OpenCC – An Open Benchmark Dataset for Corpus Callosum Segmentation and Evaluation

Authors

Affiliations

Abstract. Neuroimaging studies have revealed that the structural changes of the corpus callosum (CC) are evident in a variety of neurological diseases, such as epilepsy and autism. Segmentation of the CC from magnetic resonance images (MRI) of the brain is a crucial step in the diagnosis of various brain disorders. However, the lack of open benchmark CC datasets has hindered development of CC segmentation techniques. In this work, we present an open benchmark dataset – OpenCC – for CC segmentation and evaluation. The dataset was built through alternative application of automatic segmentation and manual refinement. The automatic segmentation is based on recent advances in deep learning – fully convolutional networks, specifically U-Net, while the manual refinement is done by domain radiologists. The resulting dataset consists of 4643 mid-sagittal (or near mid-sagittal) slices and their corresponding CC masks. Furthermore, we provided some baseline segmentation results on the OpenCC dataset by using two different deep learning segmentation approaches. The OpenCC dataset can be used for comparison and evaluation of newly developed CC segmentation algorithms. We endeavor that, through the publishing of the OpenCC dataset and baseline segmentation results, we could promote further development of CC segmentation techniques.

Keywords: Image Segmentation; Corpus Callosum; Deep Learning; Open Benchmark Dataset; Fully Convolutional Networks.

1 Introduction

The corpus callosum (CC) is a large white-matter structure in the human brain that connects the left and right cerebral hemispheres, and transfers sensory, motor and cognitive information between the two hemispheres. Recent neuroimaging studies have revealed that the structural changes of CC occur in a variety of neurological diseases, such as epilepsy [1] and autism [2]. Anatomical and structural features of the CC, such as size and shape, can be used to provide clinically relevant information on neurological diseases. Therefore, accurate segmentation of CC is important. Traditionally, the neuroimages are examined by professional radiologists, and the brain tissue and its change are manually separated from surrounding brain tissues based on their observation, knowledge and experience. However, manual delineation of CC from MRI image is time-consuming, and subject to inter- and intra-observer variations. Automatic segmentation is therefore desirable to mitigate such variations.

Automatic segmentation of CC from MRI images has been studied extensively in the past years. Most of the techniques are based on Active Contour Model (ACM). Ginneken *et al.* [3] proposed an active shape model-based technique for CC segmentation using optimal local features that are chosen by statistical analysis of training images. Jacob *et al.* [4] proposed an ACM based technique for CC segmentation by introducing a robust gradient energy term as well as a new internal energy term. Sandhu *et al.* [5] proposed a Geometric Active Contour (GAC) technique, which incorporates image intensity probability density functions of the background and object into an active contour framework. Zhou *et al.* [6] proposed an automated segmentation technique for rodent brain tissues in MRIs by using support vector machines (SVMs) to obtain prior shape knowledge of objects and incorporating an automatic shape selection into existing active shape model framework. El-Zehiry *et al.* [7] proposed a novel technique to segment CC from white matter on the midsagittal plane. And Lai *et al.* [8] proposed an automated method to extract the boundary of CC on brain MRIs by two-step processing, i.e., initialization and refinement.

Other traditional methods for CC segmentation involve automated shape and appearance model based method [9], K-means segmentation method [10], Watershed based method [11], and many others. However, existing CC segmentation techniques suffer from several drawbacks: 1) most of the ACM methods need initial close-to-target localization which is not practical in real application; 2) existing methods cannot cope with the large variability in the shape of the object and the appearance of the edges of the CC; 3) each of the methods was tested on its own individual dataset, which renders it difficult to make any meaningful performance comparison among these methods.

In the past few years, deep learning techniques, especially convolutional neural networks (CNN) [12], have dramatically improved the state-of-art in speech recognition, visual object recognition, object detection and many other applications. Recently, deep learning techniques like fully convolutional networks have been deployed for brain structures segmentation from T1-weighted MRI images [13]. These approaches can be divided into two groups: work that focuses on normal structures [14, 15] and those on brain lesions [16, 17]. In both groups, different deep learning architectures have been introduced to address domain-specific challenges. For examples, for segmentation of normal structures, Zhang *et al.* [14] presented a 2D patch-wise CNN approach to segment white matter, gray matter, and cerebrospinal fluid from multimodal MR images of infants. While Kamnitsas *et al.* [16] presented a 3D dense-inference patch-wise and multi-scale CNN architecture for brain tumors segmentation.

Superior results have been reported with deep learning approaches compared to traditional machine learning algorithms. However, most of the existing deep learning approaches were targeted at more obvious structures like white matter and grey matter from axial slices. None of these deep learning approaches focused on CC segmentation from sagittal slices mainly due to the lack of a suitable dataset for CC segmentation evaluations. The motivation behind this study is to generate and publish an open benchmark dataset for the segmentation and evaluation of the CC.

In this study, through alternative application of automatic segmentation and manual refinement, we present an open benchmark dataset – OpenCC – for CC segmentation and evaluation. The automatic segmentation is based on recent development of deep

learning – fully convolutional networks, while the manual refinement is performed by domain radiologists. The resulted dataset consists of 4643 mid-sagittal (or near mid-sagittal) slices and their corresponding CC masks. We further provide some baseline segmentation results on the OpenCC dataset by using two different deep learning segmentation approaches, which can be used for comparison and evaluation of newly developed CC segmentation algorithms.

2 Methodology

2.1 Benchmark Dataset

The MRI data used in this study were obtained from the OASIS brains dataset-1 [18]. This set consists of a cross-sectional collection of 416 subjects aged 18 to 96. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. A trained expert segmented the corpus callosum from a subset of the MRI data (53 subjects, 439 MR image slices). The MR images and the corresponding segmented corpus callosum masks are used to train a deep neural network.

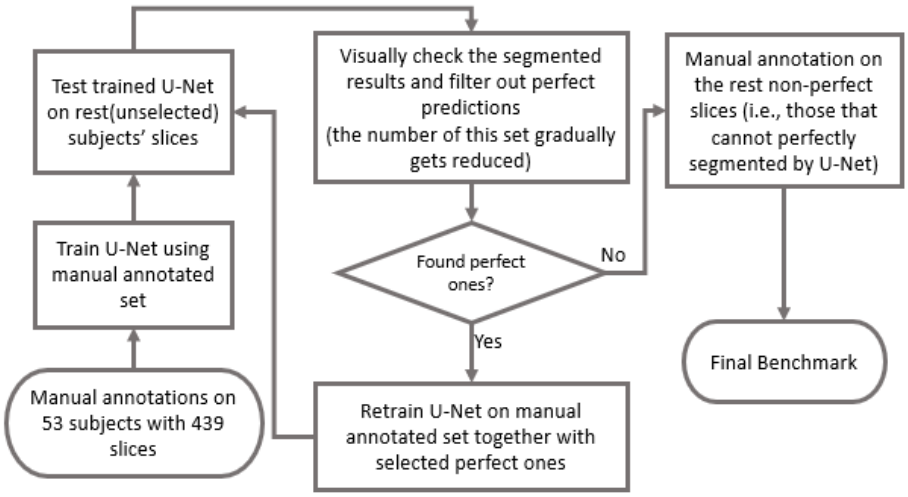

Figure 1: Workflow to generate OpenCC benchmark database

Through alternative application of automatic segmentation and manual refinement, we present an open benchmark dataset – OpenCC – for CC segmentation and evaluation. We use the popular U-Net [19] as the automatic segmentation approach, which is based on an encoder-decoder architecture. The encoder gradually reduces the spatial dimension with pooling layers and the decoder gradually recovers the object details and spatial dimension. Shortcut connections from the encoder to decoder help the decoder to recover object details better. After each iteration of automatic segmentation, the

domain expert will carefully check and refine the results. The workflow to generate OpenCC benchmark is shown in Fig.1. The detailed description is given below.

1. The domain expert manually annotates the CC for 53 patients (Patient #01-41 & #66-80), with 439 masks as the ground truth.
2. Using these 439 masks and their corresponding MR slices, we train a U-Net model for CC segmentation.
3. We test the trained U-Net model on the MR slices (using Slices #55-75 which most likely contain the CC) from the remaining patients, and segment the CC from the slices.
4. The domain expert visually inspects all the segmented CC results, and select the perfect ones, i.e., the segmented CC results match the true regions well.
5. We refine the U-Net model by training it with the selected CC slices, then test it on the remaining unselected slices.
6. We repeat Steps 4-5 until all the corpus callosum masks are deemed acceptable.
7. For any remaining slices that are unselected, i.e., the segmented CC masks do not match the true regions, the domain expert manually annotates these slices.
8. Finally, we obtain 4,643 CC masks for all the 416 subjects. The dataset is further split into training, evaluation, and testing set based on patients. The image size of both MR slices and generated masks is 256 pixels by 256 pixels.

2.2 Baseline Results

Deep learning is a fast-growing field with new architectures, variants, and algorithms appearing every few weeks. In this study, we test two latest deep learning segmentation methods on the CC benchmark dataset, and publish the results as baselines for comparison and evaluation with other newly developed CC segmentation algorithms. The first selected method is DeepLab-v3 [20], which presents an alternative to the classic encoder-decoder architectures. It advocates the usage of atrous convolutions for feature learning in multi-range contexts. The second approach is mask RCNN [21], which first performs a version of object detection to draw bounding boxes around each instance of a class, then performs semantic segmentation on each of the bounding boxes. Both methods have achieved admirable performance for segmentation tasks among existing state-of-the-art deep learning segmentation approaches.

We train DeepLab-v3 and Mask RCNN on the training and evaluation set, then perform inference on the testing set to derive the baseline results. To cater for small object segmentation, we used a loss function similar to [22] where loss = 1 - DM in which DM is the Dice Metric defined in Eqn. (1) in Section 3.2.

3 Results

3.1 Benchmark Dataset

Following the methodology in Section 2.1, we generated 4,643 CC masks for all the 416 subjects. Most subjects have 10-12 masks, while the minimum and maximum number of masks are 7 and 16, respectively. As an example, Table.1 shows the detail of the slices for 10 selected subjects with various number of CC masks. The full list of slices for all the 416 subjects is provided in the supplementary material. The visualization of some MR slices with CC masks can be found in Fig.2 (refer to first and second rows).

We further divide the benchmark into training, validation and testing sets by random splitting based on the subjects. This yielded 300 subjects for training, 50 subjects for validation, and 66 subjects for testing. The detailed list of subjects in the various sets is provided in the supplementary material. The benchmark dataset is also publicly available for download (the link is removed here due to blind review purposes). Interested readers may use this benchmark dataset to test the performance of their own CC segmentation methods.

Table.1 Slices with corresponding CC masks for Subjects #0101-0110

patient	*no. of slices*	*details*	*patient*	*no. of slices*	*details*
Subject #0101	11	*Slices #58-68*	*Subject #0106*	11	*Slices #60-70*
Subject #0102	11	*Slices #60-70*	*Subject #0107*	11	*Slices #59-69*
Subject #0103	11	*Slices #59-69*	*Subject #0108*	10	*Slices #61-70*
Subject #0104	11	*Slices #58-68*	*Subject #0109*	13	*Slices #59-71*
Subject #0105	12	*Slices #57-68*	*Subject #0110*	14	*Slices #56-69*
Total	416 subjects with 4,643 CC masks				

3.2 Baseline Results

We perform segmentation by DeepLab-v3 [20] and Mask RCNN [21] on the published openCC benchmark. For DeepLab-v3, we use the official tensorflow implementation[1], while for Mask RCNN we use the Keras implementation[2] based on tensorflow backend. The Adam optimizer with learning rate = 0.001 was used for compiling the model. The loss function 1-DM proposed in Section 2.2 was employed as the loss function for model training. The training and testing processes were performed on a workstation with four NVIDIA GeForce GTX 1080 Ti GPU cards. The training for

[1] Official tensorflow implementation of DeepLab-v3 https://github.com/tensorflow/models/tree/master/research/deeplab

[2] Kera implementation of Mask RCNN with tensorflow backend https://github.com/matterport/Mask_RCNN

DeepLab-v3 and Mask RCNN took about 3 and 5 hours, respectively, to complete for 200 epochs, and the subsequent segmentation testing took just 0.15-0.25 second per MR image by both trained models.

Table.2 Performance of DeepLab-v3 and Mask RCNN on OpenCC testing set in terms of DM and IoU evaluation metrics.

	Mean Value	Worst (Subject #)	Best (Subject #)
DeepLab-v3	DM = 0.955 IoU = 0.916	DM = 0.874 (Subject #0006) IoU = 0.779 (Subject #0006)	DM = 0.978 (Subject #0401) IoU = 0.958 (Subject #0401)
Mask RCNN	DM = 0.943 IoU = 0.895	DM = 0.872 (Subject #0380)* IoU = 0.783 (Subject #0031)*	DM = 0.966 (Subject #0127) IoU = 0.935 (Subject #0127)

* the subject with worst DM value is different from the subject with worst IoU value.

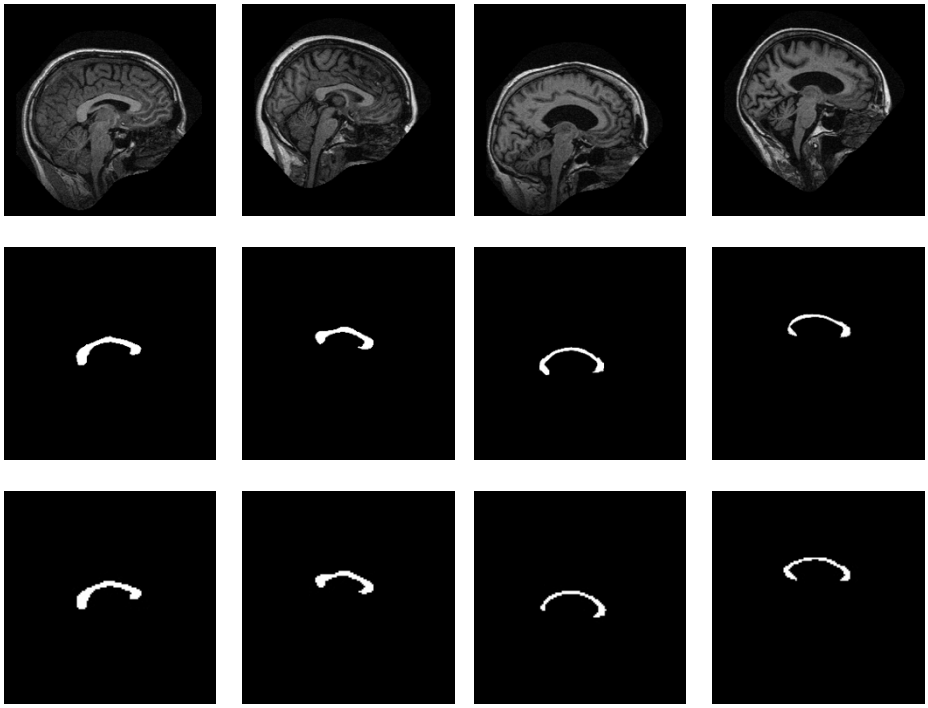

Fig.2 Visualization of the two slices (Slice #61 of Subject #0127 and Slice #64 of Subject #0378) with best IoU values (IoU = 0.980 and 0.981, respectively) as well as the two slices (Slice #72 of Subject #0428 and Slice #69 of Subject #0380) with worst IoU values (IoU = 0.544 and 0.556, respectively) by Mask RCNN. First row: input slice; second row: ground truth CC mask, third row: segmented CC mask.

To evaluate the CC segmentation performance, the most popular segmentation metrics, namely, the Dice metric (DM) and Intersection-over-Union (IoU) were used in this study, which are defined by

$$\mathrm{DM} = \frac{2(A \cap G)}{A+G} \quad \text{and} \quad \mathrm{IoU} = \frac{A \cap G}{A \cup G} \tag{1}$$

where $A \cap G$ and $A \cup G$ denote the intersection and union of automatically segmented region A and ground truth region G, respectively. Both DM and IoU are measures of the overlap between A and G, and they have values ranging from 0 to 1 such that a bigger value indicates better segmentation performance. Table.2 shows the evaluation metrics values of DeepLab-v3 and Mask RCNN on the proposed OpenCC benchmark dataset. DeepLab-v3 achieves an average DM of 0.955 and an average IoU of 0.916, which is slightly better than Mask RCNN, which achieves an average DM of 0.943 and an average IoU of 0.895. The evaluation metrics for each individual slice is provided in the supplementary material.

Fig.2 show the segmentation results on several slices (two with best IoU values and two with worst IoU values) by Mask RCNN for visualization purpose. The visualization of the results of DeepLab-v3 is attached in the supplementary material due to the space limitation here. It can be seen that for the best cases, the segmented masks are nearly identical to the ground truth masks, while for the worst cases, the segmented masks still retain the general shape of the ground truth though their thickness or the two end regions deviate from the ground truth masks.

4 Discussions and Conclusion

We implemented a workflow to generate a reliable database of segmented corpus callosum data. The goal of our approach is to minimize the need for human intervention in segmenting neuroanatomical structures. This has been achieved by progressively improving the deep learning based semantic segmentation approach. Another advantage of our workflow is that it also minimizes the amount of corrections needed for each 'mistake'. The annotator need only identify the 'good' segmentations which are used to improve the model. This is significantly less laborious than having to correct the annotations manually. It must be noted, however, that human intervention is ultimately still required to manage the challenging cases. For instance, in Fig 2, the corpus callosum has low contrast with the surrounding brain tissue, reducing segmentation performance. These ambiguous cases are challenging for even experienced human annotators. It could be expected that there may be some cases that are not represented in our dataset (image artifacts, pathological changes). Future work will evaluate the effectiveness of our approach on MR images with severe MR artifacts. We will also be looking into extending our approach to MR images with different contrasts.

References

[1] O'Dwyer, R., et al., *Differences in corpus callosum volume and diffusivity between temporal and frontal lobe epilepsy.* Epilepsy & Behavior, 2010. 19(3): p. 376-382.
[2] Hardan, A.Y., N.J. Minshew, and M.S. Keshavan, *Corpus callosum size in autism.* Neurology, 2000. 55(7): p. 1033-6.
[3] Ginneken, B.v., et al., *Active shape model segmentation with optimal features.* IEEE Transactions on Medical Imaging, 2002. 21(8): p. 924-933.
[4] Jacob, M., T. Blu, and M. Unser, *Efficient energies and algorithms for parametric snakes.* IEEE Transactions on Image Processing, 2004. 13(9): p. 1231-1244.
[5] Sandhu, R., T. Georgiou, and A. Tannenbaum. *A new distribution metric for image segmentation*. 2008.
[6] Jinghao, Z., et al. *A novel learning based segmentation method for rodent brain structures using MRI*. in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2008.
[7] El-Zehiry, N., M. Casanova, and A. Elmaghraby. *Variability of the relative corpus callosum cross sectional area between dyslexic and normally developed brains*. in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2008.
[8] Lai, R., et al. *Automated corpus callosum extraction via Laplace-Beltrami nodal parcellation and intrinsic geodesic curvature flows on surfaces*. in *2011 International Conference on Computer Vision*. 2011.
[9] Brejl, M. and M. Sonka, *Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples.* IEEE Transactions on Medical Imaging, 2000. 19(10): p. 973-985.
[10] Bhalerao, G.V. and N. Sampathila. *K-means clustering approach for segmentation of corpus callosum from brain magnetic resonance images*. in *International Conference on Circuits, Communication, Control and Computing*. 2014.
[11] Freitas, P., et al. *Watershed-Based Segmentation of the Midsagittal Section of the Corpus Callosum in Diffusion MRI*. in *2011 24th SIBGRAPI Conference on Graphics, Patterns and Images*. 2011.
[12] Y. Lecun, et al., *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 1998. 86(11): p. 2278-2324.
[13] Akkus, Z., et al., *Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions.* Journal of Digital Imaging, 2017. 30(4): p. 449-459.
[14] W. Zhang, et al., *Deep convolutional neural networks for multi-modality isointense infant brain image segmentation.* NeuroImage, 2015. 108: p. 214-224.
[15] P. Moeskops, et al., *Automatic Segmentation of MR Brain Images With a Convolutional Neural Network.* IEEE Transactions on Medical Imaging, 2016. 35(5): p. 1252-1261.
[16] K. Kamnitsas, et al., *Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation.* Medical Image Analysis, 2017. 36: p. 61-78.
[17] M. Havaei, et al., *Brain tumor segmentation with Deep Neural Networks.* Medical Image Analysis, 2017. 35: p. 18-31.
[18] D.S. Marcus, et at., *Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults*. Journal of Cognitive Neuroscience, 2007, 19(9), 1498-1507.
[19] E. Shelhamer, J. Long, and T. Darrell, *Fully Convolutional Networks for Semantic Segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. 39(4): p. 640-651.
[20] L.C. Chen, et al: Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017
[21] K. He, G. Gkioxari, P. Dollar, and R. Girshick, Mask R-CNN, arxiv: 1703.06870, 2017
[22] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, arxiv: 1606.04797, 2016