Evaluating the Impact of Image Normalization on Automatic Segmentation of Brain Regions

Thomas Buchegger *University of Bern,* Carolina Duran *University of Bern,* Stefan Weber, *University of Bern*

***Abstract*—This paper addresses the question of whether image normalization leads to better results in anatomical brain region segmentation. To answer this question, we used a random forest classifier and compared six different normalization methods with each other and the ground truth segmentation. The evaluation of the segmentation performance of our algorithm on T1- and T2-weighted MRI volumes shows no significant improvement compared to no normalization.**

January 03, 2021

I. Introduction

Every year, an estimated 13.8 million patients worldwide require neurological surgery. The majority of neurosurgical care includes traumatic brain injuries, tumors, stroke-related conditions and epilepsy. [1]

To promote and facilitate these treatments, well-designed treatment planning is required. For assessment of the best treatment, Imaging Magnetic Resonance Imaging (MRI) is widely used. Post-processes of clinical diagnosis images for treatment planning often include manual segmentation of brain regions. Unfortunately, it is time-consuming to segment and label segmentation of brain structures of the large amount of data produced by MRI manually . Furthermore, manual segmentation is affected by user variability and prone to limiting the standardisation. Thus, an automatic and reliable segmentation approach is highly recommended, desirable and will be the next evolutional step. [2] [3]

Different automatic segmentation approaches are already well used, for example, the Convolutional Neural Network [2]. Nevertheless, it is still not thoroughly analysed if the normalization step in the automatic segmentation of anatomical brain regions influences the segmentation process.

In this paper, we propose a random forest for automatic segmentation that segments and labels the five different anatomical brain regions; the thalamus, the white and grey matter, the amygdala and the hippocampus. We will explore if the normalization step in the automatic segmentation will have a significant influence.

To combine these aspects, we hypothesize that normalization has no important influence in the segmentation and labelling process of the five anatomical brain regions mentioned above. We present in this paper the acquired results of no normalization and of applying six normalization methods. We then compare the results between each segmented brain region and also compare all brain regions together with no normalization. Furthermore, we analyse the results and conclude our findings.

All authors contributed equally. Biomedical Engineering, University of Berne in Switzerland. Author's E-Mails:
thomas.buchegger@students.unibe.ch
carolina.duran@students.unibe.ch
stefan.weber1@students.unibe.ch

II. Materials and Methods

A. Medical Image Analysis pipeline

The Medical Image Analysis (MIA) pipeline taught in the MIA Laboratory lectures follows the sequence of firstly perform *Registration* to T1- and T2-weighted images (T1w and T2w images), then *Preprocess*-methods. Additionally, *Feature Extraction* followed by the *Classification* of the images is performed. At the end *Post-processing*-methods are applied. Eventually, the segmentation of one patient data set is achieved.

B. Medical Background

The five anatomical brain regions segmented in this paper are the thalamus, white and grey matter, amygdala and the hippocampus. All regions are visible in figure 1.

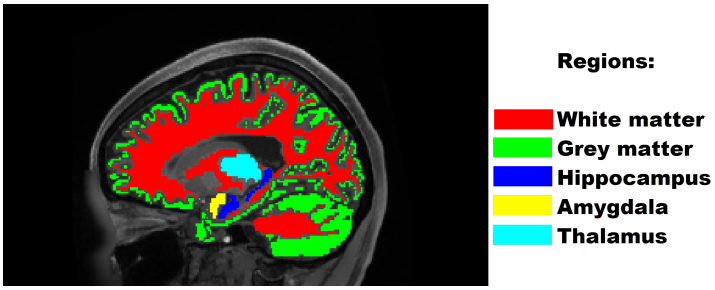

Figure 1: Ground truth picture of the five anatomical brain regions to be segmented and labelled in this project.

C. Data

The data used for this project was from the *Human Connectome Project* and has been provided by the Medical Image Analysis Lab team at the University of Bern. For anonymization, the faces and ears of the skull have been blurred. Overall, the dataset consisted of 30 MRI patient images of unrelated healthy subjects out of which 20 were used for training and 10 for testing the model. From each patient, we were given the ground truth image, the brain mask, a T1w and a T2w image. The images were generated by a 3 Tesla MRI. The ground truth has been labelled by the silver standard and with the software FreeSurfer. The atlases used are the MNI152 standard-space T1-weighted average structural template images, available at the *McConnell Brain Imaging Centre*[1] (BIC) and the *NeuroImaging & Surgical Technologies Lab*[2]. It corresponds to the MNI-ICBM atlas[2] and is derived from 152 structural images, averaged together after high-dimensional nonlinear registration into this MNI152 coordinate system. Each MRI file is of the size 118x118x217 pixels.The code was implemented in Python using scikit-learn[3] and ITK[4].

D. Registration

During *Registration*, the T1w and T2w image (floating image) is transformed with an affine transformation, such that it is similar to a given reference image. The transformation was found by an intersubjective registration from the T1w image to the provided atlas. The corresponding transformations have already been determined and are not part of this work.

E. Preprocessing

The aim of *Prerocessing* is to improve the image quality for the subsequent classification. It includes bias field correction, skull stripping, intensity normalization, histogram matching and more. Owing to the aim of this project is to analyse the impact of normalization to the realized segmentation, the used normalization methods are described in this section. For this project the brain mask for skull stripping was provided. Skull stripping was performed before a normalization method was applied to the image data. The results among all normalizations were compared to no-normalization. Primarily the used normalization methods are described. [4]

[1]URL: http://www.bic.mni.mcgill.ca/ServicesAtlases/HomePage, Date: 23.12.2020.

[2]URL: http://nist.mni.mcgill.ca/?page_id=714, Date: 23.12.2020.

[3]URL: https://scikit-learn.org/stable/, Date: 26.12.2020

[4]URL: https://itk.org/, Date: 26.12.2020

1) ZScore: While the ZScore normalization method, the mean intensity value (μ) as well as the intensity standard deviation (σ) of all pixel values of the input image are calculated. The image is normalized by subtracting μ from each pixel value and then dividing by σ. This procedure transforms the image data into an intensity distribution with a mean of 0 and a standard deviation of 1.

$$I_{New} = \frac{I - \mu}{\sigma} \tag{1}$$

2) MinMax: When applying the MinMax normalization method, the minimal intensity value of the image is subtracted from each pixel value. The result is then divided by the difference between the maximal and the minimal intensity value of the input image. This scales the intensities in a range from 0 to 1.

$$I_{New} = \frac{I - I_{min}}{I_{max} - I_{min}} \tag{2}$$

3) Whitestripe: For the Whitestripe method, equation 1 is likewise used. In contrast to the Z-Score normalization, μ and σ are obtained from the intensity values of the normal-appearing white matter (NAWM). More specifically, μ is obtained by smoothing the histogram and selecting the highest intensity peak. A 10% segment around μ corresponds to the NAWM values of the T1w image. This segment is called the whitestripe. The standard deviation is then calculated from this whitestripe values. By applying the equation 1 with the obtained μ and σ the peak of the white matter is shifted to 0 and the intensities are scaled with σ.

4) Fuzzy C-Means: By using the Fuzzy C-Means algorithm, a mask of the white matter pixel values is created. The obtained tissue mask is used to calculate the mean μ of all pixel intensity values of the white matter. Following all image intensities are scaled by μ and shifted to a constant target value c.

$$I_{New} = \frac{c \cdot I}{\mu} \tag{3}$$

5) Gaussian Mixture Model: The Guassian Mixture Model normalization method fits three Gaussian distributions to the skull stripped image intensity values. The mean μ of the Gaussian distribution of the white matter is then used to normalize the image with the same equation 3 used in the Fuzzy C-Means normalization method. c is again the target value where the mean μ is shifted to. The white matter mean μ in a T1w image is the peak with the highest intensity values. In a T2w image, the white matter mean μ is the peak with the lowest intensity values.

6) Histogram Matching: Histogram matching manipulates the histogram of the input image in such a way that the histogram of the output image matches the histogram of a given reference image. This is done by mapping the

cumulative distribution function of the input image to the reference image. The skull stripped T1w and T2w images of a subject were used as reference images.

F. Feature Extraction & Classifier

The following seven features were extracted: three coordinate features, a T1w and T2w intensity feature and a T1w and T2w gradient intensity feature.
The classifier used was a Random Forest classifier. It consists of numerous individual decision trees acting as an ensemble learning method for classification. The parameters for such a classifier are the estimator and the tree depth. The estimator indicates the maximal number of decision trees, whereas tree depth indicates the depth of each tree in the forest. Random Forest classifier tend to overfit and caution has to be given when applying. After applying a grid search, the parameters for the estimator=20 and the tree depth=190 were chosen.

G. Post-Processing

To prevent any biases or interfering in the results and thus being able analyse the influence of the different normalizations better, no post-processing methods were applied in this project.

H. Conducted experiment

To analyse the influence of the different normalizations, all parameters have been kept the same for all runs. For each run, one out of the six normalization methods has been applied. One additional run has been conducted with no normalization method. To generate reproducible results, the same random seed has been set for all runs.

I. Evaluation

To evaluate the segmentations obtained within this project the *Dice Similarity Coefficient* (DSC) as well as the *Hausdorff Distance* (HD) were applied.
DSC returns a value between 0 and 1, indicating the percentage of overall pixels of the resulted segmentation (SEG) overlapping the ground truth (GT). A result of 1 indicates a perfect segmentation. The equation states:

$$DICE(SEG, GT) = 2\frac{\mid SEG \cap GT \mid}{\mid SEG \mid + \mid GT \mid} \tag{4}$$

The HD indicates whether the margin pixels of the obtain segmentation are close to the margin pixels of the ground truth. The result is the largest distance of all pixels from one point in the segmentation to the closest point in the ground truth. A result of 0 indicates the best possible result. The equation states:

$$d_H(SEG, GT) = \max\left\{\sup_{x \in SEG}\inf_{y \in GT} d(x,y), \sup_{y \in GT}\inf_{x \in SEG} d(x,y)\right\}$$

Because of sensitivity of outlier pixels, only the lowest 95% percentile of HD values are taken into account.

III. Results

Table I presents an overview of the segmentation results for all brain regions with different normalization methods. In comparison to the ground truth segmentation, all normalization methods performed significantly worse throughout all brain regions. Comparing column per column over all normalizations, it is clear that no normalization has a significant effect on segmentation accuracy. As an example, we are looking at the white matter region. Here the mean for all DSC is ± 0.66 and the standard deviation is ± 0.04. Thus, there is clearly not much difference between the results of different normalization methods. However, the standard deviation among all DSC values have improved. Thus, all normalizations had an apparent effect on bringing σ close to zero. Worth noting is also the difference between the two metrics. A closer look at the thalamus and white matter shows the values for DSC are both similar. However, the HD for the thalamus id worse than for the white matter.

Overall, Table I shows that no normalization method performed better than the others for all brain regions together.

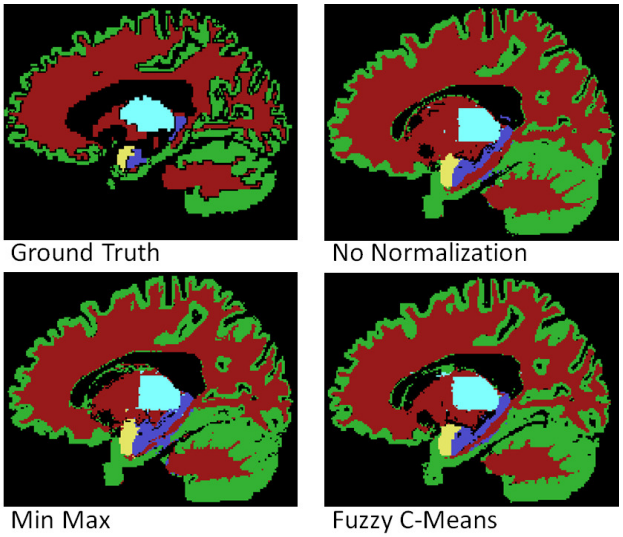

Figure 2: Comparison between the ground truth image versus no normalization (top right) and two different normalizations. All images are from the same subject and the same brain localization. In the bottom row, the resulted segmentation applying the MinMax method on the left, and the Fuzzy C-Means method on the right, are visible.

Table. I: The following table shows the results of all runs. In total there are seven runs, one without normalization and six with different normalization methods. The two metrics Dice Similarity Coefficient (DSC) and Hausdorff distance (HD) were obtained from the ten test subjects. The statistical values, mean μ and standard deviation σ, of DC and HD of the five segmented brain regions were determined.

Normalization	Metric	Thalamus	White Matter	Grey Matter	Amygdala	Hippocampus
No Normalization	DSC	μ: 0.64, σ: 0.04	μ: 0.65, σ: 0.06	μ: 0.39, σ: 0.02	μ: 0.37, σ: 0.07	μ: 0.24, σ: 0.02
	HD	μ: 24.58, σ: 15.70	μ: 4.94, σ: 0.71	μ: 5.41, σ: 1.46	μ: 18.58, σ: 4.59	μ: 22.31, σ: 3.04
ZScore	DSC	μ: 0.64, σ: 0.02	μ: 0.67, σ: 0.03	μ: 0.40, σ: 0.02	μ: 0.37, σ: 0.04	μ: 0.26, σ: 0.01
	HD	μ: 39.27, σ: 2.36	μ: 4.24, σ: 0.35	μ: 4.60, σ: 0.57	μ: 15.82, σ: 0.92	μ: 20.65, σ: 0.86
MinMax	DSC	μ: 0.66, σ: 0.03	μ: 0.65, σ: 0.03	μ: 0.39, σ: 0.02	μ: 0.37, σ: 0.04	μ: 0.24, σ: 0.02
	HD	μ: 21.04, σ: 1.73	μ: 5.53, σ: 0.31	μ: 5.04, σ: 0.92	μ: 17.45, σ: 0.72	μ: 20.43, σ: 1.63
Whitestripe	DSC	μ: 0.60, σ: 0.17	μ: 0.64, σ: 0.09	μ: 0.39, σ: 0.02	μ: 0.37, σ: 0.03	μ: 0.25, σ: 0.01
	HD	μ: 38.19, σ: 5.74	μ: 4.39, σ: 0.30	μ: 4.78, σ: 0.99	μ: 20.10, σ: 5.33	μ: 22.24, σ: 3.80
Fuzzy C-Means	DSC	μ: 0.61, σ: 0.03	μ: 0.68, σ: 0.03	μ: 0.40, σ: 0.02	μ: 0.39, σ: 0.03	μ: 0.26, σ: 0.01
	HD	μ: 41.41, σ: 0.64	μ: 4.01, σ: 0.33	μ: 4.80, σ: 0.67	μ: 17.46, σ: 2.41	μ: 20.35, σ: 1.15
Gaussian Mixture Model	DSC	μ: 0.67, σ: 0.03	μ: 0.68, σ: 0.02	μ: 0.42, σ: 0.02	μ: 0.41, σ: 0.03	μ: 0.39, σ: 0.01
	HD	μ: 38.06, σ: 1.14	μ: 4.08, σ: 0.17	μ: 3.38, σ: 0.29	μ: 17.34, σ: 3.63	μ: 17.57, σ: 1.83
Histogram Matching	DSC	μ:0.65, σ: 0.02	μ: 0.61, σ: 0.04	μ: 0.42, σ: 0.02	μ: 0.42, σ: 0.04	μ: 0.28, σ: 0.02
	HD	μ: 23.26, σ: 1.29	μ: 5.91, σ: 0.38	μ: 3.06, σ: 0.17	μ: 17.25, σ: 2.27	μ: 28.37, σ: 2.20

A selection of resulted segmentations and the ground truth segmentation of one specific subject is depicted in Figure 2. All images are taken from the same brain localization. The top left image shows the ground truth segmentation. The shape and size of the brain from the ground truth differ from the resulted images, because it has not been registered to the same atlas as the others. The more similar the resulted segmentation is to the ground truth, the better. The top right image is the segmentation with no normalization. The bottom row shows the resulted segmentation applying the MinMax method on the left, and the Fuzzy C-Means method on the right. The segmented hippocampus is more dominant in the resulted segmentation, than in the ground truth. Both segmentations look different compared to the ground truth, but look similar to the resulted normalization applying no normalization. This explains the high HDe value in Table I. Figure 2 confirms visually that there is no normalization better than the other.

IV. Discussion

Within this project, we compared six different normalization methods to answer the question of whether or not image normalization has an important influence in the segmentation and labelling process of brain regions. To facilitate brain region segmentation for treatment planning and reducing complexity and time, automatic brain region segmentation is of high importance.

Based on the obtained result, the normalization has no important influence in the segmentation and labelling process of brain regions. Nevertheless, the variance can be reduced.

Overall, no normalization method improved the segmentation significantly. We assume the data sets are already similar because they were obtained with the same MRI machine at the same hospital. Additionally, bias field correction has already been applied to all images. Another issue is that the ground truth segmentation seems not to be the best reference. In Figure 1 it is visible that the ground truth has numerous black spots where there should be a segmentation. Also, the segmentation looks irregular.

For different brain regions, different normalization methods gave slightly better results than the ground truth. However, still no normalization method gave better results for all the brain regions together.

If specific brain regions were segmented with the respective normalization that had a positive effect on them, better overall results would be obtained. Even with this approach, the segmentation results would not be significantly improved because the individual improvements are too small.

V. Conclusion

We were able to justify our hypothesis and to state that normalization has no important influence in the segmentation and labelling process of the given five anatomical brain regions.
For further steps, data sets should be obtained from different MRI machines and different hospitals and with different magnetic flux density. Additionally, the images may be in a raw condition, without any preprocess. As a further improvement, a more evolved and innovated machine learning approach could be applied - for example, a deep neural network.

References

[1] M. C. Dewan, A. Rattani, G. Fieggen, M. A. Arraez, F. Servadei, F. A. Boop, W. D. Johnson, B. C. Warf, and K. B. Park, "Global neurosurgery: the current capacity and deficit in the provision of essential neurosurgical care. executive summary of the global neurosurgery initiative at the program in global surgery and social change," *Journal of Neurosurgery JNS*, vol. 130, no. 4, pp. 1055 – 1064, 01 Apr. 2019.

[2] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[3] A. de Brébisson and G. Montana, "Deep neural networks for anatomical brain segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 20–28, 2015.

[4] J. C. Reinhold, B. E. Dewey, A. Carass, and J. L. Prince, "Evaluating the impact of intensity normalization on mr image synthesis," 2018.