficult areas of the structures in the images, and also show its importance in achieving higher accuracy. We test the proposed strategy in two well-known datasets: MICCAI 2012[4] (Landman and Warfield, 2012) and IBSR 18[5]; and compare our results with the classical and recent CNN strategies for brain structure segmentation. Additionally, we make our method publicly available for the community, accessible online at https://github.com/NIC-VICOROB/sub-cortical_segmentation.

2. Method

2.1. Input features

In our method, we employ 2.5D patches to incorporate information from three orthogonal views of a 3D volume. In our case, each patch has a size of 32×32 pixels. Although 3D patches may provide more information of surroundings for the voxel that is being classified, they are computationally and memory expensive. Thus, by using 2.5D patches, we approximate the information that is provided by a 3D patch in computational time and memory efficient manner.

Along with the appearance based features provided by the T1-w MRI, we employ spatial features extracted from a structural probabilistic atlas. In our experiments, we used the well-known Harvard–Oxford (Caviness et al., 1996) atlas template in MNI152 space distributed with the FSL package,[6] which has been built using 47 young adult healthy brains. In our method, first, T1-w image of the MNI152 template is affine registered to T1-w image of the considered datasets using a block matching approach (Ourselin et al., 2000). Then, non-linear registration of the atlas template to subject volume is applied using fast free-form deformation method (Modat et al., 2010). The deformation field obtained after the registration is used to move the probabilistic atlas into the subject space. Registration processes have been carried out using the well known and publicly available tool NiftyReg.[7] Afterwards, vectors of size 15, corresponding to seven anatomical structures with left and right parts separately and background, were

[3] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki.

[4] https://masi.vuse.vanderbilt.edu/workshop2012.

[5] https://www.nitrc.org/projects/ibsr.

[6] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki.

[7] http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg.

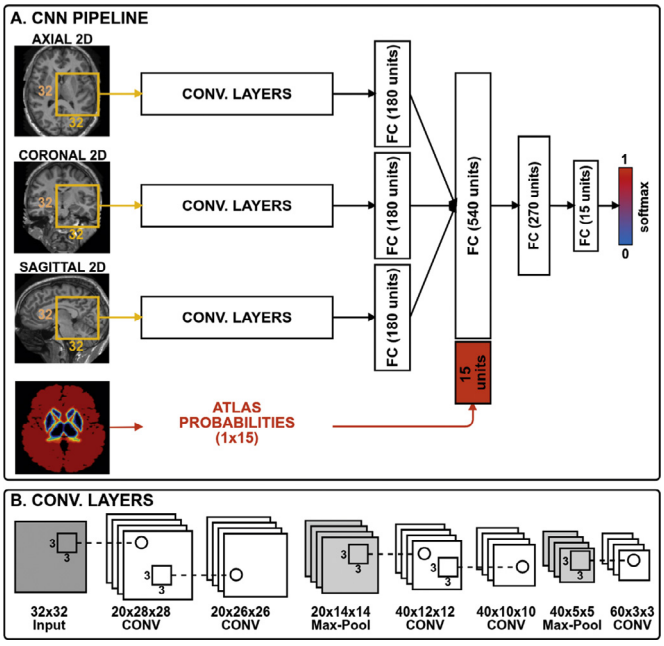


Fig. 1. The proposed 2.5D CNN architecture has three convolutional branches and a branch for spatial prior. 2D patches of size 32×32 pixels are extracted from three orthogonal views of a 3D volume. Spatial prior branch accepts a vector of size 15 with atlas probabilities for each of the 14 structures and background.

extracted from probabilistic atlas for every voxel and used as an input feature to train the network.

2.2. CNN architecture

Fig. 1 illustrates our proposed CNN architecture. It consists of three branches to process the patches extracted from axial, coronal, and sagittal views of a 3D volume, and one branch corresponding to the spatial priors. The branch for the spatial prior accepts a vector of size 15 with atlas probabilities for each structure and the background. The first three branches have the same organization of convolutional and max-pooling layers as shown in Fig. 1(B). All the feature maps of the convolutional layers are passed through the Rectified Linear Unit (ReLU) activation function (Glorot et al., 2011). For all the convolutional layers, kernels of size 3×3 are set to make the CNN deep without losing in performance and bursting the number of parameters as it has been studied in Simonyan and Zisserman (2014). Then, the outputs of the convolutional layers are flattened and followed by fully connected (FC) layers with 180 units each. Next, FC layers of each branch including atlas probabilities are fully connected to two consecutive FC layers with 540 and 270 units. The final classification layer has 15 units with the softmax activation function.

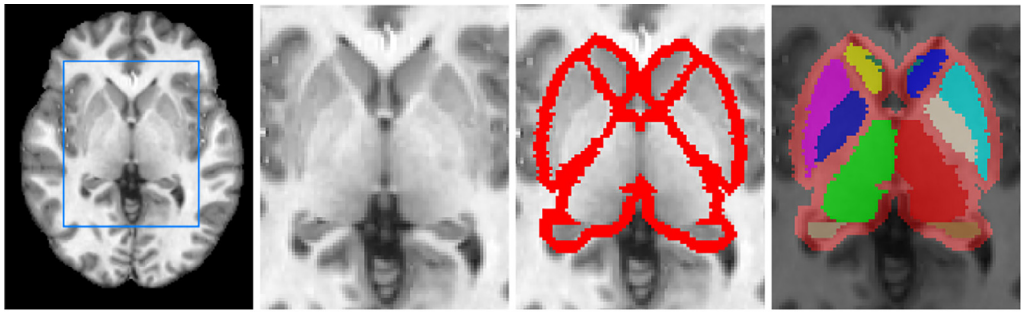
The atlas probabilities provide the network with spatial information, i.e. likelihood of an input patch belonging to one of the 14 classes or background. This information can be added either as additional input sequences (i.e. as additional channels to T1-w image patches) or later in the fully connected layers. However, when working with a high number of classes, the former way of atlas incorporation becomes impractical in terms of training/testing time due to an increase in number of trainable parameters of the network as well as a vast increase in memory usage. Accordingly, we use the latter approach, where we provide a vector of size 15 with each element corresponding to the central pixel's probability of belonging to one of the classes, which is fused with the output of the first fully connected layer after the convolutional part of the network.

2.3. CNN training

For training our network, we extract 2.5D patches from the training set and using the provided ground truth labels we optimize the kernel and fully connected layer unit weights based on the loss function. In the proposed network we employ the categorical cross-entropy loss function, which is minimized using the Adam (Kingma and Ba, 2014) optimization method. This technique automatically controls the learning rate and uses moving averages of the parameters, which allows the step size to be effectively large and converges to optimal step size without tuning it manually.

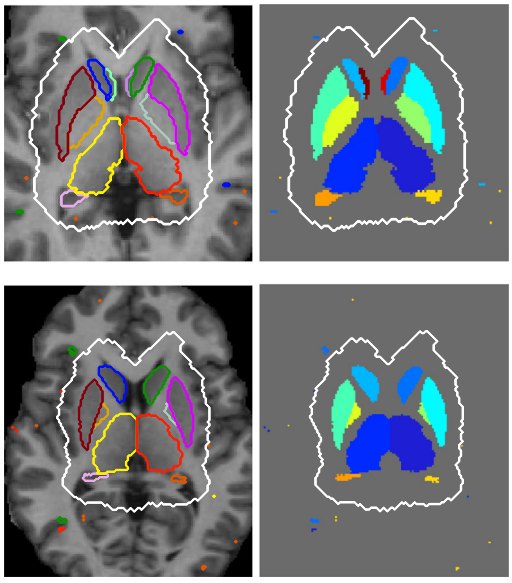
When training a CNN, it is important to take into account how the training samples are extracted from an image. Random selection of certain number of samples from an image is one of the common techniques in the literature. However, when it comes to the segmentation of the sub-cortical structures, the background (negative) samples turn out to be dispersed in the subject volume. Hence, it would lead to imperfect segmentation results on the borders of the structures, which are the most delicate areas to process due to the low contrast between the structure and the background. Therefore, we propose to extract the negative samples only from the structure boundaries as shown in Fig. 2. In doing so, we force the network to learn only from the structure boundaries and dismiss other parts of the background.

The training sample selection is performed as follows: from all the available training images, we first select the positive samples from all the voxels from the 14 sub-cortical structures. Then, the same number of negative samples are randomly selected from the structure boundaries within five voxel distance, forming a balanced dataset of sub-cortical and boundary voxels. More details about



(a) (b) (c) (d)

Fig. 2. Negative sample selection from the boundaries of the target structures. (a) T1-w image with a rectangle representing the ROI; (b) T1-w ROI; (c) structure boundaries; (d) ground truth labels with boundaries.



(a) (b)

Fig. 3. Two different examples of segmentation outputs without using ROI before post-processing. Columns: a) T1-w image and segmentation result; b) Segmentation output on solid background for better visualization of spurious outputs. ROIs are delineated in white.

batch size and number of epochs of the training process for the selected datasets will be given in Section 3.

2.4. CNN testing

To perform the segmentation of a new image volume, we extract all the patches from the image and predict class label probabilities using the trained CNN. Then, we assign a label corresponding to the maximum a posteriori probability for the central pixel of each input patch. Notice that knowing the order of the patch extraction is important to be able to reconstruct the final segmentation output. We also take advantage of the location of the sub-cortical structures, which are located in the central part of the brain. Due to the knowledge provided by the atlases, regions of interest (ROI) are automatically defined for all the subject volumes to achieve faster training and testing speeds.

Since the network has been trained with the negative samples extracted only from the structure boundaries, it produces spurious outputs in unseen areas of the background when segmenting a testing volume. In order to overcome this issue, we apply a post-processing step, where for each class only the region with the biggest volume within the ROI is preserved. For such post-processing, it is important to make sure that the volume and location of the misclassified regions are not larger than the volumes of any of the structures nor adjacent to the structure boundaries. When segmenting a new image, we send only ROI as an input to the network. In doing so, we ensure that the misclassified voxels have small size, as most of the input patches correspond to the sub-cortical area. Moreover, since the network is well trained to classify the boundaries of the structures, there will be no misclassified voxels adjacent to the structure boundaries. Fig. 3 illustrates examples, when all the patches were set as input to the network. As it can be observed, the background is well defined around the structure borders, and most of the spurious outputs appear outside the ROI.

2.5. Implementation and technical details

The proposed method has been implemented in the Python language,[8] using Lasagne[9] and Theano[10] (Bergstra et al., 2011) libraries. All experiments have been run on a GNU/Linux machine box running Ubuntu 16.04, with 32 GB RAM memory. CNN training has been carried out on a single TITAN-X GPU (NVIDIA corp, United States) with 12 GB RAM memory. The proposed method is currently available for downloading at our research website.[11]

3. Results

This section presents the results obtained by the proposed method on two datasets. The first dataset is the one provided in the MICCAI Multi-Atlas Labeling challenge[12] (Landman and Warfield, 2012) and the second is a publicly available dataset from the Internet Brain Segmentation Repository[13] (IBSR). Details of these datasets and the corresponding results will be given in Sections 3.2 and 3.3 respectively.

3.1. Evaluation measures

For evaluating the proposed method, we selected two metrics that are commonly used in the literature. These are overlap and spatial distance-based metrics, which show similarity and discrepancy of automatic and manual segmentations. The first measurement is Dice Similarity Coefficient (DSC) (Dice, 1945) defined as the following for automatic segmentation A and manual segmentation B:

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|}. \quad (1)$$

DSC measures the overlap of the segmentation with the ground truth on a scale between 0 and 1, where the former shows no overlap and the latter represents 100% overlap with the ground truth.

For the spatial distance based metric, Hausdorff Distance (HD) is used in our experiments. This metric is defined as a function of the Euclidean distances between the voxels of A and B as:

$$HD(A,B) = \max(h(A,B), h(B,A)),$$
$$h(A,B) = \max_{a \in A} \min_{b \in B} ||a - b||. \quad (2)$$

In other words, HD is the maximum distance from all the minimum distances between boundaries of segmentation and boundaries of the ground truth.

Similarly to Wachinger et al. (2018), we used Wilcoxon signed-rank test to test the statistical significance of: 1) the differences in DSC and HD between our and state-of-the-art methods; and 2) the effect of using spatial features and the proposed sample selection technique.

3.2. MICCAI 2012 dataset

This dataset consists of 35 T1-w MRI volumes split into 15 cases for training and 20 cases for testing. Manually segmented ground truth for each image is available as well, which contains 134 structures overall. In our experiments, we extracted 14 classes corresponding to seven sub-cortical structures with left and right parts separately. All the subject volumes have even voxel spacing of 1 mm^3 with a size of $256 \times 256 \times 256$ voxels in axial, sagittal, and coronal views respectively.

[8] https://www.python.org/.
[9] http://lasagne.readthedocs.io.
[10] http://deeplearning.net/software/theano/.
[11] https://github.com/NIC-VICOROB/sub-cortical_segmentation.
[12] https://masi.vuse.vanderbilt.edu/workshop2012.
[13] https://www.nitrc.org/projects/ibsr.

Table 1
MICCAI 2012 dataset results. Mean DSC ± standard deviation and HD ± standard deviation values for each structure obtained using FreeSurfer, FIRST, PICSL, and our method. Structure acronyms are: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R) and average value (Avg.). Highest DSC and HD values for each structure are shown in bold.

Method	FreeSurfer Fischl (2012)		FIRST Patenaude et al. (2011)		PICSL Wang and Yushkevich (2013)		Our method	
Structure	DSC	HD	DSC	HD	DSC	HD	DSC	HD
Tha.L	0.830 ± 0.018	4.94 ± 1.01	0.889 ± 0.018	4.65 ± 0.90	0.920 ± 0.013	**3.22 ± 0.99**	**0.921 ± 0.018**	3.39 ± 1.13
Tha.R	0.849 ± 0.021	4.76 ± 0.75	0.890 ± 0.017	4.39 ± 0.92	**0.924 ± 0.008**	**3.11 ± 0.79**	0.920 ± 0.016	3.31 ± 1.01
Cau.L	0.808 ± 0.079	9.89 ± 3.09	0.797 ± 0.046	3.56 ± 1.30	0.885 ± 0.074	3.44 ± 1.89	**0.894 ± 0.071**	**3.32 ± 2.00**
Cau.R	0.801 ± 0.042	10.39 ± 3.09	0.837 ± 0.117	4.16 ± 1.37	0.887 ± 0.065	3.60 ± 1.67	**0.892 ± 0.057**	**3.51 ± 1.67**
Put.L	0.771 ± 0.039	6.31 ± 1.09	0.860 ± 0.060	3.79 ± 1.76	0.909 ± 0.042	3.07 ± 1.40	**0.916 ± 0.023**	**2.63 ± 1.09**
Put.R	0.799 ± 0.026	5.85 ± 0.84	0.876 ± 0.080	3.26 ± 1.23	0.908 ± 0.046	2.91 ± 1.41	**0.914 ± 0.031**	**2.75 ± 0.99**
Pal.L	0.693 ± 0.189	3.89 ± 1.07	0.815 ± 0.088	2.89 ± 0.71	**0.873 ± 0.032**	2.52 ± 0.54	0.843 ± 0.101	**2.38 ± 0.76**
Pal.R	0.792 ± 0.085	3.45 ± 0.98	0.799 ± 0.060	3.18 ± 0.93	**0.874 ± 0.047**	**2.49 ± 0.59**	0.861 ± 0.049	2.59 ± 0.61
Hip.L	0.784 ± 0.054	6.35 ± 1.87	0.809 ± 0.022	5.49 ± 1.66	0.871 ± 0.024	**4.34 ± 1.66**	**0.876 ± 0.020**	4.48 ± 2.02
Hip.R	0.794 ± 0.025	6.19 ± 1.59	0.810 ± 0.140	4.80 ± 1.66	0.869 ± 0.022	4.01 ± 1.45	**0.879 ± 0.020**	**3.76 ± 1.23**
Amy.L	0.585 ± 0.064	5.05 ± 0.97	0.721 ± 0.053	3.54 ± 0.72	0.832 ± 0.026	**2.44 ± 0.29**	**0.833 ± 0.032**	2.39 ± 0.39
Amy.R	0.576 ± 0.076	5.43 ± 0.90	0.707 ± 0.054	4.11 ± 0.75	0.812 ± 0.033	**2.72 ± 0.50**	**0.821 ± 0.027**	2.72 ± 0.69
Acc.L	0.630 ± 0.055	4.28 ± 1.11	0.699 ± 0.089	6.81 ± 8.76	0.790 ± 0.050	2.57 ± 0.67	**0.799 ± 0.052**	**2.39 ± 0.64**
Acc.R	0.443 ± 0.065	5.47 ± 1.02	0.678 ± 0.081	3.93 ± 1.75	0.783 ± 0.058	2.65 ± 0.76	**0.791 ± 0.067**	**2.54 ± 0.65**
Avg.	0.725 ± 0.137	5.87 ± 2.48	0.799 ± 0.094	4.18 ± 2.76	0.867 ± 0.061	3.08 ± 1.27	**0.869 ± 0.064**	**3.01 ± 1.30**

3.2.1. Experimental details

Skull-stripping was applied to extract the brain and cut out other parts appearing in the MRI such as eyes, skull, skin, and fat using the BET algorithm (Smith, 2002). The spatial intensity variations on the MRI volumes were corrected using a bias field correction algorithm – N4ITK (Tustison et al., 2010), which is included in the publicly available ITK[14] toolkit. Both preprocessing methods were run with default parameters.

In our experiments, we trained a single model using the available training set of 15 images, while we tested the other 20 images as provided in the original MICCAI 2012 Challenge. From the training set, we extracted around 1.5M (750K of sub-cortical voxels and 750K of boundary voxels) sample patches of size 32×32 pixels from three orthogonal views, where around 1.1M (75%) were used for training and 400K samples for validation (25%). The extracted patches were passed to the network for training in batches of size 128. The network was set to train for 200 epochs, yet, we applied early stopping of the training process to prevent over-fitting. The training process was automatically terminated when the validation accuracy did not increase after 20 epochs.

3.2.2. Comparison with other available methods

The performance of the proposed approach is compared with widely used tools in medical practice – FreeSurfer and FIRST. We also compared the performance of our method with the one of PICSL (Wang and Yushkevich, 2013) method, which is a multi-atlas based segmentation strategy that uses joint fusion technique with corrective learning. PICSL was the winner of the MICCAI 2012 Challenge for brain structure segmentation and still shows the best results on this dataset. We used the default parameters for the methods of FreeSurfer and FIRST to produce segmentation masks for the testing volumes. Accordingly, the training and testing split matches the configuration we used for evaluating the proposed method. We have to note that, with this dataset, there were no individually reported numerical results for each of the sub-cortical structure in other CNN based approaches.

3.2.3. Results

Table 1 shows overall and per structure mean DSC and HD values on the MICCAI 2012 dataset. According to the results, our method showed significantly ($p < 0.001$) higher DSC of 0.869 than FIRST and FreeSurfer which yielded 0.799 and 0.725 overall mean DSC, respectively. Moreover, as it can be observed, the HD values showed similar behavior as DSC, where the proposed approach significantly outperformed both of these methods ($p < 0.001$), in average, with a reduction of 1.17 mm and 2.86 mm with respect to FIRST and FreeSurfer. Also, the DSC and HD results of our method with respect to FreeSurfer and FIRST were significantly higher for all the structures individually. Our method did not show a significant difference in comparison with PICSL in terms of DSC ($p > 0.05$), having similar mean of 0.867 and 0.869 for PICSL and our method, respectively. However, there was a significant improvement for the left caudate, right putamen, right hippocampus, and left accumbens structures ($p < 0.05$). The average HD values of our approach and PICSL also confirmed previous DSC numbers, but no significant increase per structure was observed. Fig. 4 shows a qualitative comparison of segmentation outputs from FreeSurfer, FIRST, PICSL, and our method. As it can be observed, FreeSurfer provided less accurate segmentation output with coarse structure boundaries. FIRST produced smooth segmentation on the borders, however, the overlap between the ground truth was poor. Our method's segmentation output was similar to the one of PICSL's and both of the methods had consistent structure boundaries, which were not far from the ground truth. Fig. 5c depicts an example of low DSC score (0.61) produced by our method for the right caudate structure. As it can be seen from the T1-w image, the intensities above the caudate structure are similar to the ones of the actual structure region defined by the manual segmentation (Fig. 5a). This irregularity led to an apparent atlas registration error, where a region outside the structure was defined with high atlas probabilities (Fig. 5b). Even though our network takes both – the intensities and the atlas probabilities – into account, these kinds of pathological cases may lead to inaccurate segmentation results. However, this is also a common issue for other methods as seen in Fig. 5d, where an atlas based method (PICSL) also fails in accurately segmenting this structure.

Apart from having similar results to the best performing method on this dataset, our strategy gained a good improvement in training and segmentation times. According to Landman and Warfield (2012), PICSL took 330 CPU hours for training 138 classifiers used for correcting systematic errors. Reported segmentation time of PICSL with optimal parameters was more than 50 minutes per subject volume (Wang and Yushkevich, 2013). In comparison with the above, the execution time of our CNN strategy was around 8 hours for training and less than 5 min for testing, including the atlas registration.

[14] https://itk.org/.

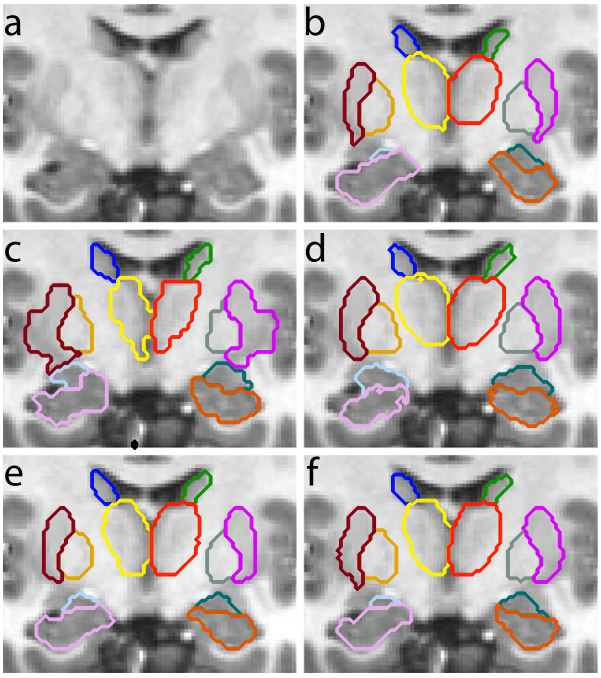


Fig. 4. Qualitative comparison of segmentation outputs obtained by FreeSurfer, FIRST, PICSL, and our method on MICCAI 2012 dataset. a) T1-w image; b) Ground truth; c) FreeSurfer; d) FIRST; e) PICSL; f) Our method. Visible structures on coronal view: thalamus, caudate, pallidum, putamen, hippocampus, and amygdala.

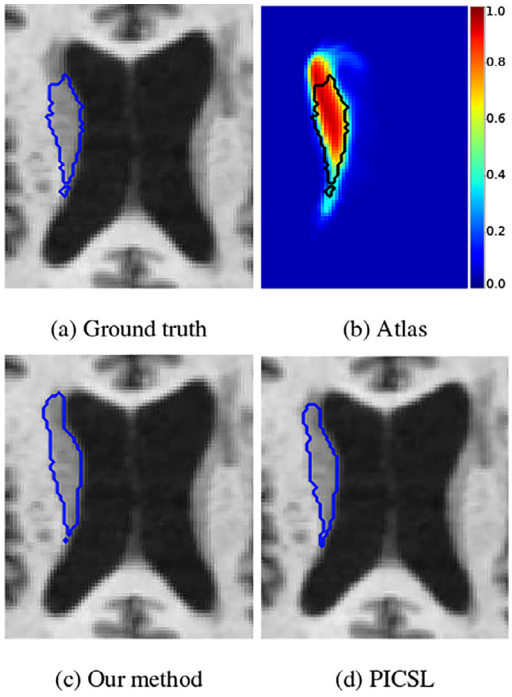


Fig. 5. Example of a segmentation result with a low DSC value for the right caudate structure. a) manual segmentation; b) probabilistic atlas with overlaid manual segmentation shown in black; c) our method; d) PICSL.

3.3. IBSR 18 dataset

This dataset consists of 18 T1-w subject volumes with manually segmented ground truth with 32 classes. Similarly to the MICCAI 2012 dataset, we extracted 14 classes corresponding to seven sub-cortical brain structures with left and right parts separately. The subject volumes of this dataset have dimensionality of $256 \times 256 \times 128$ and different voxel spacings: $0.84 \times 0.84 \times 1.5$ mm^3, $0.94 \times 0.94 \times 1.5$ mm^3, and $1.00 \times 1.00 \times 1.5$ mm^3. Images in this dataset have lower contrast and resolution in comparison with the MICCAI 2012 dataset, which makes the segmentation task even more challenging.

3.3.1. Experimental details

For the experiments with this dataset, we followed the same preprocessing steps as done with the MICCAI 2012 dataset, which included skull-stripping and bias field correction. Since there was no training and testing split on this dataset, we performed our experiments using a leave-one-subject-out cross-validation scheme. For each 17-1 fold, we extracted around 1.1M patches from each of the three orthogonal views, divided into 825K (75%) training and 220K (25%) validation sets. Each model was trained for 200 epochs applying also early stopping policy in the training process after 20 epochs.

3.3.2. Comparison with other available methods

For this dataset, our results will be compared against: 1) to the commonly used FreeSurfer and FIRST methods including the statistical significance test, since the evaluation values for each subject volume were computed by us using the corresponding tools; and 2) to recent CNN approaches of Shakeri et al. (2016), Mehta et al. (2017) (BrainSegNet), Bao and Chung (2016) (MS-CNN), and Dolz et al. (2018). The results for the recent methods were taken from their corresponding papers exactly as they have been reported. We have to mention that most of the CNN based methods report results only for a specific group of sub-cortical structures, but do not show or consider the results for the other, yet important, sub-cortical structures. Note also that the comparison on HD metric is present only for FreeSurfer, FIRST and our method, but not for other considered methods because most of the approaches do not report HD values.

3.3.3. Results

Table 2 shows the mean DSC and HD values for each of the evaluated methods. Our method showed a better performance in comparison to both FreeSurfer and FIRST methods for all the sub-cortical structures. The overall DSC mean of our method was significantly higher than both of the methods ($p < 0.001$), with mean DSC of 0.740, 0.808, and 0.843 for FreeSurfer, FIRST and the proposed strategy, respectively. In terms of HD values, our method showed overall mean of 4.49, whereas FreeSurfer and FIRST yielded 5.21 and 4.50, respectively. The proposed strategy significantly outperformed FreeSurfer with ($p < 0.001$), however the difference with FIRST was not significant ($p > 0.05$). As shown in Table 2, FreeSurfer performed worst for almost all the structures, while FIRST and our method showed similar performance. On both thalamus structures, our method showed lowest score in comparison with the other methods, however it yielded better HD for the small structures like amygdala, accumbens, and hippocampus. In general, HD metric is very sensitive to outliers, hence, a few misclassified voxels can cause considerable reduction in performance as seen in the results for the thalamus structure in our method.

Compared to other CNNs, our approach outperformed the method proposed by Shakeri et al. (DSC = 0.808) on the eight evaluated structures. Similarly, the performance of the proposed approach was also superior on the six structures evaluated in the work of Mehta et al. (DSC = 0.841). Further, we compare our method with MS-CNN, which has reported average DSC values for six structures for left and right parts together (overall DSC = 0.807). Our method's mean DSC on these structures was 0.859, which was higher than the result of MS-CNN (0.807) and yielded higher DSC scores for all the structures. Finally, when compared with the work of Dolz et al., our method showed a compa-

Table 2
Comparison of our method with the state-of-the-art methods as well as previous CNN approaches on IBSR dataset in terms of DSC, HD, and standard deviation. Structure acronyms are: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R). "–" represents no results were reported on corresponding structure. The average (Avg.) values show mean DSC for the presented structure DSC scores. Highest DSC and HD values for each structure are shown in bold.

Method	FreeSurfer		FIRST		Shakeri	BrainSegNet	MS-CNN	Dolz	Our method	
Struct.	DSC	HD	DSC	HD	DSC	DSC	DSC	DSC	DSC	HD
Tha.L	0.815 ± 0.056	5.367 ± 1.168	0.893 ± 0.017	**3.819 ± 0.850**	0.866 ± 0.023	0.88 ± 0.050	0.889	**0.92**	0.910 ± 0.014	7.159 ± 0.402
Tha.R	0.864 ± 0.022	**4.471 ± 1.245**	0.885 ± 0.012	4.273 ± 1.137	0.874 ± 0.021	0.90 ± 0.029			0.914 ± 0.016	7.256 ± 0.571
Cau.L	0.796 ± 0.050	6.435 ± 1.939	0.783 ± 0.044	4.128 ± 1.575	0.778 ± 0.053	0.86 ± 0.047	0.849	**0.91**	0.896 ± 0.018	**4.054 ± 1.412**
Cau.R	0.809 ± 0.048	8.201 ± 2.443	0.870 ± 0.027	**3.687 ± 0.791**	0.783 ± 0.068	0.88 ± 0.048			0.896 ± 0.020	4.153 ± 1.061
Put.L	0.789 ± 0.038	5.310 ± 0.923	0.869 ± 0.020	**4.421 ± 1.185**	0.838 ± 0.026	**0.91 ± 0.022**	0.875	0.90	0.900 ± 0.014	5.216 ± 1.788
Put.R	0.829 ± 0.031	4.716 ± 1.189	0.880 ± 0.010	4.725 ± 1.814	0.824 ± 0.039	**0.91 ± 0.023**			0.904 ± 0.012	**4.577 ± 0.410**
Pal.L	0.632 ± 0.171	4.652 ± 1.294	0.810 ± 0.033	**3.477 ± 0.572**	0.763 ± 0.031	0.81 ± 0.089	0.787	**0.86**	0.825 ± 0.050	3.849 ± 0.574
Pal.R	0.774 ± 0.032	3.966 ± 0.793	0.809 ± 0.037	3.990 ± 1.075	0.736 ± 0.055	0.83 ± 0.086			0.829 ± 0.046	**3.700 ± 0.576**
Hip.L	0.760 ± 0.036	5.787 ± 1.264	0.806 ± 0.023	5.571 ± 1.592	–	0.81 ± 0.065	0.788	–	**0.851 ± 0.024**	**4.177 ± 1.087**
Hip.R	0.767 ± 0.060	5.615 ± 1.600	0.817 ± 0.023	4.349 ± 0.984	–	0.83 ± 0.071			**0.851 ± 0.024**	**4.124 ± 0.824**
Amy.L	0.661 ± 0.069	5.521 ± 1.517	0.742 ± 0.064	4.648 ± 1.950	–	0.76 ± 0.087	0.654	–	**0.763 ± 0.052**	**4.326 ± 0.822**
Amy.R	0.690 ± 0.067	4.720 ± 1.553	0.757 ± 0.062	4.402 ± 1.493	–	0.71 ± 0.087			**0.768 ± 0.058**	**4.292 ± 1.064**
Acc.L	0.604 ± 0.071	3.634 ± 0.783	0.684 ± 0.098	7.770 ± 8.803	–	–	–	–	**0.744 ± 0.053**	**3.026 ± 0.676**
Acc.R	0.574 ± 0.074	4.507 ± 1.077	0.703 ± 0.076	3.733 ± 1.482	–	–			**0.752 ± 0.047**	**2.995 ± 0.609**
Avg.	0.740 ± 0.110	5.207 ± 1.761	0.808 ± 0.080	4.499 ± 2.810	0.808 ± 0.063	0.841 ± 0.064	0.807	0.898	0.843 ± 0.071	4.493 ± 1.533

Table 3
Effect of spatial features and the proposed sample selection technique. MICCAI 2012 dataset. Random sampling – method without using the sample selection from boundaries (including the spatial priors). No atlas – method without incorporating atlas priors (using the sampling technique). Final method – proposed method that includes both the spatial features and the sampling technique. Structure acronyms are: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R). The values with an asterisk (*) indicate that the final method obtained significantly higher results than that of the strategy without atlas priors. Highest DSC values for each structure are shown in bold.

Method	Random sampling	No atlas	Final method
Tha.L	0.860 ± 0.013	0.911 ± 0.024	**0.921 ± 0.017***
Tha.R	0.862 ± 0.014	0.917 ± 0.017	**0.920 ± 0.016**
Cau.L	0.831 ± 0.067	0.880 ± 0.103	**0.894 ± 0.071***
Cau.R	0.834 ± 0.048	0.864 ± 0.131	**0.892 ± 0.057**
Put.L	0.871 ± 0.024	0.900 ± 0.073	**0.916 ± 0.023***
Put.R	0.872 ± 0.027	0.913 ± 0.029	**0.914 ± 0.031**
Pal.L	0.784 ± 0.040	**0.852 ± 0.086**	0.843 ± 0.101
Pal.R	0.775 ± 0.057	0.833 ± 0.099	**0.861 ± 0.049***
Hip.L	0.778 ± 0.034	0.871 ± 0.019	**0.876 ± 0.020***
Hip.R	0.770 ± 0.026	0.876 ± 0.018	**0.879 ± 0.020***
Amy.L	0.709 ± 0.025	0.824 ± 0.037	**0.833 ± 0.032***
Amy.R	0.716 ± 0.054	0.819 ± 0.035	**0.821 ± 0.027**
Acc.L	0.744 ± 0.060	0.796 ± 0.052	**0.799 ± 0.052**
Acc.R	0.689 ± 0.091	0.753 ± 0.106	**0.791 ± 0.067***
Avg.	0.792 ± 0.076	0.858 ± 0.083	**0.869 ± 0.064***

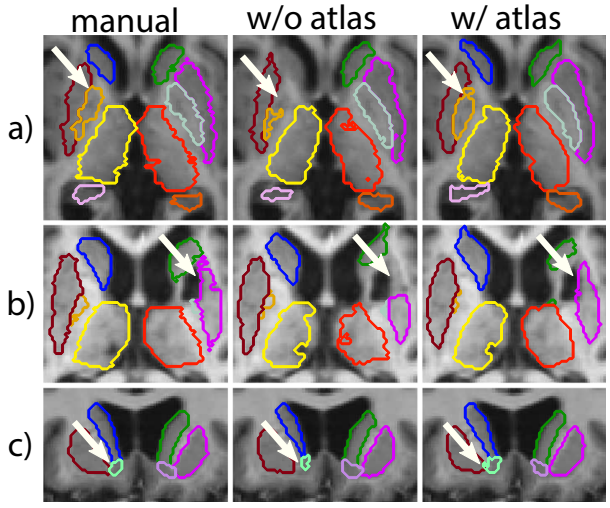


Fig. 6. Comparison of segmentation outputs for difficult areas of the (a) pallidum, (b) putamen, and (c) accumbens structures in some of the images from MICCAI 2012 dataset using the proposed method with and without the spatial priors. Regions of remarkable improvement when employing the atlas priors are indicated with arrows.

rable performance, although this last work showed slightly higher averaged DSC values for the four biggest structures.

3.4. Effect of the spatial priors

We ran experiments using the proposed method with and without spatial priors to determine the effect of using such features to the segmentation performance on both datasets. For this experiment, we analyzed the results in terms of DSC on the MICCAI 2012 dataset. We did not present the results of this experiment for the IBSR 18 dataset for simplicity, since it produced a similar outcome. In order to test our network without the spatial features, we modified the architecture (Fig. 1) by removing the branch of atlas probabilities and keeping only three branches of convolutional layers.

Table 3, shows DSC results of our method with random sampling, without using spatial features, and the final method. Inclusion of the spatial features significantly improved the overall DSC ($p < 0.001$), as well as the results for almost all the structures. The segmentation difference can be seen from Fig. 6, where difficult areas of the pallidum, putamen, and accumbens structures were segmented better by the method that comprised the spatial features. Hence, the spatial priors helped to overcome difficult areas, producing more accurate segmentation for some images that had intensity and shape irregularities that could not be observed in any of the training images. Although the spatial priors are effective to overcome these sort of issues, it could be misleading in certain cases, where the irregularity is extremely large - as shown in Fig. 6b, where a hole is present in the left pallidum structure. The final method obtained lower score for the left pallidum on this subject volume, which downgraded the average DSC for this structure (Table 3). On the other hand, our method without the spatial priors segmented this area better than the final approach, however, the overall difference was not significant ($p = 0.101$).

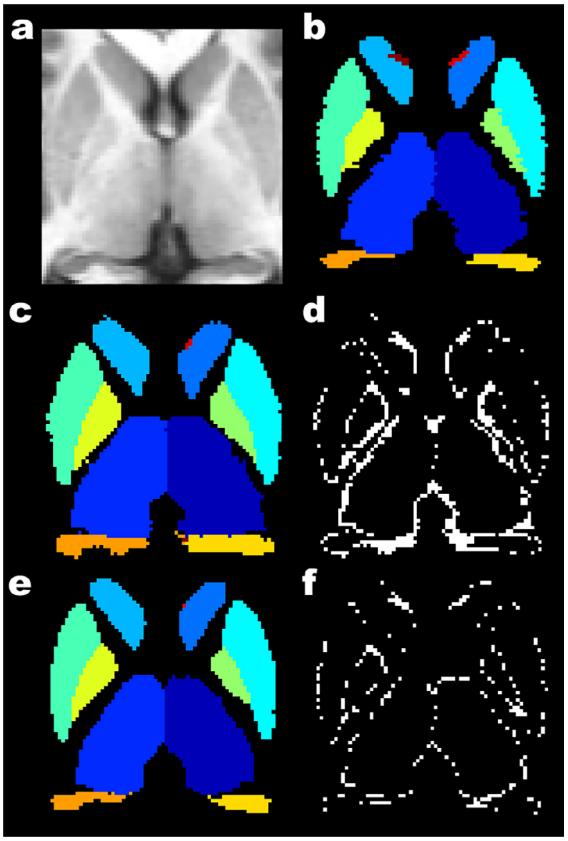


Fig. 7. Illustration of misclassification occurrence on borders. MICCAI 2012 dataset. (a, b) T1-w image and manual segmentation; (c, d) segmentation using random sample selection and difference from ground truth; (e, f) segmentation using the sample selection from borders and difference from ground truth.

3.5. Effect of sample selection

In this section, we show the effect of sample selection from structure boundaries using the MICCAI 2012 dataset. For this experiment, random sample selection from all the brain tissues has been used for training the network. For every epoch, we extracted the same number of voxels (1.5M), split equally into the sub-cortical structures (750K) and background (750K). Here, background voxels were randomly selected from whole brain volume, instead of selecting only from structure boundaries (see Fig. 2d). The network was again trained for 200 epochs using the same configuration. Spatial features were also included in training.

Table 3 shows the results corresponding to this experiment. Mean DSC obtained with our network without using the sample selection technique was 0.792 compared to 0.869 of the final approach. Accordingly, the proposed sample selection technique significantly improved the network's performance in average as well as for each of the structures ($p < 0.001$). Fig. 7 illustrates the segmentation results produced by our final approach and without applying sampling from borders. As it can be seen from the difference between ground truth and segmentation masks, the final strategy produced better segmentation on the boundaries than random sample selection method. In fact, the difference of our segmentation and the ground truth mask was not substantial, but only a few voxels. We also can observe that the intensities on the border voxels of the structures are mostly confounding. Therefore, assigning these voxels to the structure or background is highly dependent on ground truth.

4. Discussion

In this paper, we have proposed a fully automated 2.5D patch-based CNN approach that combines both convolutional and a priori spatial features for accurate segmentation of the sub-cortical brain structures. In our approach, a structural sub-cortical atlas has been registered into the image space to extract the spatial probability of each voxel, and, later, fused with the extracted convolutional features in the fully connected layers. The inclusion of the spatial information increases the execution time by adding atlas registration. However, it allows us to filter out misclassified regions that have bigger size than the actual structures in the segmentation output, which may appear in unobserved areas (i.e. not included in the training phase) of the brain as a consequence of applying restricted sampling. As seen in all the experiments, the addition of the spatial priors and the restricted sampling strategy have a significant effect on the accuracy of the proposed method, outperforming or showing a comparable performance to both classic as well as other novel deep learning approaches for segmenting the sub-cortical structures.

Compared to other state-of-the-art techniques such as FreeSurfer and FIRST, the spatial agreement of the proposed method with the manual segmentation is clearly higher in all evaluated datasets. As seen in other radiological tasks, this reinforces the effectiveness of CNN techniques when manual expert annotations are available. On the MICCAI 2012 dataset, our method shows an excellent performance, slightly over-performing the best challenge participant strategy – PICSL. Although not directly evaluated, our method clearly reduces the training and inference time. However, it has to be noted that most of the execution time of PICSL is due to highly computational registration processes which were carried out on CPU, while our method relies on GPU processors to speed-up training. Other CNN methods have also been evaluated on the MICCAI 2012 database (Wachinger et al., 2018; Mehta et al., 2017). However, these works do not report exact evaluation values for sub-cortical structures, hence, no direct comparison can be established.

In contrast, different CNN methods that have been evaluated using the IBSR 18 dataset have reported exact numerical values. When compared to other CNN approaches, our method also showed a significant increase in the performance with respect to most of them, and a comparable behavior with the method proposed by Dolz et al. However, as seen in Section 3.3, previous studies do not always deal with all sub-cortical structures, restricting a more detailed comparison with respect to our proposal. Additionally, the training methodology also differed among the strategies. In this aspect, although all our experiments were carried out using the leave-one-out approach, we also repeated our IBSR 18 experiments using a six-fold (15 training and three testing) validation strategy to perform a fair comparison with some of the considered methods. The complete results of the six-fold validation strategy were not depicted in the paper for simplicity, but, our network achieved similar results with only 0.005 of difference in DSC with respect to the leave-one-out strategy, showing the robustness of the proposed approach to changes in the number of training images.

According to the experimental results, employing the spatial features to the CNN significantly improved the performance of the network. The atlas priors showed to be useful in guiding the network when segmenting the difficult areas. As we have seen in Section 3.4, CNN that leveraged the spatial priors coped with these intensity based difficulties. Accordingly, by providing the atlas probabilities, we make sure that the anatomical shape and structure are taken into account before assigning a label to a voxel. Since the sub-cortical structures follow the similar anatomical structure in all patients, the inclusion of the spatial features

makes the segmentation approach more robust to irregularities in intensity based features obtained from T1-w images by providing additional location-based information. Despite being prone to the inherent errors in image registration and not showing as much DSC improvements as in border-selective sampling (Table 3), the addition of these a priori spatial class probabilities, or other explicit fused problem-specific information, may have other direct benefits such as reduction of the effect of low contrast, poor resolution, presence of noise, and artifacts close to the structure boundaries. Some examples of improvements in this regard were illustrated in Fig. 6.

Our results also showed the importance of sampling and class balancing in the training process. By feeding the network with only the most difficult negative samples, we ensured that useful samples were used in the training process. When compared to the rest of CNN approaches, our method without restricted sampling yielded a similar performance to other methods such as the one of Shakeri et al. (2016) and MS-CNN (Bao and Chung, 2016) even if trained on the same conditions, which highlights the effectiveness of the used sampling strategy. As a counterpart, these kind of approaches tend to generate false positive regions outside the subcortical space, due to the lack of contextual spatial information of the whole brain. Within our proposal, we took advantage of the already computed spatial priors to reduce the segmentation to only a region of interest containing the sub-cortical structures, which reduced remarkably the inference time. Remaining false positive voxels were then post-processed by maintaining only the biggest region for each class.

Our study comprises some limitations. Although our analysis shows that incorporating a-priori atlas information is effective on segmentation of the sub-cortical structures, there is room for further analysis of this approach in other brain segmentation tasks. Furthermore, the addition of atlas probabilities requires nonlinear registration, which may be tedious and prone to errors if applied on extreme cases such as advanced pathological subjects with a high degree of atrophy. Additionally, the extrapolation of our sample selection technique to other more general brain segmentation tasks should also be studied. As part of supervised training strategies, the accuracy of CNN methods tend to decrease significantly in other image domains (i.e. different MRI scanner, image protocol, etc.) than the ones used for training. Nevertheless, there is still a little evidence of the capability of CNN methods in radiological tasks with small or none datasets, which highlights the need of further studying this issue to increase the accuracy of such approaches. With no more evidence in this field, FIRST may be more appropriate in these scenarios when few or no training data is available. Another constraint involves the applicability of the proposed method on datasets of images with neurological diseases comprising, for instance, white matter lesions, which affect brain structure segmentation (González-Villà et al., 2017).

5. Conclusion

In this paper, we have presented a novel CNN based deep learning approach for accurate and robust segmentation of the sub-cortical brain structures that combines both convolutional and prior spatial features for improving the segmentation accuracy. In order to increase the accuracy of the classifier, we have proposed to train the network using a restricted sample selection to force the network to learn the most difficult parts of the structures. As seen from all the experiments carried out on the public MICCAI 2012 and IBSR 18 datasets, the addition of the spatial priors and the restricted sampling strategy have a significant impact on the effectiveness of the proposed method, outperforming or showing a comparable performance to state-of-the-art methods such as FreeSurfer, FIRST and different recently proposed CNN approaches. In order to encourage the reproducibility and the use of the proposed method, a public version is available to download for the neuroimaging community at our research website.

Acknowledgments

Kaisar Kushibar and Jose Bernal hold FI-DGR2017 grant from the Catalan Government with reference numbers 2017FI_B00372 and 2017FI_B00476, respectively. This work has been partially supported by La Fundació la Marató de TV3, by Retos de Investigación TIN2014-55710-R, TIN2015-73563-JIN, and DPI2017-86696-R from the Ministerio de Ciencia y Tecnologia. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN-X PASCAL GPU used in this research.

References

Altshuler, L.L., Bartzokis, G., Grieder, T., Curran, J., Mintz, J., 1998. Amygdala enlargement in bipolar disorder and hippocampal reduction in schizophrenia: an MRI study demonstrating neuroanatomic specificity. Arch. Gen. Psychiatry 55, 663–664.

Bao, S., Chung, A.C., 2016. Multi-scale structured CNN with label consistency for brain MR image segmentation. Comput. Methods Biomech. Biomed. Eng. 1–5.

Bergstra, J., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O., Desjardins, G., Goodfellow, I., Bergeron, A., Bengio, Y., Kaelbling, P., 2011. Theano: deep learning on GPUs with python. J. Mach. Learn. Res. 1, 1–48.

Brébisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–28.

Caviness V. S., Jr, Meyer, J., Makris, N., Kennedy, D.N., 1996. MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. J. Cognit. Neurosci. 8, 566–587.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2016. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv:1606.00915.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach.Intell. 23, 681–685.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vision Image Understanding 61, 38–59.

Debernard, L., Melzer, T.R., Alla, S., Eagle, J., Van Stockum, S., Graham, C., Osborne, J.R., Dalrymple-Alford, J.C., Miller, D.H., Mason, D.F., 2015. Deep grey matter MRI abnormalities and cognitive function in relapsing-remitting multiple sclerosis. Psychiatry Res. 234, 352–361.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Dolz, J., Desrosiers, C., Ayed, I.B., 2018. 3D fully convolutional networks for subcortical segmentation in MRI: a large-scale study. Neuroimage 170, 456–470.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118.

Fischl, B., 2012. Freesurfer. Neuroimage 62, 774–781.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Fox, N., Warrington, E., Freeborough, P., Hartikainen, P., Kennedy, A., Stevens, J., Rossor, M.N., 1996. Presymptomatic hippocampal atrophy in alzheimer's disease: a longitudinal MRI study. Brain 119, 2001–2007.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I.W., Sanchez, C.I., Litjens, G., de Leeuw, F.E., van Ginneken, B., Marchiori, E., Platel, B., 2017. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci. Rep. 7, 5110.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 315–323.

González-Villà, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., Lladó, X., 2016. A review on brain structures segmentation in magnetic resonance imaging. Artif. Intell. Med. 73, 45–69.

González-Villà, S., Valverde, S., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Oliver, A., Lladó, X., 2017. Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation. Neuroimage 15, 228–238.

Hazlett, H.C., Gu, H., Munsell, B.C., Kim, S.H., Styner, M., Wolff, J.J., Elison, J.T., Swanson, M.R., Zhu, H., Botteron, K.N., et al., 2017. Early brain development in infants at high risk for autism spectrum disorder. Nature 542, 348–351.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Kikinis, R., Shenton, M.E., Iosifescu, D.V., McCarley, R.W., Saiviroonporn, P., Hokama, H.H., Robatino, A., Metcalf, D., Wible, C.G., Portas, C.M., et al., 1996. A digital brain atlas for surgical planning, model-driven segmentation, and teaching. IEEE Trans. Visual. Comput.Graph. 2, 232–241.

Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Lafferty, J., McCallum, A., Pereira, F., et al., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, vol. 1, pp. 282–289.

Landman, B., Warfield, S., 2012. MICCAI 2012 workshop on multi-atlas labeling. In: Medical Image Computing and Computer Assisted Intervention Conference.

Lawrie, S.M., Whalley, H.C., Job, D.E., Johnstone, E.C., 2003. Structural and functional abnormalities of the amygdala in schizophrenia. Ann. New York Acad. Sci. 985, 445–460.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. S., Hipp, J. D., Peng, L., Stumpe, M. C., 2017. Detecting cancer metastases on gigapixel pathology images. arXiv:1703.02442.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Mak, E., Bergsland, N., Dwyer, M., Zivadinov, R., Kandiah, N., 2014. Subcortical atrophy is associated with cognitive impairment in mild parkinson disease: a combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis. Am. J. Neuroradiol. 35, 2257–2264.

Mehta, R., Majumdar, A., Sivaswamy, J., 2017. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. J. Med. Imaging 4. 024003–024003

Milham, M.P., Nugent, A.C., Drevets, W.C., Dickstein, D.S., Leibenluft, E., Ernst, M., Charney, D., Pine, D.S., 2005. Selective reduction in amygdala volume in pediatric anxiety disorders: a voxel-based morphometry investigation. Biol. Psychiatry 57, 961–966.

Milletari, F., et al., 2017. Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. Comput. Vision Image Understanding.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. 98, 278–284.

Ourselin, S., Roche, A., Prima, S., Ayache, N., 2000. Block matching: a general framework to improve robustness of rigid registration of medical images. In: MICCAI, volume 1935. Springer, pp. 557–566.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56, 907–922.

Phillips, J.L., Batten, L.A., Tremblay, P., Aldosary, F., Blier, P., 2015. A prospective, longitudinal study of the effect of remission on cortical thickness and hippocampal volume in patients with treatment-resistant depression. Int. J. Neuropsychopharmacol. 18. pyv037.

Pitiot, A., Delingette, H., Thompson, P.M., Ayache, N., 2004. Expert knowledge-guided segmentation system for brain MRI. Neuroimage 23, S85–S96.

Sarraf, S., Tofighi, G., et al., 2016. DeepAD: Alzheimers Disease Classification via Deep Convolutional Neural Networks Using MRI and fMRI. bioRxiv, p. 070441.

Shakeri, M., Tsogkas, S., Ferrante, E., Lippe, S., Kadoury, S., Paragios, N., Kokkinos, I., 2016. Sub-cortical brain structure segmentation using F-CNN's. In: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on. IEEE, pp. 269–272.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. Arxiv:arXiv:1409.1556.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved n3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320.

Wachinger, C., Reuter, M., Klein, T., 2018. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. Neuroimage 170, 434–445.

Wang, H., Yushkevich, P.A., 2013. Groupwise segmentation with multi-atlas joint label fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 711–718.