Automatic Segmentation of Brain Structures for Treatment Planning Optimization and Target Volume Definition

Marco Langhans¹,²,³(✉), Tobias Fechter¹,², Dimos Baltas¹,², Harald Binder⁴, and Thomas Bortfeld¹,⁵

¹ Division of Medical Physics, Department of Radiation Oncology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg im Breisgau, Germany
marco.langhans@mail.de
² German Cancer Consortium (DKTK), Partner Site Freiburg, Freiburg im Breisgau, Germany
³ Department of Radiation Oncology, Radiologie Vechta, Vechta, Germany
⁴ Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau, Germany
⁵ Division of Radiation Biophysics, Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

Abstract. The MICCAI Challenge 2020 “Anatomical Brain Barriers to Cancer Spread: Segmentation from CT and MR Image” was about segmenting brain structures automatically for further use in the definition of the Clinical Target Volume (CTV) of glioblastoma patients and treatment planning optimization in radiation therapy. This paper describes the methods of the team “FREI”. A 3D U-Net style deep learning network was used to achieve human-like segmentation accuracy for most of the structures within seconds.

Keywords: Medical image segmentation · Brain structures · U-Net · Deep learning · MICCAI

1 Introduction

The MICCAI Challenge 2020 “Anatomical Brain Barriers to Cancer Spread: Segmentation from CT and MR Images” consists of two tasks. The first one asks for the automatic segmentation of brain structures of glioblastoma patients for further use in Clinical Target Volume (CTV) definition. The CTV includes a margin of up to 2 cm around the Gross Tumor Volume (GTV), which is the visible part of the tumor. The margin takes the potential presence of non-visible tumor cells into account; it excludes certain structures that are impenetrable...
for tumor cells. Recently [1] described an automated method to define the CTV using an expansion model taking into account brain structures as anatomical barriers. Those barrier structures are falx cerebri, tentorium cerebelli, sagittal and transverse brain sinuses, cerebellum and ventricles. So far these structures have to be segmented manually which is time-consuming and also leads to a high inter-user variability.

Task 2 strives to segment structures (brainstem, eyes, optical nerves, chiasm, lacrimal glands and cochleas) for treatment planning optimization. These structures are so-called organs-at-risk (OAR) that need to be spared from receiving high levels of radiation dose to reduce side effects.

2 Materials and Methods

2.1 Pre-processing

**Data Overview.** The organizers of the MICCAI 2020 challenge provided a data set from 45 patients for training. Data from 15 additional patients were released as a test set to determine the leaderboard. For the final leaderboard data from another 15 patients were released. For all patients, imaging data consisting of one Computer Tomography (CT) and two Magnetic Resonance Imaging (MRI) scans were provided by the organizers. The two MRI datasets included one T1 and one T2 weighted scan. The labels for Task 1 are cerebellum, falx, sinuses, tentorium and ventricles. They were defined using the T1 weighted MRI.

For Task 2 the labels were defined on the CT scan and consist of the following structures: brainstem, chiasm, cochlea (left), cochlea (right), eye (left), eye (right), lacrimal (left), lacrimal (right), optical nerve (left) and optical nerve (right). The test sets 1 and 2 did not contain any labelmaps.

The available data showed registration errors between CT and MRI and also between the T1 and T2 images. Due to the specifics of how the structures were defined, the CT did not add useful information for Task 1 and the MRI did not add useful information for Task 2. To test if better results could be obtained after better registration of the imaging data sets, an affine registration algorithm, SimpleElastix [2,3], was applied. The multi-modality approach using CT and MRI (T1 and T2) for both tasks still created bias and resulted in lower accuracy, thus didn’t improve the outcome. In consequence, the multi-modality approach has not been pursued. A possible way to solve the registration error problem could be a deformable registration algorithm (e.g. [4]), which will be considered in future work. In this work the authors used a uni-modal approach to solve the task, thus for Task 1 the T1 weighted MRI data and Task 2 the CT was used.

**Brain Region.** To support the segmentation algorithm to distinguish between regular brain tissue and the respective label masks, an automatic brain segmentation algorithm [5] was used to create the brain mask for each patient individually. The brain mask was added to the existing label maps as a further
class. Additionally via thresholding of Hounsfield Units, the remaining patient was classified as a further separate class (body contour). With respect to Task 1 there was no gain in accuracy ascertainable, therefore both additional classes were only added to Task 2.

**Additional Pre-processing and Creation of the Training Data.** Structures belonging to paired organs (such as the left and right eye) were merged into a single structure (eye) to avoid biases during training. We found that it was easier for the network to accurately detect and segment a structure like the eye, without also specifying it as left or right eye. The body side could be easily determined in a post-processing step using the location of the structure. First attempts without using the merging approach led to a frequent misclassification of the structure’s side.

All patients were imported using SimpleITK [6], further handled using the numpy array format and cropped to the non-zero region to keep memory usage low. All image values above the 99th and below the 1st percentile were clipped to avoid biases through single pixels with very high or low values. Afterwards the MR images were normalized by Z-score normalization:

$$z = \frac{x - \mu}{\sigma}$$

(1)

with $z$: normalised intensity, $x$: intensity, $\mu$: mean value, $\sigma$: standard deviation.

Necessary metadata was saved into a separate file to maintain data like spacing, coordinates of the non-zero region, direction, origin and original shape. The metadata is needed to transfer the plain arrays back to it’s original format. There was no need to re-sample the images because they all had the same pixel size of 1.2 mm.

### 2.2 Training

**Deep Learning Architecture.** A 3D U-Net [7,8] based network (see Fig. 1) was utilized within the PyTorch framework. Each level block (yellow) consisted of a Group Normalization [9] + Parametric Rectified Linear Unit (PReLU) [10] + 3D Convolution layer (kernel size = 3, stride = 1, padding = 1) with residual connections [11]. Group Normalization was preferred to Batch Normalization because Batch Normalization turned out to be unstable for small batch sizes (one in this case). The red blocks are down- and the blue blocks upsample layers. The first hidden layer started with 24 channels. During test-time a final Softmax layer (violet) was applied.

**Optimization.** A regular dice loss function (see Eq. 2) was applied and optimized using the ADAM optimizer. The initial learning rate was 0.001, coefficients used for computing running averages of the gradient and its square were 0.9 and 0.999, respectively. No weight decay was applied.

$$DiceLoss = 1 - \frac{2 \cdot |X \cap Y|}{|X| + |Y|}$$

(2)
To estimate the performance of the model, a 5-fold cross validation was carried out. 1000 random batches were defined as one epoch. In the first step, the model was trained for 40 epochs. Afterwards the trained model was trained for an additional 40 epochs using the full data set (without validation set) and applied to Test Set 1 and 2 (each n = 15).

**Data Loader and Augmentation.** Due to the small amount of training data (n = 45), heavy data augmentation was performed using an already implemented data loader [12]. Random cropping, mirroring, scaling, rotations, elastic deformations, gamma transformations, blurring and gaussian noises were applied to the training set. During Training/Validation, patches of $128 \times 128 \times 128$ were randomly cropped from the dataset.

**Evaluation.** The organizer used two metrics to evaluate the predictions: the commonly known volumetric dice score (see Eq. 2) and the surface dice score. The surface dice score doesn’t take the whole volume into account, but the surface volume given a tolerance (2 mm in this challenge). Both metrics ranged between 0 (no overlap) and 1 (perfect overlap).

### 2.3 Prediction

To predict the final labelmap, a total of 5 patches (128, 128, 128 pixels) were fed into the network. The patches were chosen to sample the whole patient space thus resulting in four quadrants and a center one. Afterwards all labelmaps were merged into one.

### 2.4 Post-processing

The predictions of both tasks consisted of small false positive dots for structures for a few patients. The removal was done using a “keeping the largest element in an array” method. With respect to Task 2 all false positive dots were removed reliably. Task 1 consists of masks that were partially not connected (e.g. see
ventricle in Fig. 3). Thus applying the method resulted in removing true positive volumes as well. As a consequence the method was only applied to all structures of all predictions of Task 2.

3 Results

The following results have been calculated by the organizers as part of the challenge.

Figure 2(a) shows the quantitative results of Task 1. Most of the structures of Task 1 attained mean dice scores over 0.8 (except sinuses with 0.77), which indicates that the segmentation network works well. The cerebellum is perfectly segmented as all patients were close to 1.0, which shows that there is almost no variability. The surface dice scores are, as expected, even better and most of the structures achieved scores above 0.9.

Figure 2(b) shows results for Task 2. For most structures the results look very good. Structures like the cochleas and lacrimal glands are prone to error due to their relatively small size. Small shifts translate to a large error. It is also noticeable that small structures exhibit a broader spread of both the regular and surface dice score. In general, if the tolerance window is broader (e.g. surface
Fig. 3. Predicted segmentations of Task 1 projected on different slices (transversal a-e and coronal f) of the T1 MRI scan. Cerebellum (red), ventricles (cyan), falx cerebri (green), tentorium cerebelli (yellow), sinuses (blue) (Color figure online)
the metric strongly increases. Again, structures like the brainstem, similar to the cerebellum in Task 1, are clearly definable due to their sharp demarcation to other tissue. It is interesting that also the chiasm, which is, especially using only the CT, difficult to segment manually, provides mean dice scores of 0.9. Qualitative examples are shown in Fig. 3 and 4.

Fig. 4. Predicted segmentations of Task 2 projected on two slices of the CT-scan. Brainstem (red), eyes (magenta, cyan), cochleas (yellow, blue), lacrimal glands (dark blue, white), optical nerves (light and dark brown), chiasm (green) (Color figure online)

4 Discussion and Conclusion

In this work it has been shown that deep neural networks can segment various brain structures in an accurate way. Most of the structures attained volume dice scores of above 0.8, which is approximately the inter-user variability
([14] showed this for comparable brain structures). In other words: the networks achieve human like accuracy.

It is interesting that the dice scores are not symmetrical (with respect to Task 2). One would expect that the left and right hand side of all paired organs are segmented with similar performance, which is not the case. As these structures are also the smallest ones, it leads to the assumption that even slight deviations lead to a large error. Additional investigation will be done in future work.

Further structures like the left eye or the optical nerves, which are clearly appreciable visually, are inferior to structures like the chiasm, which are difficult to identify visually. With respect to the left eye, the surface dice is closer to its right hand side counterpart, which indicates that the geometry itself is captured well, but is segmented slightly too large or small, with respect to the ground truth. Consider a small structure (e.g. a sphere), which has the same shape of the ground truth, but the radius is 2 mm less. Evaluation using the surface dice score would result in a perfect overlap (if a tolerance of 2 mm is given). Although the evaluation using the volumetric dice score would result in a larger error if the volume is small enough.

In this work a uni-modal approach was performed. The structures of Task 2, which are further used for treatment optimization, could be easily derived using the treatment planning CT. No complex registration involving MRI data is needed. In addition, it is more accurate to use the definition of organs at risk on the planning CT, as the calculation of the dose distribution is also done on the CT. Shifts of contours would lead to an under- or overestimation of dose in organs at risk and could even harm the patient. However, the performance may be increased if a deformable registration between CT and MRI is applied. Especially structures that don’t show a sharp demarcation in CT, but are appreciable in MRI, may benefit.

Advantages over manual segmentation include a more objective assessment of treatment outcomes, especially with respect to adverse side effects. Because those side effects can dependent on how (correctly or incorrectly) the organs were contoured for treatment planning, a standardized procedure provides more consistency. Also the manual segmentation of a structure set, with various organs at risk, is a time-consuming process. Using the automatic approach the structures are segmented almost instantaneously (within a few seconds).

A limitation of the study is the dependency on the ground truth, as the results strongly depend on how the ground truth is defined. If there are raters providing different styles of how to segment a certain structure it is difficult for the network to learn.

As the split of the merging approach is done using the central axis of the patient it could provide incorrect segmentations for patients showing a deformed brain. Especially with respect to brain tumor patients the presence of deformations is not uncommon.
References

1. Shusharina, N., Soederberg, J., Edmunds, D., Löfman, F., Shih, H., Bortfeld, T.: Automated delineation of the clinical target volume using anatomically constrained 3D expansion of the gross tumor volume. Radiother. Oncol. 146, 37–43 (2020). https://doi.org/10.1016/j.radonc.2020.01.028

2. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Phium, J.P.W.: elastix: a toolbox for intensity based medical image registration. IEEE Trans. Med. Imaging 29(1), 196–205 (2010)

3. Shamonin, D.P., Bron, E.E., Lelieveldt, B.P.F., Smits, M., Klein, S., Staring, M.: Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer’s disease. Front. Neuroinform. 7(50), 1–15 (2014)

4. Balakrishnan, G., Zhao, A., Sabuncu, M., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging 38, 1788–1800 (2019)

5. Kleesiek, J., Urban, G., Hubert, A., et al.: Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. Neuroimage 2016(129), 460–469 (2016)

6. Lowekamp, B.C., Chen, D.T., Ibáñez, L., Blezek, D.: The design of simpleITK. Front. Neuroinform. 7, 45 (2013)

7. 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

8. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49

9. Wu, Y., Kaiming H.: Group normalization. arXiv.org/abs/1803.08494, 11 June 2018

10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, pp. 1026–1034 (2015). https://doi.org/10.1109/ICCV.2015.123

11. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual U-Net. IEEE Geosci. Remote Sens. Lett. 15(5), 749–753 (2018). https://doi.org/10.1109/LGRS.2018.2802944

12. Isensee, F., et al.: batchgenerators - a python framework for data augmentation (2020). https://doi.org/10.5281/zenodo.3632567

13. Iqbal, H.: HarisIqbal88/PlotNeuralNet v1.0.0. (2018)

14. Liedlgruber, M., et al.: Variability issues in automated hippocampal segmentation: a study on out-of-the-box software and multi-rater ground truth. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), Dublin, pp. 191–196 (2016)