ilastik for Multi-modal Brain Tumor Segmentation

Jens Kleesiek[1,2,3], Armin Biller[1,3], Gregor Urban[2], Ullrich Köthe[2], Martin Bendszus[1], and Fred A. Hamprecht[2]

[1] Division of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany
[2] Heidelberg University HCI/IWR, Heidelberg, Germany
[3] Division of Radiology, German Cancer Research Center, Heidelberg, Germany
kleesiek@uni-heidelberg.de

Abstract. We present the application of *ilastik*, the open source interactive learning and segmentation toolkit, for brain tumor segmentation in multi-modal magnetic resonance images. Even without utilizing the interactive nature of the toolkit, we are able to achieve Dice scores comparable to human inter-rater variability and are ranked in the top-5 results for the BraTS 2013 challenge data set, where no ground truth is publicly available. As careful intensity calibration is crucial for discriminative models, we propose a cerebrospinal fluid (CSF) normalization technique for pre-processing, which appears to be robust and effective. Further, we evaluate different post-processing methods for the random forest (RF) predictions obtained with *ilastik*.

Keywords: Multi-modal MRI, Brain tumor segmentation, BraTS challenge

1 Introduction

Segmenting brain tumors from multi-modal imaging data is a very challenging medical image analysis task due to the fact that magnetic resonance imaging (MRI) is usually not quantitative and lesion areas are mostly defined through intensity changes relative to surrounding normal tissue. Furthermore, the task is complicated by partial volume effects and various artifacts, e.g. due to the inhomogeneities of the magnetic field or motion of the patient during the examination. Hence, it is not surprising that even manual segmentations by experts exhibit significant intra- and inter-rater variability, which is estimated to be up to 20 % and 28 %, respectively [8].

The state-of-the-art brain tumor segmentation methods can roughly be divided in discriminative and generative approaches. For a comprehensive recent overview please see Menze et al. [9]. In general, the task of a discriminative method is to perform a tissue classification of unseen data, based on the raw data and voxel-wise or regionally extracted features. For training, supervised approaches usually rely on labels that were assigned by human expert raters and are considered to resemble ground truth. In the current study, we mostly

follow this canonical approach, but introduce important variations during pre- and post-processing (see Sec. 2). The core of the proposed segmentation pipeline is *ilastik*[4] that allows predictions in close to real time [10]. The generic framework of *ilastik* has been used successfully in different domains, e.g [6, 7]. Instead of exploiting the intended usage of *ilastik*, i.e. interactive machine learning via a convenient graphical user interface, we non-interactively generate project files with random labels drawn from the annotated training data and then use the pixel classification workflow in batch prediction mode for training and prediction. The pixel classification workflow is based on a random forest (RF) classifier [3]. Although possible, user interaction beyond pre-recorded groundtruth- and CSF-labeling (see below) is not required. The proposed pipeline achieves accuracies comparable to human raters and, at the time of writing, is ranked in the top-5 of all submitted results for the BraTS 2013 challenge data set.

In this workshop paper we elucidate the proposed method in detail (Sec. 2), report (Sec. 3) and discuss (Sec. 4) the results achieved for the BraTS 2013 training and challenge data set [9].

2 Materials and Methods

2.1 Data

We use the BraTS 2013 training and challenge data set provided via the Virtual Skeleton Database (VSD) [5]. The synthetic data was excluded, because it i) was not evaluated in the 2013 challenge and ii) the synthetic data sets "are less variable in intensity and less artifact-loaded than real images" [9].

The data stems from MR scanners of different vendors and with different field strengths. It comprises co-registered native and contrast enhanced T1-weighted images, as well as T2-weighted and T2-FLAIR images. The images contain low grade (LG) and high grade (HG) tumors. For a detailed description please see Menze et al. [9].

2.2 Pre-processing

The pre-processing comprises two steps. First we employ histogram normalization as implemented by the *HistogramMatching* routine of 3D-Slicer[5]. As reference images we used the four different modalities of an arbitrary data set (HG0001). To exclude the background during matching, all voxels whose grayscale values were smaller than the mean grayscale value were excluded. Next, we normalized each individual modality with the mean value of the CSF. To obtain these values we interactively trained *ilastik* with ten randomly chosen data sets from the training set. This two class classification (CSF vs. rest) is a fairly easy task, because CSF exhibits an unambiguous combination of intensity values in the multi-modal images (dark in T1, T1c and FLAIR but bright in T2). The effect of this proposed two-step normalization technique can be seen in Fig. 1.

[4] `https://github.com/ilastik`
[5] `http://www.slicer.org`

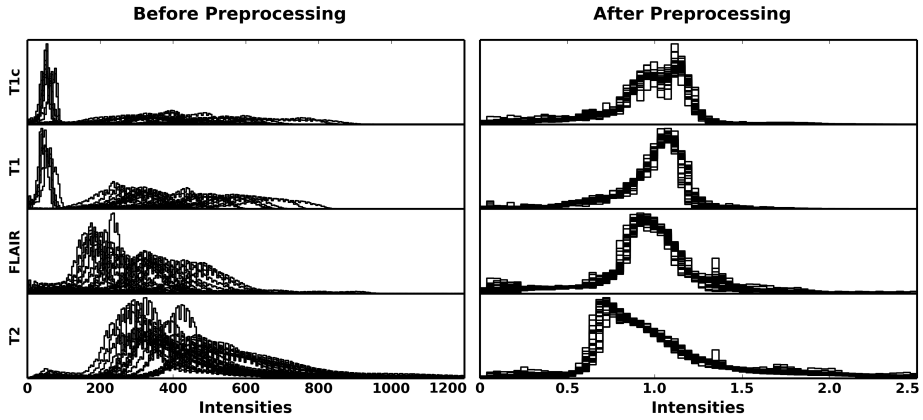

Fig. 1. Effect of the proposed two-step normalization technique. On the left side histograms of the raw intensity values of the BraTS 2013 training set (LG and HG, $N = 30$) are plotted separately for each modality. The right side shows the histograms after normalization with CSF.

After normalization we augmented the four base sequences by subtracting each modality from every other. In combination with the original four images this yields a stack of ten volumes that consecutively are used for voxel-wise feature computation. For each channel we calculated the Laplacian of Gaussian (scale 1.0), the structure tensor eigenvalues (scale 1.6) and the Hessian of Gaussian eigenvalues (scale 1.6), as implemented in the *ilastik* feature selection applet.

2.3 Pixel Classification

The *ilastik* project consists of three core software libraries: *volumina*, *lazyflow* and *ilastik*. *Lazyflow* provides threading utilities for distributing concurrent workloads across multiple cores. To achieve close to real time computations in interactive mode, this library ensures, that only computations are preformed that are strictly required to produce an output for the actually displayed data. Visualization of the multi-dimensional data, that possibly can be larger than RAM, is realized with *volumina*. These two frameworks are then orchestrated to an integrated software tool via the *ilastik* library.

Pixel classification is one of the available workflows. It relies on ten random forests with 10 trees each that are trained in parallel and eventually are merged into a single forest. Gini impurity is used as a split criterion and the number of randomly chosen features at each split is proportional to the square root of the total number of features.

To use *ilastik* in an automatic fashion, we created project files off-line. For each of the four tumor classes (edema, enhancing, non-enhancing and necrosis) up to 200 training samples, i.e. multi-dimensional feature vectors, were randomly chosen from the provided ground truth labels of every training data set. Another

1000 random samples were taken from the normal tissue of each training data set. Further, we introduced 'air' as an additional class that was granted an additional 20 labels. Different classifiers were trained for LG and HG tumors.

2.4 Post-processing

For post-processing we evaluated different strategies with increasing computational costs. In the simplest case we use simple Gaussian smoothing to clean-up the RF predictions. A more sophisticated approach relies on a guided filter as proposed by He et al. [4]. This is an edge-preserving filter that does not suffer from gradient reversal artifacts as for instance a bilateral filter and it can be computed in linear time. We also employ graph-cut optimization via the α-expansion algorithm [2] to adjust the labels. For this purpose we transformed the pseudo-probabilities P of the RF into unary potentials:

$$U(\mathbf{x}) = -\log(P(\mathbf{x}))\ . \tag{1}$$

If the labels of two variables differ we assign a cost of $c = 0.4$. The computations are realized with the *OpenGM* library [1].

A common downstream processing of the labels consists of identifying connected components (CC) and discarding all those that are < 3000 voxels. This is realized with the VIGRA library[6].

2.5 Evaluation of the Results

For comparison of the predicted segmentations we computed different standard measures, with an emphasize on the Dice coefficient as suggested in Menze et al. [9]. This metric characterizes the voxel-wise overlap of two segmented regions, by normalizing the number of true positives with the average size of the two regions. To evaluate the performance on the BraTS 2013 training data we performed leave-one-out cross-validation (LOO-CV) and used the Comparison and Validation of Image Computing (COVALIC) toolkit[7] to obtain the comparison metrics. This toolkit is also used by the challenge organizers for the evaluation. The challenge data, for which no ground truth is publicly available, was evaluated through the challenge website[8].

3 Results

Results for the LOO-CV of the training data are summarized in Tab. 1, for the challenge data in Tab. 2. For a description of the different post-processing methods please see Sec. 2.4.

[6] `https://github.com/ukoethe/vigra`
[7] `https://github.com/InsightSoftwareConsortium/covalic`
[8] `http://www.virtualskeleton.ch`

Table 1. Dice scores for BratTS 2013 training data with LOO-CV

Method	whole	LG/HG	core	LG/HG	active
Human Rater [9]	85	84/88	75	67/93	74
ilastik	75	73/76	60	58/61	65
ilastik + CC	80	78/81	64	60/66	69
ilastik + Gaussian Smoothing + CC	84	82/84	68	61/71	72
ilastik + Guided Filter + CC	83	81/84	68	61/72	71
ilastik + OpenGM + CC	83	81/84	67	61/70	72

Table 2. Dice scores for BratTS 2013 challenge data (only HG)

Method	whole	core	active
Best 2013	87	78	74
Current Best	**92**	**79**	**76**
ilastik + OpenGM + CC	87	76	74

4 Discussion

Our results (Tab. 2) on the 2013 challenge data set are comparable to the inter-rater variability reported for the BraTS data [9]. At the time of writing they are ranked in the top-5 of all submitted results. On the training data we perform slightly worse (rank 7). This might be explained by the fact that we omitted the synthetic data, for which higher Dice scores were reached as for similar real data [9].

In contrast to most methods reported in Menze et al. [9], we do not perform a bias field correction with N4ITK [11] during pre-processing, because it did not improve our result on the training data. Instead, we propose to perform intensity normalization with the mean CSF value, which proved to be a robust and effective technique (Fig. 1).

The evaluation of the different post-processing methods on the training set with LOO-CV (Tab. 2) shows the added value of "cleaning-up" the RF predictions. The three different methods used, exhibit a similar performance but come at different computational costs. Especially, simple Gaussian smoothing is a fast and effective method.

Looking at our segmentations in detail, we noticed the presence of 'holes', which –according to our predictions– correspond to islands of healthy neuronal tissue. From a neuro-oncological point of view this is plausible and can not be ruled out per se. However, due to the labeling instructions for the experts [9], it is not very likely that those kind of islands occur in the ground truth data. Primarily aiming at an interactive clinical workflow, we decided not to fill these holes with a computational method, which supposedly would improve our challenge results further.

Future work aims at integrating the insights obtained during the challenge into an *ilastik* workflow that can be easily deployed in clinical routine and for clinical trials.

Acknowledgments. We thank Thorsten Beier, Christoph Decker, Markus Döring, Burçin Eröcal, Carsten Haubold and the entire ilastik team for technical help and valuable comments. This work was supported by a postdoctoral fellowship from the Medical Faculty of the University of Heidelberg.

References

1. Andres, B., Beier, T., Kappes, J.H.: OpenGM: A C++ library for discrete graphical models. ArXiv e-prints (2012), `http://arxiv.org/abs/1206.0111`
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23(11), 1222–1239 (Nov 2001), `http://dx.doi.org/10.1109/34.969114`
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
4. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Trans Pattern Anal Mach Intell 35(6), 1397–409 (Jun 2013)
5. Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P.: The virtual skeleton database: an open access repository for biomedical research and collaboration. J Med Internet Res 15(11), e245 (2013)
6. Kreshuk, A., Koethe, U., Pax, E., Bock, D.D., Hamprecht, F.A.: Automated detection of synapses in serial section transmission electron microscopy image stacks. PLoS One 9(2), e87351 (2014)
7. Kroeger, T., Mikula, S., Denk, W., Koethe, U., Hamprecht, F.A.: Learning to segment neurons with non-local quality measures. Med Image Comput Comput Assist Interv 16(Pt 2), 419–27 (2013)
8. Mazzara, G.P., Velthuizen, R.P., Pearlman, J.L., Greenberg, H.M., Wagner, H.: Brain tumor target volume determination for radiation treatment planning through automated mri segmentation. Int J Radiat Oncol Biol Phys 59(1), 300–12 (May 2004)
9. Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), submitted to IEEE Transactions on Medical Imaging
10. Sommer, C., Straehle, C., Koethe, U., Hamprecht, F.A.: "ilastik: Interactive learning and segmentation toolkit". In: 8th IEEE International Symposium on Biomedical Imaging (ISBI) (2011)
11. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: improved n3 bias correction. IEEE Trans Med Imaging 29(6), 1310–20 (Jun 2010)