

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 whether image normalization leads to better results in 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 a neurological surgery. The majority of neurosurgical care include traumatic brain injuries, tumors, stroke-related conditions and epilepsy. [1] To promote and facilitate this treatments, well-designed treatment planning is required. To assess 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, to segment and label by manual operator-guided segmentation the brain structures of the large amount of data produced by MRI is complicated and time-consuming. Furthermore, it 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 evolutionary step. [2] [3]

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

In this paper, we propose a randomized 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.

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

pus. We will explore if the normalization step in the automatic segmentation will have a significant influence.

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

II. MATERIALS AND METHODS

A. Medical Image Analysis pipeline

The Medical Image Analysis (MIA) pipeline taught in the MIA Lab lectures follow the sequence of firstly perform *Registration* to T1- and T2-weighted images (T1w and T2w image), then *Pre-Process*-methods. Additionally *Feature Extraction* followed by the *Classification* of the images are performed. At the end *Post-Processing*-methods are performed. Eventually, the segmentation of one patient's data set is achieved.

B. Medical Background

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

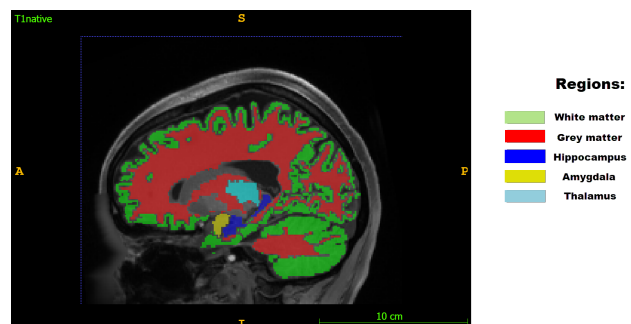


Figure 1: Ground truth picture of the five anatomical brain regions to be segmented and labelled in this project.

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 the ground truth image, the brain mask, a T1w and a T2w image were available. 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 atlas used are the MNI152 standard-space T1-weighted average structural template images available from the *McConnell Brain Imaging Centre*¹ (BIC) and the *NeuroImaging & Surgical Technologies Lab*². It corresponds to the MNI-ICBM atlas² 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-learn³ and ITK⁴.

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 *Pre-Processing* is to improve the image quality for the subsequent classification. It includes among others bias field correction, skull stripping, intensity normalization and/or histogram matching. Owing to the fact that the analysis and application of the impact of normalization to the realized segmentation, the used normalization techniques are described in this section. For this project the brain mask for skull stripping was provided. Skull stripping was performed before six different normalizations were applied to the image data. The results among all normalizations were compared to no-normalization. Primarily the used normalization methods are described. [?]

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

$$I_{New} = \frac{I - \mu}{\sigma} \quad (1)$$

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

$$I_{New} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (2)$$

3) *Whitestrife*: For the Whitestrife 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). Thus, 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 whitestrife. The standard deviation is then calculated from this whitestrife 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 of all pixel intensity values of the white matter. This mean is called μ . Following all image intensities are scaled by μ and shifted to a constant target value c .

$$I_{New} = \frac{c \cdot I}{\mu} \quad (3)$$

5) *Gaussian Mixture Model*: The Gaussian Mixture Model normalization 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 ?? as in the Fuzzy C-Means normalization method. c is again the target value where the mean μ is shifted to. The mean μ in a T1w image corresponding to the white matter 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.

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

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

³URL: <https://scikit-learn.org/stable/>, Date: 26.12.2020

⁴URL: <https://itk.org/>, Date: 26.12.2020

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. As a reference image the skull stripped T1w and T2w images of a subject were used.

F. Feature Extraction & Classifier

The following seven features were being 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 to be chosen are the estimator and the tree depth. The estimator indicates the maximal number of decision trees. The tree depth indicates the depth of each tree in the forest. This classifier tends to over-fit and caution has to be given when applying this classifier. 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 to better analyse the influence of the different normalizations, 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 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:

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

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. The equation states:

$$d_H(SEG, GT) = \max \left\{ \sup_{x \in SEG} \inf_{y \in GT} d(x, y), \sup_{y \in GT} \inf_{x \in SEG} d(x, y) \right\} \quad (5)$$

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

III. RESULTS

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	$\mu: 0.64, \sigma: 0.04$	$\mu: 0.65, \sigma: 0.06$	$\mu: 0.39, \sigma: 0.02$	$\mu: 0.37, \sigma: 0.07$	$\mu: 0.24, \sigma: 0.02$
	HD	$\mu: 24.58, \sigma: 15.70$	$\mu: 4.94, \sigma: 0.71$	$\mu: 5.41, \sigma: 1.46$	$\mu: 18.58, \sigma: 4.59$	$\mu: 22.31, \sigma: 3.04$
ZScore	DSC	$\mu: 0.64, \sigma: 0.02$	$\mu: 0.67, \sigma: 0.03$	$\mu: 0.40, \sigma: 0.02$	$\mu: 0.37, \sigma: 0.04$	$\mu: 0.26, \sigma: 0.01$
	HD	$\mu: 39.27, \sigma: 2.36$	$\mu: 4.24, \sigma: 0.35$	$\mu: 4.60, \sigma: 0.57$	$\mu: 15.82, \sigma: 0.92$	$\mu: 20.65, \sigma: 0.86$
MinMax	DSC	$\mu: 0.66, \sigma: 0.03$	$\mu: 0.65, \sigma: 0.03$	$\mu: 0.39, \sigma: 0.02$	$\mu: 0.37, \sigma: 0.04$	$\mu: 0.24, \sigma: 0.02$
	HD	$\mu: 21.04, \sigma: 1.73$	$\mu: 5.53, \sigma: 0.31$	$\mu: 5.04, \sigma: 0.92$	$\mu: 17.45, \sigma: 0.72$	$\mu: 20.43, \sigma: 1.63$
Whitestripe	DSC	$\mu: 0.60, \sigma: 0.17$	$\mu: 0.64, \sigma: 0.09$	$\mu: 0.39, \sigma: 0.02$	$\mu: 0.37, \sigma: 0.03$	$\mu: 0.25, \sigma: 0.01$
	HD	$\mu: 38.19, \sigma: 5.74$	$\mu: 4.39, \sigma: 0.30$	$\mu: 4.78, \sigma: 0.99$	$\mu: 20.10, \sigma: 5.33$	$\mu: 22.24, \sigma: 3.80$
Fuzzy C-Means	DSC	$\mu: 0.61, \sigma: 0.03$	$\mu: 0.68, \sigma: 0.03$	$\mu: 0.40, \sigma: 0.02$	$\mu: 0.39, \sigma: 0.03$	$\mu: 0.26, \sigma: 0.01$
	HD	$\mu: 41.41, \sigma: 0.64$	$\mu: 4.01, \sigma: 0.33$	$\mu: 4.80, \sigma: 0.67$	$\mu: 17.46, \sigma: 2.41$	$\mu: 20.35, \sigma: 1.15$
Gaussian Mixture Model	DSC	$\mu: 0.67, \sigma: 0.03$	$\mu: 0.68, \sigma: 0.02$	$\mu: 0.42, \sigma: 0.02$	$\mu: 0.41, \sigma: 0.03$	$\mu: 0.39, \sigma: 0.01$
	HD	$\mu: 38.06, \sigma: 1.14$	$\mu: 4.08, \sigma: 0.17$	$\mu: 3.38, \sigma: 0.29$	$\mu: 17.34, \sigma: 3.63$	$\mu: 17.57, \sigma: 1.83$
Histogram Matching	DSC	$\mu: 0.65, \sigma: 0.02$	$\mu: 0.61, \sigma: 0.04$	$\mu: 0.42, \sigma: 0.02$	$\mu: 0.42, \sigma: 0.04$	$\mu: 0.28, \sigma: 0.02$
	HD	$\mu: 23.26, \sigma: 1.29$	$\mu: 5.91, \sigma: 0.38$	$\mu: 3.06, \sigma: 0.17$	$\mu: 17.25, \sigma: 2.27$	$\mu: 28.37, \sigma: 2.20$

Table 1 presents an overview of the segmentation results for all brain regions with the different normalization methods. In comparison to the ground-truth segmentation, all normalizations performed significantly worse throughout all brain regions. A perfect score for DSC would be $\mu = 1$ and $\sigma = 0$ and for Hausdorff Distance $\mu = 0$ and $\sigma = 0$. Comparing column per column over all normalizations, it is clear that no normalization has a significant effect on segmentation accuracy. For White Matter, for example, the mean for all DCs is ± 0.65 and the standard deviation is ± 0.05 , so clearly not much difference. What improved, however, is the standard deviation among all dice coefficient values. So all normalizations had an apparent effect on bringing σ really close to zero. Also worth noting is the difference between the two metrics. A closer look on thalamus and white matter shows the values for DSC are both close, the Hausdorff Distance, however, is for thalamus much worse than for white matter.

Overall, Table 1 clearly shows that no normalization method performed better than the others and it also does not matter which brain region is segmented as the results are all similar.

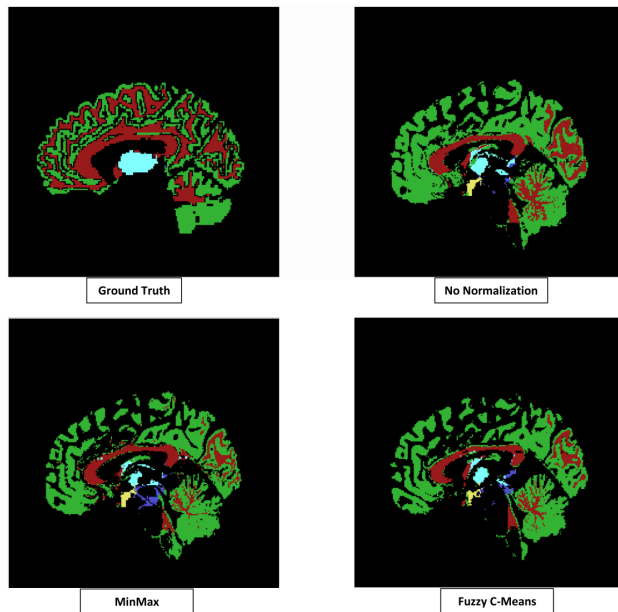


Figure 2: Comparison of no ground truth vs. different normalizations

A selection of results is depicted in Figure 2. Top left shows the ground-truth segmentation, so as similar our normalization look, the better. Top right is the segmentation with no normalization and it is obvious that the result is much worse. The bottom row with Min-Max and Fuzzy C-Means normalization look completely

different than the ground truth as well, yet similar to no normalization. One obvious difference is the lack of thalamus, which explains the very high hausdorff distance value in Table 1. As already mentioned that Table 1 shows that there is no normalization better than the other, Figure 2 confirms this visually.

IV. DISCUSSION

Aims of the Discussion part:

- Highlight importance of your work (highlight novelty /impact etc
- To interpret your results in relation to your original problem
- To put your work into the context of existing work
- To present any limitations of the presented work
- To make future recommendations
- To provide a conclusion of the work

1) Importance of the work

- a) Summarise your results
- b) Reiterate the importance of the work (novelty , impact etc)

2) Interpretation of results

- a) Interpret your results focussing on the problem described in the introduction. What do the results mean for the described problem?
- b) Explain any unusual/important findings (be careful if not your original investigative subject)

3) Provide context

- a) Describe your results in relation to others and try to explain any discrepancies
- b) Emphasize how your results support or refute your hypotheses current thinking in the field. Were results as expected? If not why and what does this mean?

4) Limitations of your work

- a) Describe any limitations /deficiencies of your work and what impact they have on the findings
- b) Suggest possible future solutions

V. CONCLUSION

- Summarise your findings and relate your findings back to your hypothesis / aim / objective and to your problem.
- Based on your findings, suggest next steps towards solving your problem

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.