Reproducible White Matter Tract Segmentation Using 3D U-Net on a Large-scale DTI Dataset

Bo Li[1,2](✉), Marius de Groot[2], Meike W. Vernooij[2], M. Arfan Ikram[2], Wiro J. Niessen[2,3], and Esther E. Bron[2]

[1] Northeastern University, Shenyang, China
[2] Erasmus MC, Rotterdam, The Netherlands
{b.li,e.bron}@erasmusmc.nl
[3] Delft University of Technology, Delft, the Netherlands

Abstract. Tract-specific diffusion measures, as derived from brain diffusion MRI, have been linked to white matter tract structural integrity and neurodegeneration. As a consequence, there is a large interest in the automatic segmentation of white matter tract in diffusion tensor MRI data. Methods based on the tractography are popular for white matter tract segmentation. However, because of the limited consistency and long processing time, such methods may not be suitable for clinical practice. We therefore developed a novel convolutional neural network based method to directly segment white matter tract trained on a low-resolution dataset of 9149 DTI images. The method is optimized on input, loss function and network architecture selections. We evaluated both segmentation accuracy and reproducibility, and reproducibility of determining tract-specific diffusion measures. The reproducibility of the method is higher than that of the reference standard and the determined diffusion measures are consistent. Therefore, we expect our method to be applicable in clinical practice and in longitudinal analysis of white matter microstructure.

Keywords: White Matter · Tract · Low resolution · DTI
Diffusion measurements · Segmentation · Convolution neural network
3D

1 Introduction

White matter (WM) tracts are the neural fibers enabling communication among brain regions. The changes in which have increasingly been associated with cognitive dysfunction and neurodegeneration. For improving understanding of neurodegenerative process and the study of pathogenesis triggered by abnormal changes, a quantitative description of WM tract is essential. Therefore, a precise segmentation method used for quantifying WM tract is needed [1].

Y. Shi et al. (Eds.): MLMI 2018, LNCS 11046, pp. 205–213, 2018.
https://doi.org/10.1007/978-3-030-00919-9_24

Tract segmentation is typically performed by tractography followed by a filtering step based on the prior information. After tractography reconstruction, millions of possible pathways are filtered into specific tract either via tract-specific thresholds [2], anatomical atlas based mask [3,4] or neighboring anatomical labels based prior probability [5]. However, these steps result in accumulating intermediate errors, multiple environment settings and limited consistency due to the property of tractography and, therefore, limit their application in clinical practice.

The U-Net architecture [6] has shown good performances in several segmentation tasks. Based on 3D U-Net, the newer V-Net [7] made further improvements by introducing residual function, strided convolution and convolution transpose operations. Recently a U-Net based WM tract segmentation method [8] showed competitive results to tractography-based methods. Model in [8] was trained on a high resolution dataset of only 20 subjects. In this paper we develop a method based on a large dataset of lower resolution data, and evaluate the potential of the method in this setting.

This work presents a novel deep learning method for direct segmentation of white matter tract. Our method was evaluated on the tasks of FMI and CST segmentation and determining diffusion measures. We will evaluate whether this method is reproducible and can be used to provide more insight into the role of WM microstructure in neurodegeneration.

2 Methods

2.1 Model

We built our model based on the 3D U-Net architecture. We added batch normalization after each convolution layer and replaced Relu activation function with PRelu. The used convolution layers are 3D with a kernel size of $3 \times 3 \times 3$.

As input to the model, voxel-wise diffusion tensor elements were used. The input was fed in random batches during each training iteration to increase the robustness. Its batch generation was "on-the-fly" paralleled to the training process for efficiency. The method outputs a binary segmentation of a specific tract.

2.2 Dataset

The method was developed based on the dataset of Rotterdam Study, an ongoing, population based cohort study [9]. After quality assessment, 9149 MRI scans from 4983 non-demented subjects were available for this work. Scans were performed at 1.5 Tesla. The diffusion weighted images (DWIs) were acquired with a maximum b-value of $1000\,\mathrm{s/mm}^2$ in 25 gradient directions. Voxel size was resampled from $2.2 \times 3.3 \times 3.5\,\mathrm{mm}^3$ to $1\ \mathrm{mm}^3$.

We assign these scans into an optimization set (D1), a validation set (D2) and a reproducibility set (D3). Their sizes are as follows: $D1a_{train}$ 864 subjects, $D1a_{test}$ 218 subjects, size is same for $D1b_{train}$ and $D1b_{test}$ but with different subjects; $D2_{train}$ 7162 scans (including D1), $D2_{validate}$ 200 subjects and $D2_{test}$ 1036 subjects; $D3_{test}$ 80 subjects. The subjects (mean age of 69.7 years) in $D3_{test}$ had been scanned twice (mean interval of 19.3 days). A separate cohort was used for $D2_{validate}$ and $D2_{test}$ to ensure this is completely independent from $D2_{train}$. Additionally, all $D3_{test}$ related scans, which are their other rounds of scan, were excluded from D2 for the purpose of reproducibility evaluation.

2.3 Preprocessing

DWIs were corrected motion and eddy currents by co-registering all diffusion weighted volumes to the $b = 0$ volume with Elastix [10]. Diffusion tensors were estimated with ExploreDTI [11]. Diffusion measures, such as fractional anisotropy (FA) and mean diffusivity (MD), were computed based on the estimated tensors.

The diffusion tensor imaging (DTI) was used because this was the most suitable model for low-resolution DWIs. To evaluate location and structure information, with FLIRT [12] we registered the MNI_152 template and T1 weighted image (T1) to DTI space where most features were computed. Tissue masks including WM and gray matter (GM) were applied on all features. Due to the large image size and computation limitation of 3D convolution, we computed the region of interest (ROI) as input based on tract bounding boxes. The ROI sizes are 96 × 64 × 64 (FMI) and 64 × 96 × 128 (CST).

2.4 Reference Standard

As reference standard we used a clinical-accepted method [2], which consists of probabilistic tractography and tract-specific thresholds. Manual annotation can not be obtained as WM tracts are not visible on imaging and the semi-manual annotation on tractography images is also unrealistic for such a large dataset.

The method was evaluated on FMI and CST tract, since they are significantly related to aging [2], anatomically distinctive and have different degrees of difficulty for segmentation [8].

2.5 Evaluation Metrics

Segmentation accuracy was quantified by the Dice coefficient (DC). Binary segmentations were created from the probabilistic output by thresholding by 0.5.

To evaluate reproducibility, tract-specific metrics were compared between our method and the reference standard. Median FA and MD were individually computed inside the segmented tract, then averaged over $D3_{test}$. We computed the R^2 value of ordinary least squares (OLS) regression for measures in both scans. Cohen's kappa (K), which measures inter-rater agreement, was computed by rigidly aligning the FA image of rescan to the space of the first scan. We used t-test to compare K and paired scan-rescan differences of FA, MD and volumes with those of the reference standard, and used paired t-test to compare whether the measures determined by our method are consistent in both scans.

3 Experiments

The experiments were ran on one node of Cartesius, Dutch national supercomputer, with the Intel E5-2450 v2 CPU and NVidia Tesla K40m GPU.

3.1 Method Optimization Experiments

We optimized the method using the FMI tract on three key elements: (1) input, (2) the loss function and tract weight, and (3) network architecture. Experiments (1) and (3) were performed on $D1a$ and $D1b$, trained with default parameters of optimizers; experiment (2) was performed on $D2$. The following paragraphs will describe these optimization experiments.

We trained the V-Net based model with eleven different inputs. Nadam optimizer [13] and weighted inner product [14] loss function (L_{wip}) were used. The choice of input is based on DC and computation consumption. Since there are 25 diffusion weighted volumes in our raw DWIs and the number of volumes increases with resolution, e.g. 270 volumes in 7T scanner, it's an essential step to choose a concentrated and generalized input that works on different scanners. We considered diffusion tensor, FA, MD, location and T1 for training the model. we experimented to find the efficient input to avoid overlapping information and to reduce our high computation load due to the 3D convolution and large dataset.

To compare the L_{wip} and weighted cross entropy (L_{wce}) loss function, we trained two V-Net based models using

$$L_{wip} = -\frac{1}{N}\sum_{i=1}^{N} W * r_i * p_i + (1 - r_i) * (1 - p_i) \tag{1}$$

and

$$L_{wce} = -\frac{1}{N}\sum_{i=1}^{N} W * r_i * log(p_i) + (1 - r_i) * log(1 - p_i), \tag{2}$$

where $r_i \rightarrow \{0, 1\}$ is the reference standard, $p_i \rightarrow \{0, 1\}$ is the binarized prediction, N is the voxel number of the input and $W = [1, 3, 5, 10, 100]$ is the weight of tract. Due to the great frequency imbalance between classes, we evaluated different weights (W) for FMI segmentation, ranging from 1 to the mean frequency ratio of non-tract relative to tract, i.e. 100. Models were trained using Adam optimizer with an initial learning rate of 0.1, which was automatically reduced by 50% once the validation loss stopped improving for 10 epochs.

Similarly, to investigate if the newer V-Net architecture performs better than 3D U-Net in WM tract segmentation, two separate models were trained using diffusion tensor input and L_{wip}. Furthermore, to avoid the chance that one gradient descent algorithm works better for a particular back-propagation pathway, we doubled the number of experiments using Adam [15] and Nadam optimizer.

3.2 Validation Experiments

The optimized method was trained for FMI and CST tract on $D2_{train}$ to evaluate accuracy ($D2_{test}$) and reproducibility ($D3_{test}$). For $D3_{test}$, because of the short time interval between two scans (i.e. 19.3 days on average), tract segmentations and diffusion metrics are expected to be identical. We computed the paired scan-rescan differences, mean, standard deviation, R^2 value for FA, MD and volumes inside the segmented tracts in both scans and the Cohen's kappa to evaluate reproducibility of both segmentation and determining diffusion measures.

4 Results

4.1 Method Optimization Results

Figure 1 (left) presents the test DC of FMI for different combinations of input images. The figure shows that all combinations gave similar performances. Therefore, we used the simplest and most computation-efficient input, i.e. tensor only.

The performance when varying the loss function (L_{wip} and L_{wce}) and tract weight is provided in Fig. 2. L_{wip} in combination with $W = 3$ gave the best result. Both loss functions had instable performance when $W > 5$, especially L_{wce}. Based on the comparison, we used L_{wip} ($W = 3$) in the remainder of the experiments.

Since Fig. 1 (right) shows that the 3D U-Net architecture in combination with the Adam optimizer yielded a better performance than the other methods using either a V-Net architecture or the Nadam optimizer, we will adopt this combination in our method.

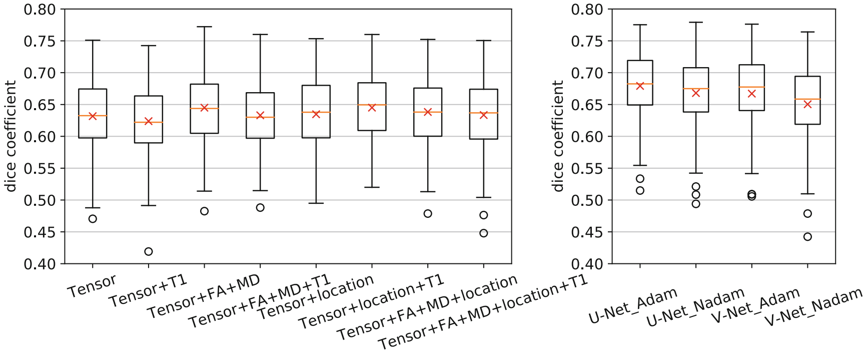

Fig. 1. Test dice coefficient of FMI for different: (left). Input images. The "Location" is an image of voxel-wise coordinates on MNI_152 template; (right). Architecture and optimizer.

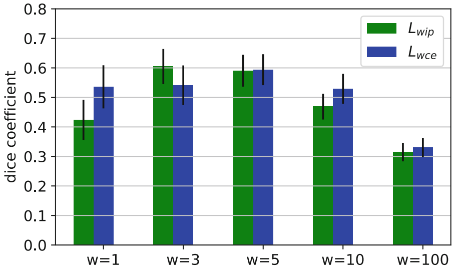

Fig. 2. Test dice coefficient of FMI using L_{wip} and L_{wce} loss function. W indicates the weight of tract.

4.2 Validation Results

Figure 3 provides a visualization of our segmentation result. It overlaps with the reference standard in (a) and (c) for FMI and right CST, respectively. The mean test DC of FMI is 0.66 (SD 0.06), that of CST is 0.77 (SD 0.03).

Figure 3(b)(d) provide its overlaps with segmented rescan, which was registered by rigidly aligning the FA images. Table 1 gives the reproducibility statistics. Typically, a $K > 0.60$ indicates "substantial" agreement between raters, and a $K > 0.80$ for "almost perfect" [16]. Our mean K for FMI longitudinal-segmentations achieved 0.74 and 0.80 for CST. The R^2 and K show that our method has better reproducibility than reference. Moreover, there was no difference in our longitudinal-measures (FA, MD, volume, paired t-test, $p > .1$). Our mean FA and MD are consistent with that of the reference. These results show that our method is applicable in longitudinal analysis of WM microstructure.

Figure 4 provides subject-wise reproducibility in determining diffusion measures. The Bland-Altman plots show that almost all differences are within the 95% limits of agreement and the mean of which is close to zero, indicating no consistent bias in longitudinal-measures. Additionally, Fig. 4 (right) shows that the MD is a discriminative feature for FMI and CST tract.

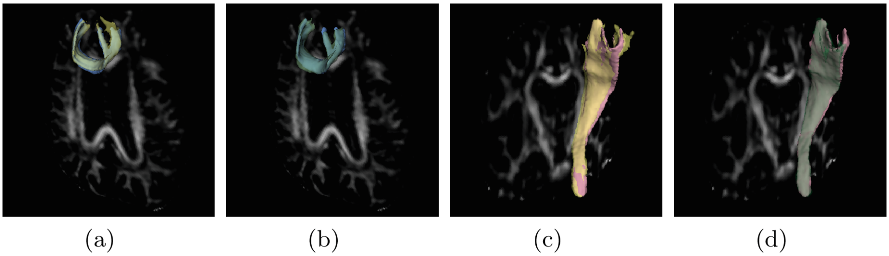

Fig. 3. Visualization of segmentation results: (a) FMI (blue) and reference (yellow), $DC = 0.67$; (b) FMI of the first scan and rescan (green), $K = 0.79$; (c) right CST (pink) and reference, $DC = 0.76$; (d) right CST of the first scan and rescan, $K = 0.84$.

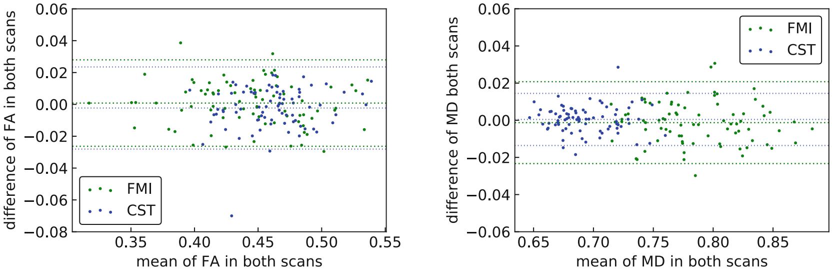

Fig. 4. The Bland-Altman plots. Difference (y-axis) and mean (x-axis) of diffusion measures (left) FA, (right) MD inside the segmented FMI and CST tract in both scans.

Table 1. Tract-specific reproducibility statistics. MD x $10^{-3}mm^2/s$. The "ref" indicates reference standard; "prop" indicates proposed method; SD standard deviation; "diff (SD)" indicates averaged absolute-differences between both scans; "mean (SD)" indicates mean value over all scans; R^2 is the R^2 value of OLS regression for measures in both scans; "Vol" indicates tract volume in ml; "K" indicates the Cohen's kappa; * significantly ($99\% - CI$) improved from the reference, t-test, $p < .01$.

	FA			MD			Vol			K
	diff	mean	R^2	diff	mean	R^2	diff	mean	R^2	
FMI										
ref	.012 (.009)	.44 (.04)	.89	.0082 (.007)	.79 (.04)	.93	.26 (.21)	3.3 (.53)	.66	.64 (.02)
prop	.011 (.008)	.44 (.05)	.91	.0089 (.007)	.79 (.04)	.93	.23 (.16)	3.8 (.58)	.79	.74* (.01)
CST										
ref	.011 (.008)	.46 (.03)	.83	.0053 (.005)	.70 (.03)	.92	.64 (.50)	6.1 (.93)	.39	.72 (.04)
prop	.009 (.003)	.46 (.03)	.84	.0052 (.004)	.69 (.03)	.93	.41*(.26)	6.5 (.69)	.52	.80* (.07)

5 Discussion

We developed and evaluated a novel deep learning method for direct WM tract segmentation. The method was trained and applied on a large set of low resolution DTI images and showed very good reproducibility. Therefore it can be applied to longitudinal imaging studies to investigate the process of neurodegeneration in WM microstructure as can be assessed with diffusion MRI.

Strengths of this study are the large size of dataset, which is representative of clinical variation, and the reproducibility validation in both segmentation and determining diffusion measures. Reproducibility is an essential indicator of a method that can be applied in clinical practice to ensure reliable and reproducible results. Moreover, comparing with the tractography-based methods, our direct method enables to segment a 3D tract in 0.5 s, and therefore avoid the processing time and storage space of tractography for researchers who only focus on the analysis of diffusion measures.

Based on the results of FMI segmentation we concluded that both the dependency of train and test datasets and their respective sizes are important for the resulting performance. If a much older training ($D1b_{train}$) than testing ($D2_{test}$) dataset is used, performance is suboptimal ($DC = 0.61$, $D1b_{train}/D2_{test}$ vs. $DC = 0.68$, $D1b_{train}/D1b_{test}$). On the other hand, a large, diverse and test-independent training dataset increases the robustness and difficulty of learning at the same time ($DC = 0.66$, $D2_{train}/D2_{test}$).

The paper by Wasserthal et al. [8] is the only published deep learning method of WM tract segmentation that we are aware of. Our test DC of the right CST ($DC = 0.77$) is lower than that reported in [8] ($DC = 0.83$). The main differences between two works are: they takes high resolution (7T) based input and semi-manual annotated reference, stacks four 2D models and is tested on only 5 subjects; while ours is applicable for a low-resolution dataset (1.5T), uses one 3D model and tested on a train-independent cohort of 1036 subjects. We suspect that the differences in the quality of the reference standard and the data are the main causes of this performance difference.

A limitation of our method is that we take a single tract ROI. This is mainly because of the large whole input size of 210 × 211 × 123 × 6 and the limitation of 3D convolution, which used for preserving the continuity of the tract. Another limitation is our low quality reference standard. Validation is difficult since the semi-manual annotation can not be obtained for such a large dataset.

For future work, our method will be applied in a dementia population. We will tackle the computation limitation of taking whole brain volume as input.

We conclude that our direct WM tract segmentation method has very good reproducibility and comparable performance to the reference standard. This is the first deep learning based method of WM tract segmentation developed on such a large-scale dataset. Our method can lead toward a faster, more lightweight way of diffusion measures analysis, thereby, reducing the time-consuming of segmentation, the complexity of pipeline setting and the required storage space.

References

1. OSullivan, M., et al.: Evidence for cortical disconnection as a mechanism of age-related cognitive decline. Neurology **57**(4), 632–638 (2001)
2. de Groot, M., et al.: Tract-specific white matter degeneration in aging: the Rotterdam study. Alzheimer's Dement **11**(3), 321–330 (2015)
3. Lawes, I.N.C., et al.: Atlas-based segmentation of white matter tracts of the human brain using diffusion tensor tractography and comparison with classical dissection. Neuroimage **39**(1), 62–79 (2008)
4. O'Donnell, L.J., Westin, C.F.: Automatic tractography segmentation using a high-dimensional white matter atlas. IEEE Trans. Med. Imaging **26**(11), 1562–1575 (2007)
5. Yendiki, A., Reuter, M., Wilkens, P., Rosas, H.D., Fischl, B.: Joint reconstruction of white-matter pathways from longitudinal diffusion MRI data with anatomical priors. Neuroimage **127**, 277–286 (2016)
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

7. Milletari, F., et al.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), pp. 565–571. IEEE (2016)
8. Wasserthal, J., et al.: Direct white matter bundle segmentation using stacked u-nets. arXiv preprint arXiv:1703.02036 (2017)
9. Hofman, A., et al.: The Rotterdam study: 2016 objectives and design update. Eur. J. Epidemiol. **30**(8), 661–708 (2015)
10. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2010)
11. Leemans, A., et al.: Exploredti: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. Int. Soc. Mag. Reson. Med. **209**, 35–37 (2009)
12. Jenkinson, M., et al.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage **17**(2), 825–841 (2002)
13. Dozat, T.: Incorporating nesterov momentum into adam (2016)
14. Choi, S.S., Cha, S.H., Tappert, C.C.: A survey of binary similarity and distance measures. J. Syst. Cybern. Inf. **8**(1), 43–48 (2010)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics pp. 159–174 (1977)