This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

A generative probabilistic model and discriminative extensions for brain lesion segmentation – with application to tumor and stroke

Bjoern H. Menze[1,2,3,4,5], Koen Van Leemput[6,7], Danial Lashkari[1], Tammy Riklin-Raviv[8], Ezequiel Geremia[2], Esther Alberts[4,5], Philipp Gruber[9], Susanne Wegener[9], Marc-André Weber[10], Gabor Székely[3], Nicholas Ayache[2], and Polina Golland[1]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA
[2]Asclepios Research Project, INRIA Sophia-Antipolis, France
[3]Computer Vision Laboratory, ETH Zurich, Switzerland
[4]Institute for Advanced Study, TU München, Munich, Germany
[5]Department of Computer Science, TU München, Munich, Germany
[6]Department of Radiology, Massachusetts General Hospital, Harvard Medical School, USA
[7]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[8]Electrical and Computer Engineering Department, Ben Gurion University, Beer-Sheva, Israel
[9]Department of Neurology, University Hospital, Zurich, Switzerland
[10]Department of Diagnostic Radiology, Heidelberg University Hospital, Germany

***Abstract*—We introduce a generative probabilistic model for segmentation of brain lesions in multi-dimensional images that generalizes the *EM segmenter*, a common approach for modelling brain images using Gaussian mixtures and a probabilistic tissue atlas that employs expectation-maximization (EM) to estimate the label map for a new image. Our model augments the probabilistic atlas of the healthy tissues with a latent atlas of the lesion. We derive an estimation algorithm with closed-form EM update equations. The method extracts a latent atlas prior distribution and the lesion posterior distributions jointly from the image data. It delineates lesion areas individually in each channel, allowing for differences in lesion appearance across modalities, an important feature of many brain tumor imaging sequences. We also propose discriminative model extensions to map the output of the generative model to arbitrary labels with semantic and biological meaning, such as "tumor core" or "fluid-filled structure", but without a one-to-one correspondence to the hypo- or hyper-intense lesion areas identified by the generative model.**

We test the approach in two image sets: the publicly available BRATS set of glioma patient scans, and multimodal brain images of patients with acute and subacute ischemic stroke. We find the generative model that has been designed for tumor lesions to generalize well to stroke images, and the generative-discriminative model to be one of the top ranking methods in the BRATS evaluation.

I. Introduction

Gliomas are the most frequent primary brain tumors. They originate from glial cells and grow by infiltrating the surrounding tissue. The more aggressive form of this disease is referred to as "high-grade" glioma. The tumor grows fast and patients often have survival times of two years or less, calling for immediate treatment after diagnosis. The slower growing "low-grade" disease comes with a life expectancy of five years or more, allowing the aggressive treatment to be delayed. Extensive neuroimaging protocols are used before and after treatment, mapping different tissue contrasts to evaluate the progression of the disease or the success of a chosen treatment strategy. As evaluations are often repeated every few months, large longitudinal datasets with multiple modalities are generated for these patients even in routine clinical practice. In spite of the need for accurate information to guide decision making process for an treatment, these image series are primarily evaluated using qualitative criteria – indicating, for example, the presence of characteristic hyper-intense intensity changes in contrast-enhanced T1 MRI – or relying on quantitative measures that are as basic as calculating the largest tumor diameter that can be recorded in a set of axial images.

While an automated and reproducible quantification of tumor structures in multimodal 3D and 4D volumes is highly desirable, it remains difficult. Glioma is an infiltratively growing tumor with diffuse boundaries and lesion areas are only defined through intensity changes *relative* to surrounding normal tissues. As a consequence, the outlines of tumor structures cannot be easily delineated – even manual segmentations by expert raters show a significant variability [1] – and common MR intensity normalization strategies fail in the presence of extended lesions. Tumor structures show a significant amount of variation in size, shape, and localization, precluding the use of related mathematical priors. Moreover, the mass effect induced by the growing lesion may lead to displacements of the normal brain tissues, as well as a resection cavity that is present after treatment, limits the reliability of prior knowledge available for the healthy parts of the brain. Finally, a large variety of imaging modalities can be used for mapping tumor-related tissue changes, providing different types of biological information, such as differences in tissue water (T_2-MRI, FLARI-MRI), enhancement of contrast agents (post-Gadolinium T_1-MRI), water diffusion (DTI), blood perfusion (ASL-, DSC-, DCE-MRI), or relative concentrations of se-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

lected metabolites (MRSI). A segmentation algorithm must adjust to any of these, without having to collect large training sets, a common limitation for many data-driven learning methods.

Related Prior Work

Brain tumor segmentation has been the focus of recent research, most of which is dealing with glioma [2], [3]. Few methods have been developed for less frequent and less aggressive tumors [4], [5], [6], [7]. Tumor segmentation methods often borrow ideas from other brain tissue and other brain lesion segmentation methods that have achieved a considerable accuracy [8]. Brain lesions resulting from traumatic brain injuries [9], [10] and stroke [11], [12] are similar to glioma lesions in terms of size and multimodal intensity patterns, but have attracted little attention so far. Most automated algorithms for brain lesion segmentation rely on either generative or discriminative probabilistic models at the core of their processing pipeline. Many encode prior knowledge about spatial regularity and tumor structures, and some offer longitudinal extensions for 4D image volumes to exploit longitudinal image sets that are becoming increasingly available [13], [14].

Generative probabilistic models of spatial tissue distribution and appearance have enjoyed popularity for tissue classification as they exhibit good generalization performance [15], [16], [17]. Encoding spatial prior knowledge for a lesion, however, is challenging. Tumors may be modeled as outliers relative to the expected shape [18], [19] or to the image signal of healthy tissues [16], [20]. In [16], for example, a criterion for detecting outliers is used to generate a tumor prior in a subsequent EM segmentation that treats the tumor as an additional tissue class. Alternatively, the spatial prior for the tumor can be derived from the appearance of tumor-specific markers, such as Gadolinium enhancements [21], [22], or from using tumor growth models to infer the most likely localization of tumor structures for a given set of patient images [23]. All these segmentation methods rely on registration to align images and the spatial prior. As a result, joint registration and tumor segmentation [17], [24] and joint registration and estimations of tumor displacement [25] have been investigated, as well as the direct evaluation of the deformation field for the purpose of identifying the tumor region [7], [26].

Discriminative probabilistic models directly learn the differences between the appearance of the lesion and other tissues from the data. Although they require substantial amounts of training data to be robust to artifacts and variations in intensity and shape, they have been applied successfully to tumor segmentation tasks [27], [28], [29], [30], [31]. Discriminative approaches proposed for tumor segmentation typically employ dense, voxel-wise features from anatomical maps [32] or image intensities, such as local intensity differences [33], [34] or intensity profiles, that are used as input to inference algorithms such as support vector machines [35], decision trees ensembles [32], [36], [37], or deep learning approaches [38], [39]. All methods require the imaging protocol to be exactly the same in the training set and in the novel images to be segmented. Since local intensity variation that is characteristic of MRI is not estimated during the segmentation process, as in most generative mixture models, calibration of the image intensities becomes necessary which is already a difficult task in the absence of lesions [40], [41], [42].

Advantageous properties of generative and discriminative probabilistic models have been combined for a number of applications in medical imaging: generative approaches can be used for model-driven dimensional reduction to form a low-dimensional basis for a subsequent discriminative method, for example, in whole brain classification of Alzheimer's patients [43]. Vice versa, a discriminative model may serve as a filter to constrain the search space for employing complex generative models in a subsequent step, for example, when fitting biophysical metabolic models to MRSI signals [44], or when fusing evidence across different anatomical regions in the analysis of contrast-enhancing structures [45]. Other approaches improve the output of a discriminative classification of brain scans by adding prior knowledge on the location of subcortical structures [46] or the skull shape [47] through generative models. The latter approach for skull stripping showed superior robustness in particular on images of glioma patients [48]. To the best of our knowledge no generative-discriminative model has been used for tumor analysis so far, although the advantages of employing a secondary discriminative classifier on the probabilistic output of a first level discriminative classifier, which considers prior knowledge on expected anatomical structures of the brain, has been demonstrated in [32].

Spatial regularity and spatial arrangement of the 3D tumor sub-structure is used in most generative and discriminative segmentation techniques, often in a postprocessing step and with extensions along the temporal dimension for longitudinal tasks: Local regularity of tissue labels can be encoded via boundary modeling within generative [16], [49] and discriminative methods [27], [28], [50], [49], or by using Markov random field priors [30], [31], [51]. Conditional random fields help to impose structures on the adjacency of specific labels and, hence, impose constraints on the wider spatial context of a pixel [29], [35]. 4D extensions enforce spatial contiguity along the time dimension either in an undirected fashion [52], or in a directed one when imposing monotonic growth constraints on the segmented tumor lesion acting as a non-parametric growth model [13], [53], [14]. While all these segmentation models act locally, more or less at the pixel level, other approaches consider prior knowledge about the global location of tumor structures. They learn, for example, the relative spatial arrangement of tumor structures such as tumor core, edema, or enhancing active components, through hierarchical models of super-voxel clusters [54], [34], or by relating image patterns with phenomenological tumor growth models that are adapted to the patient [25].

Contributions

In this paper we address three different aspects of multimodal brain lesion segmentation, extending preliminary work we presented earlier in [55] [56], [57]:

- We propose a new generative probabilistic model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

for channel-specific tumor segmentation in multi-dimensional images. The model shares information about the spatial location of the lesion among channels while making full use of the highly specific multimodal, i.e., multivariate, signal of the healthy tissue classes for segmenting normal tissues in the brain. In addition to the tissue type, the model includes a latent variable for each voxel encoding the probability of observing a tumor at that voxel, similar to [49], [50]. The probabilistic model formalizes qualitative biological knowledge about hyper- and hypo-intensities of lesion structures in different channels. Our approach extends the general *EM segmentation* algorithm [58], [59] using probabilistic tissue atlases [60], [15], [61] for situations when specific spatial structures cannot be described sufficiently through population priors.

- We illustrate the excellent generalization performance of the generative segmentation algorithm by applying it to MR images of patient with ischemic stroke, which – to the best of our knowledge – is one of the first automated segmentation algorithms for this major neurological disease.
- We extend the generative model to a joint generative-discriminative method that compensates for some of the shortcomings of both the generative and the discriminative modeling approach. This strategy enables us to predict clinically meaningful tumor tissue labels and not just the channel-specific hyper- and hypo-intensities returned by the generative model. The discriminative classifier uses the output of the generative model, which improves its robustness against intensity variation and changes in imaging sequences. This generative-discriminative model defines the state-of-the-art on the public BRATS benchmark data set [1].

In the following we introduce the probabilistic model (Section II), derive the segmentation algorithm and additional biological constraints, and we describe the discriminative model extensions (Section III). We evaluate the properties and performance of the generative and the generative-discriminative methods on a public glioma dataset (Sections IV and V, respectively), including an experiment on the transfer of the generative model to images from stroke patients. We conclude with a discussion of the results and of future research directions (Section VI).

II. A Generative Brain Lesion Segmentation Model

Generative models consider prior information about the structure of the observed data and exploit such information to estimate latent structure from new data. The *EM segmenter*, for example, models the image of a healthy brain through three tissue classes [60], [15], [61]. It encodes their approximate spatial distribution through a population atlas generated by aligning a larger set of reference scans, segmenting them manually, and averaging the frequency of each tissue class in a given voxel within the chosen reference frame. Moreover, it assumes that all voxels of a tissue class have about the same image intensity which is modeled through a Gaussian distribution. This segmentation method, whose parameters can be estimated very efficiently through the expectation maximization (EM) procedure, treats image intensities as nuisance parameters which makes it robust in the presence of the characteristic variability of the intensity distributions of MR images. Moreover, since the method formalizes the image content explicitly through the probabilistic model, it can be combined with other parametric transformations, for example, for registration [62] or bias field correction [15], and account for the related changes in the observed data. Generative models with tissue atlases used as spatial priors are at the heart of most advanced image segmentation models in neuroimaging [63], [64].

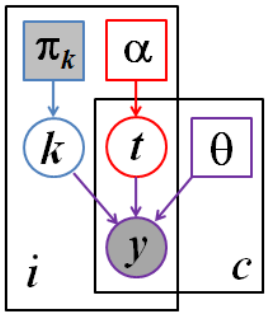

Fig. 1. Graphical model for the proposed segmentation approach. Voxels are indexed with i, channels are indexed by c. The known prior π_k determines the label k of the normal healthy tissue. The latent atlas α determines the channel-specific presence of tumor $\boldsymbol{t}$. Normal tissue label k, tumor labels $\boldsymbol{t}$, and intensity distribution parameters θ jointly determine the multimodal image observations $\boldsymbol{y}$. Observed (known) quantities are shaded. The segmentation algorithms aims to estimate $p(t_i^c|\boldsymbol{y})$, along with the segmentation of healthy tissue $p(k_i|\boldsymbol{y})$.

Population atlases cannot be generated for tumors as their location and extensions vary significantly across patients. Still, the tumor location is similar in different MR images of the *same* patient and a patient-specific atlas of the lesion class could be generated. Segmentation and atlas building can be performed simultaneously, in a joint estimation procedure [50]. Here, the key idea is to model the lesions through a separate latent atlas class. Combined with the standard population atlas of the normal tissues and the standard EM segmentation framework, this extends the *EM segmenter* to multimodal or longitudinal images of patients with a brain lesion. The generative model is illustrated in Fig. 1.

A. The Probabilistic Generative Model

Normal healthy tissue classes: We model the *normal* healthy tissue label k_i of voxel i in the healthy part of the brain using a spatially varying probabilistic atlas, or prior $p(k_i = k)$ that is constructed from prior examples. At each voxel $i \in \{1, \ldots, I\}$, this atlas indicates the probability of the tissue label k_i to be equal to tissue class $k \in \{1, \ldots, K\}$ (Fig. 1, blue). The probability of observing tissue label k at voxel i is modeled through a categorical distribution

$$p(k_i = k) = \pi_{ki}, \tag{1}$$

where $\sum_k \pi_{ki} = 1$ for all i and $\pi_{ki} \geq 0$ for all i, k. The tissue label k_i is shared among all C channels at voxel i. In our experiments we assume $K = 3$, representing gray matter (G), white matter (W) and cerebrospinal fluid (CSF), as illustrated in Fig. 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

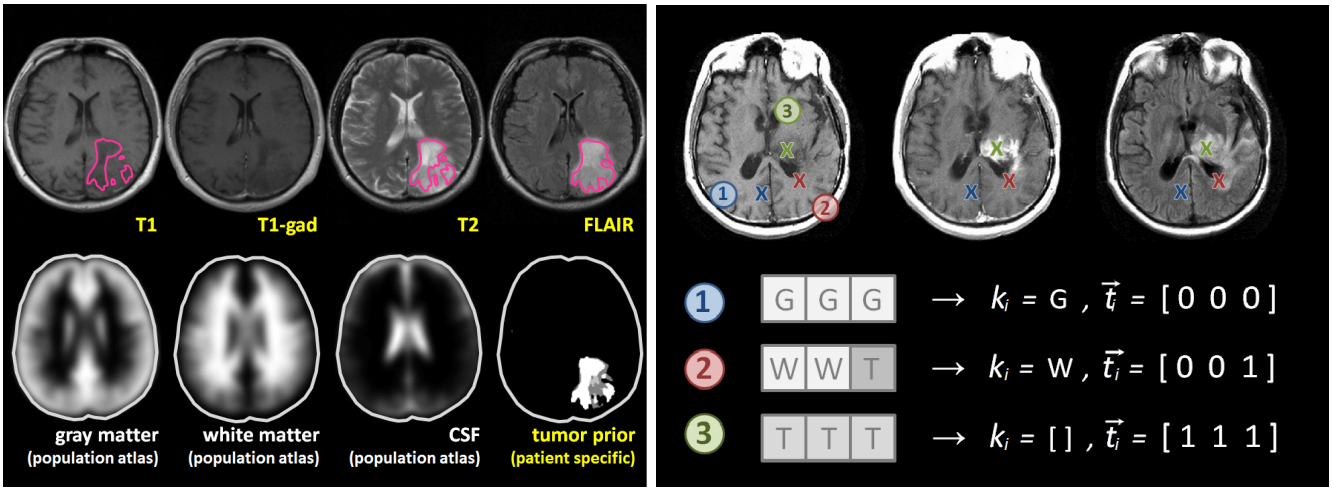

Fig. 2. Illustration of the probabilistic model. The left panel shows images of a low-grade glioma patient with lesion segmentations in the different channels (outlined in magenta); in the bottom row it shows the probabilistic tissue atlases used in the analysis, and the patient-specific tumor prior inferred from the segmentations in the different channels. The right panel shows three voxels i with different labels in T1-, T1c- and FLAIR-MRI for a high-grade glioma patient. In voxel 1 all three channels show the characteristic image intensity of gray matter (G). In voxel 2 white matter (W) is visible in the first two channels, but the third channel contains a tumor-induced change (T), here due to edema or infiltration. In voxel 3 all channels exhibit gray values characteristic of tumor: a hypo-intense signal in T1, a hyper-intense gadolinium uptake in T1c – indicating the most active regions of tumor growth – and a hyper-intense signal in the FLAIR image. The initial tissue class k_i remains unknown. Both k_i and $\boldsymbol{t}_i$ are to be estimated. Inference is done by introducing a transition process – with latent prior α_i (Fig. 1) – which is assumed to have induced the channel-specific tissue changes implied by $\boldsymbol{t}_i^c = 1$ in the tumor label vector $\boldsymbol{t}_i$.

Tumor class: We model the *tumor* label using a spatially varying latent probabilistic atlas α [49], [50] that is specific to the given patient (Fig. 1, red). At each voxel i, this atlas contains a scalar parameter α_i that defines the probability of observing a tumor at that voxel, forming the 3D parameter volume $\boldsymbol{\alpha}$. Parameter α_i is unknown and is estimated as part of the segmentation process. We define a latent tumor label $t_i^c \in \{0, 1\}$ that indicates the presence or absence of tumor-induced changes in channel $c \in [1, \ldots, C]$ at voxel i, and model it as a Bernoulli random variable with parameter α_i. We form a binary tumor label vector $\boldsymbol{t}_i = [t_i^1, \ldots, t_i^C]^T$ (where $[\cdot]^T$ indicates the transpose of the vector) of the tumor labels in all C channels, that describes tumor presence in voxel i with probability

$$p(\boldsymbol{t}_i; \alpha_i) = \prod_c p(t_i^c; \alpha_i) = \prod_c \alpha_i^{t_i^c} \cdot (1 - \alpha_i)^{1 - t_i^c}. \tag{2}$$

Here, we assume tumor occurrence to be independent from the type of the underlying healthy tissue. We will introduce conditional dependencies between the underlying tissue class and the likelihood of observing tumor-induced intensity modifications in Sec. II-C.

Observation model: The *image observations* y_i^c are generated by Gaussian intensity distributions for each of the K tissue classes and the C channels, with mean μ_k^c and variance v_k^c, respectively (Fig. 1, purple). In the tumor (i.e., if $t_i^c = 1$), the normal observations are replaced by intensities from another set of channel-specific Gaussian distributions with mean μ_T^c and variance v_T^c representing the tumor class. Letting $\boldsymbol{\theta}$ denote the set of mean and variance parameters for normal tissue and tumor classes, and $\boldsymbol{y}_i = [y_i^1, \ldots, y_i^C]^T$ denote the vector of the intensity observations at voxel i, we form the data likelihood:

$$\begin{aligned} p(\boldsymbol{y}_i | \boldsymbol{t}_i, k_i; \boldsymbol{\theta}) &= \prod_c p(y_i^c | t_i^c, k_i; \boldsymbol{\theta}) \\ &= \prod_c \left[\mathcal{N}(y_i^c \ ; \ \mu_{k_i}^c, v_{k_i}^c)^{1 - t_i^c} \cdot \mathcal{N}(y_i^c \ ; \ \mu_T^c, v_T^c)^{t_i^c} \right], \end{aligned} \tag{3}$$

where $\mathcal{N}(\cdot \ ; \ \mu, v)$ is the Gaussian distribution with mean μ and variance v.

Joint model: Finally, the *joint probability* of the atlas, the latent tumor class, and the observed variables is the product of the components defined in Eqs. (1-3):

$$p(\boldsymbol{y}_i, \boldsymbol{t}_i, k_i; \boldsymbol{\theta}, \alpha_i) = p(\boldsymbol{y}_i | \boldsymbol{t}_i, k_i; \boldsymbol{\theta}) \cdot p(\boldsymbol{t}_i; \alpha_i) \cdot p(k_i). \tag{4}$$

We let $\boldsymbol{Y}$ denote the set of the C image volumes, $\boldsymbol{T}$ denote the corresponding C volumes of binary tumor labels, $\boldsymbol{K}$ denote the tissue labels, and $\boldsymbol{\alpha}$ denote the parameter volume. We obtain the joint probability over all voxels $i \in I$ by forming $p(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{K}; \boldsymbol{\theta}, \boldsymbol{\alpha}) = \prod_{i \in I} p(\boldsymbol{y}_i, \boldsymbol{t}_i, k_i; \boldsymbol{\theta}, \alpha_i)$, assuming that all voxels represent independent observations of the model.

B. Maximum Likelihood Parameter Estimation

We derive an expectation-maximization scheme that jointly estimates the model parameters $\{\boldsymbol{\theta}, \boldsymbol{\alpha}\}$ and the posterior distribution of tissue labels k_i and tumor labels t_i. We start by seeking maximum likelihood estimates of the model parameters $\{\boldsymbol{\theta}, \alpha\}$:

$$\langle \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}} \rangle = \arg\max_{\langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle} p(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\alpha}) \tag{5}$$

$$= \arg\max_{\langle \boldsymbol{\theta}, \boldsymbol{\alpha} \rangle} \prod_{i=1}^{M} p(\boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\alpha}), \tag{6}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

and

$$p(\boldsymbol{y}_i;\boldsymbol{\theta},\boldsymbol{\alpha}) = \sum_{\boldsymbol{t}_i}\sum_{k_i} p(\boldsymbol{y}_i,\boldsymbol{t}_i,k_i;\boldsymbol{\theta},\boldsymbol{\alpha}) \quad (7)$$

$$= \sum_{\boldsymbol{s}_i} p(\boldsymbol{y}_i,\boldsymbol{s}_i;\boldsymbol{\theta},\boldsymbol{\alpha}), \quad (8)$$

where label vector $\boldsymbol{s}_i = [s_i^1,\ldots,s_i^C]^T$ indicates tumor $s_i^c = T$ in all channels with $t_i^c = 1$, and normal tissue $s_i^c = k_i$ for all other channels. As an example with three channels, illustrated in Fig. 2 (voxel 2), suppose $\boldsymbol{t}_i = [0,0,1]$ and $k_i = W$ indicating tumor in channel 3 and image intensities relating to white matter in channels 1 and 2. This results in the tissue label vector $\boldsymbol{s}_i = [W,W,T]$.

E-step: In the E-Step of the algorithm, making use of given estimates of the model parameters $\{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}}\}$, we compute the posterior probability of all $K * 2^C$ tissue label vectors $\boldsymbol{s}_i$. Expanding Eq. (4), we use $\boldsymbol{t}_i(\boldsymbol{s}_i)$ and $k_i(\boldsymbol{s}_i)$ that are corresponding to $\boldsymbol{s}_i$ to simplify notation:

$$p(\boldsymbol{s}_i|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}}) \;\propto\; \pi_{ki}\prod_c \Big[\{\alpha_i\,\mathcal{N}(y_i^c\;;\;\hat{\mu}_T^c,\hat{v}_T^c)\}^{t_i^c}\cdot \{(1-\alpha_i)\,\mathcal{N}(y_i^c\;;\;\hat{\mu}_k^c,\hat{v}_k^c)\}^{1-t_i^c}\Big], \quad (9)$$

and $\sum_{s_i} p(\boldsymbol{s}_i|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}}) = 1$ for all i. Using the tissue label vectors, we can calculate the probability that tumor is visible in channel c of voxel i by summing over all the configurations $\boldsymbol{t}_i$ for which $s_i^c = T$ (or equivalently $t_i^c = 1$):

$$p(s_i^c = T|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}}) = \sum_{\boldsymbol{s}_i}\delta(s_i^c,T)\; p(\boldsymbol{s}_i|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}}), \quad (10)$$

where δ is the Kronecker delta that is equal to 1 for $s_i^c = T$ and 0 otherwise. In the same way we can estimate the probability for the healthy tissue classes k:

$$p(s_i = k|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}}) = \sum_{\boldsymbol{s}_i}\max_c\{\delta(s_i^c,k)\}\; p(\boldsymbol{s}_i|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}}), \quad (11)$$

where $\max_c\{\delta(s_i^c,k)\}$ indicates that one or more of the C channels of label vector s_i contain k.

M-step: In the M-Step of the algorithm, we update the parameter estimates using closed-form update expressions that guarantee increasingly better estimates of the model parameters [65]. The updates are intuitive: the latent tumor prior $\hat{\alpha}_i$ is an average of the corresponding posterior probability estimates

$$\hat{\alpha}_i \leftarrow \sum_{\boldsymbol{s}_i} p(\boldsymbol{s}_i|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})\left[\frac{1}{C}\sum_c \delta(s_i^c,T)\right] = \frac{1}{C}\sum_c p(s_i^c = T|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}}), \quad (12)$$

and the intensity parameters $\hat{\mu}_k^c$ and $\hat{v}_k^c$ are set to the weighted statistics of the data for the healthy tissues ($k = 1,\ldots,K$)

$$\hat{\mu}_k^c \leftarrow \frac{\sum_i p(s_i^c = k|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})\, y_i^c}{\sum_i p(s_i^c = k|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})}, \quad (13)$$

$$\hat{v}_k^c \leftarrow \frac{\sum_i p(s_i^c = k|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})\,(y_i^c - \hat{\mu}_k^c)^2}{\sum_i p(s_i^c = k|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})}. \quad (14)$$

Similarly, for the parameters of the tumor class (T), we obtain

$$\hat{\mu}_T^c \leftarrow \frac{\sum_i p(s_i^c = T|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})\, y_i^c}{\sum_i p(s_i^c = T|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})}, \quad (15)$$

$$\hat{v}_T^c \leftarrow \frac{\sum_i p(s_i^c = T|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})\,(y_i^c - \hat{\mu}_T^c)^2}{\sum_i p(s_i^c = T|\boldsymbol{y}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\alpha}})}. \quad (16)$$

We alternate between updating the parameters $\{\boldsymbol{\theta},\boldsymbol{\alpha}\}$ and the computation of the posteriors $p(\boldsymbol{s}_i|\boldsymbol{y}_i;\boldsymbol{\theta},\boldsymbol{\alpha})$ until convergence, which is typically reached after 10-15 iterations.

C. Enforcing Additional Biological Constraints

Expectation-maximization is a local optimizer. To overcome problems with initialization, we enforce desired properties of the solution by replacing the exact computation with an approximate solution that satisfied additional constraints. These constraints represent our prior knowledge about tumor structure, shape or growth behaviour[1].

Conditional dependencies on tumor occurrence: A possible limitation in the generalization of our probabilistic model is the dimensionality of tissue label vector $\boldsymbol{s}_i$ that has $K*2^C$ possible combinations in Equation (9) and, hence, the computational demands and memory requirements that grow exponentially with the number of channels C in multimodal data sets. To this end, we may want to impose prior knowledge on $p(\boldsymbol{t}_i|k_i)$ and $p(\boldsymbol{t}_i)$ by only considering label vectors $\boldsymbol{s}_i$ that are biologically plausible. First, instead of assuming independence between tissue class and tumor occurrence, we assume conditional dependencies, such as $p(t_i^c = 1|k_i = CSF) = 0$ for all c. We impose this dependency by removing, in this example, all tumor label vectors containing both CSF and tumor from the list of vectors $\boldsymbol{s}_i$ that are included in Equation (9). Second, we can impose constraints on the co-occurrence of tumor-specific changes in the different image modalities (rather than assuming independence here as well), and exclude additional tumor label vectors. We consider, for example, that the edema visible in T2 will always coincide with the edema visible in FLAIR, or that lesions visible in T1 and T1c are always contained within lesions that are visible in T2 and FLAIR.

Together, these constraints reduce the total number of label vectors $\boldsymbol{s}_i$ to be computed in Equation (9), for a standard glioma imaging sequences with T1c, T1, T2, and FLAIR, from $K * 2^C = 3 * 2^4 = 48$ to as few as ten vectors: three *healthy* vectors with $\boldsymbol{t} = [0,0,0,0]$ (corresponding to $[G,G,G,G]$, $[W,W,W,W]$, and $[CSF,CSF,CSF,CSF]$); *edema* with tumor-induced chances visible in FLAIR in the forth channel $\boldsymbol{t} = [0,0,0,1]$ (with $[W,W,W,T]$ and $[G,G,G,T]$); *edema* with tumor-induced changes visible in both FLAIR and in T2 $\boldsymbol{t} = [0,0,1,1]$ (with $[W,W,T,T]$ and $[G,G,T,T]$); the *non-enhancing tumor core* with changes in T1, T2, FLAIR, but without hyper-intensities in T1c $\boldsymbol{t} = [0,1,1,1]$ ($[W,T,T,T]$

[1]An implementation of the generative tumor segmentation algorithm in Python is available from http://ibbm.in.tum.de.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

and $[G, T, T, T]$); the *enhancing tumor core* with hyper-intensities in T1c and additional changes in all other channels $\boldsymbol{t} = [1, 1, 1, 1]$ ($[T, T, T, T]$).

Hyper- and hypo-intense tumor structures: During the iterations of the EM algorithm we enforce that tumor voxels are hyper- or hypo-intense with respect to the current average image intensity μ_k^c of the white matter tissue (hypo-intense for T1, hyper-intense for T1c, T2, FLAIR). Similar to [51], we modify the probability that tumor is visible in channel c of voxel i by comparing the observed image intensity y_i^c with the previously estimated $\hat{\mu}_k^c$ prior to calculating updates for parameters $\boldsymbol{\theta}$ (Eq. 16). We set the probability to zero if the intensity does not align with our expectations:

$$\hat{p}(s_i^c = T|\boldsymbol{y}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}) = \begin{cases} p(s_i^c = T|\boldsymbol{y}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}) & \text{if } y_i^c > \hat{\mu}_k^c, \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

For hypo-intensity constraints we modify the posterior probability updates in the same way, using $y_i^c < \hat{\mu}_k^c$ as a criterion.

Spatial regularity of the tumor prior: Little spatial context is used in the basic model, as we assume the tissue class $\boldsymbol{s}_i$ in each voxel to be independent from the class labels of other voxels. Atlas $\boldsymbol{\pi_k}$ encourages spatially continuous classification for the healthy tissue classes by imposing similar priors in neighboring voxels. To encourage spatial regularity of the tumor labels, we extend the latent atlas α to include a Markov random field (MRF) prior:

$$\hat{p}(\boldsymbol{T}; \beta, \boldsymbol{\alpha}) \propto \prod_c \prod_i \Big[\alpha_i^{t_i^c} (1-\alpha_i)^{1-t_i^c} \cdot \exp\big[-\frac{\beta}{2} \sum_{j \in \mathtt{N}_i} \{ t_i^c (1-t_j^c) + t_j^c (1-t_i^c) \} \big] \Big]. \quad (18)$$

Here, $\mathtt{N}_i$ denotes the set of the six nearest neighbors of voxel i, and β is a parameter governing how similar the tumor labels tend to be at the neighboring voxels. When $\beta = 0$, there is no interaction between voxels and the model reduces to the one described in Section II. By applying a mean-field approximation [66], we derive an efficient approximate algorithm. We let

$$n_i^c = \sum_{j \in \mathtt{N}_i} p(s_j^c = T|\boldsymbol{y}_j; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}) \quad (19)$$

denote the currently estimated "soft" count of neighbors that contain tumor in channel c. The mean-field approximation implies

$$\hat{p}(\boldsymbol{s}_i|\boldsymbol{y}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}) \propto \pi_{ki} \prod_c \Big[\{ \gamma_i^c \, \mathcal{N}(y_i^c \, ; \, \hat{\mu}_T^c, \hat{v}_T^c) \}^{t_i^c} \cdot \{ (1-\gamma_i^c) \, \mathcal{N}(y_i^c \, ; \, \hat{\mu}_k^c, \hat{v}_k^c) \}^{1-t_i^c} \Big] \quad (20)$$

where $\gamma_i^c = \alpha_i / \big[\alpha_i + (1-\alpha_i) \exp(-\beta(2n_i^c - 6)) \big]$, replacing the previously defined Eq. (9), using a channel-specific γ_i^c as a modification of α_i that features the desired spatial regularity.

III. Discriminative Extensions

High-level context at the organ or lesion level, as well as regional information, is not considered in the segmentation process of the generative model. Although we use different constraints to incorporate local neighbourhood information, the generative model treats each voxel as an independent observation and estimates class labels only from very local information. To evaluate global patterns, such as the presence of characteristic artifacts or tumor sub-structures of specific diagnostic interest, we present two alternative discriminative probabilistic methods that make use of both local and non-local image information. The first one, acting at the regional level, is improving the output of the generative model and maintaining its hyper- and hypo-intense lesion maps, while the second one, acting at the voxel level, is transforming the generative model output to any given set of biological tumor labels.

A. The Probabilistic Discriminative Model

We employ an algorithm that predicts the probability of label $l \in L$ for a given observation j which is described by feature vector $\boldsymbol{x}_j = [x_j^1, \ldots, x_j^P]^T$ derived from the segmentations of the generative model. We seek to address two slightly different problems. In the first task, class labels L indicate whether a segmented region j is a result of a characteristic artifact rather than of tumor-induced tissue changes, essentially indicating false positive regions in the segmentations of the generative algorithm that should be removed from the output. In the second task, class labels L represent dense, voxel-wise labels with a semantic interpretation, for example structural attributes of the tumor that do not coincide with the hyper- and hyper-intense segmentations in the different channels, but labels such as "necrosis", or "non-enhancing core". We test both cases in the experimental evaluation, using on channel-wise tumor probabilities $p(s_i^c = T|\boldsymbol{y}_i)$ and on normalized intensities $\boldsymbol{y}_i$ to derive input features for the discriminative algorithms.

To model relations between l_j and $\boldsymbol{x}_j$ for observation $j \in N$, we choose random forests, an ensemble of D randomized decision trees [67]. We use the random forest classifier as it is capable of handling irrelevant predictors and, to some degree, label noise. During training each tree uses a different set of samples X_m^n. It consists of n randomly sampled observations X^n that only contain features from a random subspace of dimensionality $m = log(P)$, where P is the number of features. We learn an ensemble of D different discriminative classifiers, indexed by d, that can be applied to new observations $\boldsymbol{x}_j$ during testing, with each tree predicting the membership $L(d)_j$. When averaging over all D predictions that we obtain for the individual observation, we obtain an estimate of $p(l_j|\boldsymbol{x}_j) = 1/D \sum L(d)_j$. We choose logistic regression trees as discriminative base classifiers for our ensemble, as the resulting oblique random forests perform multivariate splits at each node and are, hence, better capable of dealing with the correlated predictors derived from a multi-modal image data set [68]. For both discriminative approaches we use an ensemble with $D = 255$ decision trees.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

B. A Discriminative Approach Acting at the Regional Level

As many characteristic artifacts have, at the pixel level, a multimodal image intensity patterns that is similar to the one of a lesion, we design a discriminative probabilistic method that is postprocessing and "filtering" the basic output of the generative model. In addition to the pixel-wise intensity pattern, it evaluates *regional* statistics of each connected tumor area, such as volume, location, shape, signal intensities. It replaces commonly used postprocessing routines for quality control that evaluate hand-crafted rules on lesion size or shape and location by a discriminative probabilistic model, similar to [44].

Features and labels: The discriminative classifier acts at the regional level to remove those lesion areas from the output of the generative model that are not associated with tumor, but that stem from arbitrary biological or imaging variation of the voxel intensities. To this end we identify all R isolated regions in the binary tumor map of the FLAIR image (containing voxels i with $p(s_i^{FLAIR} = T|\boldsymbol{y}_i) > 1/2$). We choose FLAIR since it is the most inclusive image modality. As artifacts may be connected to lesion areas, we over-segment larger structures using a watershed algorithm, subdividing regions with connections that are less than 5mm in diameter to reduce the number mixed regions containing both tumor pixels and artifact patterns. For each individual region $r \in 1 \dots R$ we calculate a feature vector $\boldsymbol{x}_r$ that includes volume, surface area, surface-to-volume ratio, as well as regional statistics that are minimum, maximum, mean and median of the normalized image intensities in the four channels. We scale the image intensities for each channels linearly to match the distribution of intensities in a reference data set. We also determine the absolute and the relative number of voxels i with $p(s_i^{T1c} = T|\boldsymbol{y}_i) > 1/2$ within region r, i.e., the volume of the active tumor. We calculate the linear dimensions of the region in axial, sagittal, and transversal direction, the maximal ratio between these three values indicating eccentricity, and the relative location of the region with respect to the center of the brain mask, as well as minimum, maximum, mean and median distances of the regions's voxels from the skull, as a measure of centrality within the brain. We then determine the total number of FLAIR lesions for the given patient and assign this number as another feature to each lesion, together with its individual rank with respect to volume both in absolute numbers and as a normalized rank within $[0, 1]$.

Overall, we construct $P = 39$ features for each region r (Fig. 6). To assign labels to each region, we inspect them visually and assign those overlapping well with a tumor area to the true positive "tumor" class $L_r = 1$, all other to the false positive "artifact" class $L_r = 0$. When labeling regions in the BRATS training data set (Sec. V), all regions labeled as true positives have at least 30% overlap with the "whole tumor" annotation of the experts.

C. A Discriminative Approach Acting at the Voxel Level

The generative model returns a set of probabilistic maps indicating the presence of hypo- or hyper-intense modifications of the tissue. In most applications and imaging protocols, however, it is necessary to localize arbitrary tumor structures – with biological interpretations and clinically relevant semantic labels, such as "edema", "active tumor" or "necrotic core". These structures do not correspond one-by-one to the hypo- and hyper-intense lesions, but have to be inferred by evaluating spatial context and tumor structure as well. We use the probabilistic output of the generative model, together with few secondary features that are derived from the same probabilistic maps and image intensity features, as input to a classifier predicting the posterior probability of the desired semantic labels. The discriminative classifier evaluates local and non-local features to map the output of the generative model to semantic tumor structure and to infer the most likely label L for each given voxel, similar to [32].

Features and labels: To predict a dense set of semantic labels L we extract the following set of features $\boldsymbol{x}_j$ for each voxel j: the tissue prior probabilities $p(k_j)$ for the $K = 3$ tissue classes ($\boldsymbol{x}_j^k$); the tumor probability $p(s_j^c = T)$ for all $C = 4$ channels ($\boldsymbol{x}_j^c$), and the $C = 4$ image intensities after they have been scaled linearly to the intensities of a reference data set ($\boldsymbol{x}_j^{im}$). From these data we derive two types of features. First, we construct the differences of local image intensities or probabilities for all three types of input features ($\boldsymbol{x}_j^k$, $\boldsymbol{x}_j^c$,$\boldsymbol{x}_j^{im}$). This feature captures the difference between the image intensity or probability $\boldsymbol{x}_j$ of voxel j and the corresponding image intensity or probability of another voxel k. For every voxel j in our volume we calculate these differences $\boldsymbol{x}_j^{diff} = \boldsymbol{x}_j - \boldsymbol{x}_k$ for 20 different directions, with spatial offsets in between 3mm to 3cm, i.e., distances that correspond to the extension of most relevant tumor structures. To reduce noise the subtracted values of $\boldsymbol{x}_k$ are extracted after smoothing the image intensities locally around voxel k (using a Gaussian kernel with 3mm standard deviation). We calculate differences between tumor or tissue probability at a given voxel and the probability of the same location on the contralateral side. Second, we evaluate the geodesic distance between voxel j and specific image features that are of particular interest in the analysis. The path is constrained to areas that are most likely gray matter, white matter or tumor as predicted by the generative model. More specifically, we use the distance of $\boldsymbol{x}_j^{\delta tissue}$ of voxel j to the boundary of the the brain tissue (the interface of white and gray matter with CSF), the distance $\boldsymbol{x}_j^{\delta edm}$ to the boundary of the T2 lesion representing the approximate location of the edema. This latter distance $\boldsymbol{x}_j^{\delta edm}$ is calculated independently for voxels outside ($\boldsymbol{x}_j^{\delta edm+}$) and inside ($\boldsymbol{x}_j^{\delta edm-}$) the edema. In the same way, we calculate $\boldsymbol{x}_j^{\delta act+}$ and $\boldsymbol{x}_j^{\delta act-}$ representing the inner and outer distance to the next T1c hyper-intensity. We calculate the number of voxels that are labeled as "edema" or "active tumor" among the direct neighbours of voxel j ($\boldsymbol{x}_j^{edm,act}$), and determine the x-y-z location of the voxel in the co-registered NMI space ($\boldsymbol{x}_j^{MNI}$).

Overall, we construct $P = 651$ image features $\boldsymbol{x}_j = [\boldsymbol{x}_j^k, \boldsymbol{x}_j^c, \boldsymbol{x}_j^{im}, \boldsymbol{x}_j^{diff}, \boldsymbol{x}_j^{\delta}, \boldsymbol{x}_j^{edm,act}, \boldsymbol{x}_j^{MNI}]$ for each voxel j and, when adapted to the BRATS training data set (Sec. V), five labels L_j as provided by clinical experts.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

IV. Experiment 1: Properties and performance of the generative model

In a first experiment, we evaluate the relevance of different components and parameters of the probabilistic model, compare it with related generative approaches,and evaluate the performance on the public BRATS glioma dataset, and test the generalization in a transfer to a related application dealing with stroke lesion segmentation.

A. Data and Evaluation

Glioma data: We use the public BRATS 2012-2013 dataset that provides a total of 45 annotated multimodal glioma image volumes [1]. Training datasets consist of 10/20 low/high-grade cases with native T1, Gadolinium-enhanced T1 (T1c), T2 and FLAIR MR image volumes. The test dataset contains no labels, but can be evaluated by uploading image segmentations to a server; it includes 4/11 low-grade/high-grade cases. Experts have delineated tumor edema, Gadolinium-enhancing "active" core, non-enhancing solid core, cystic/necrotic core. We co-register the probabilistic MNI tissue atlas of SPM99 with the T1 image of each dataset using the FSL software, and sampled to 1mm^3 isotropic voxel resolution. We perform a bias field correction using a polynomial spline model (degree 3) together with a multivariate tissue segmentation using an EM segmenter that is robust against lesions[2] [51]. Image intensities of each channel in each volume are scaled linearly to match the histogram of a reference.

Stroke data: Images are acquired in patients with acute and subacute ischemic stroke. About half of the 18 datasets comprise T1, T2, T1c and FLAIR images in patients in the sub-acute phase, acquired about one or two days after the event; another half comprises T1, T1c, T2 base diffusion and mean diffusivity (MD) images acquired in acute stroke patients within the first few hours after the onset of symptoms. For both groups the imaging sequences return tissue contrasts of normal tissues and lesion areas that are similar to hyper- and hypo-intensities expected in glioma sequences; stroke lesions are characterised here by T1 hypo-, T1c hyper-, T2 hyper- and FLAIR /MD hyper-intense changes. For the quantitative evaluation of the algorithm, we delineate the lesion in every 10th axial, sagittal, and coronal slice, in each of the four modalities. In addition, we annotate about 10% of the 2D slices twice to estimate variability. We register the probabilistic atlas and perform a model based bias field correction as for the glioma data. Image intensities are scaled to the same reference as for the glioma cases.

Evaluation: To measure segmentation performance in the experiments with this dataset, we combine the set of four tumor labels (edema and the three tumor core subtypes) to one binary "complete lesion" label map. We compare this map with the hyper-intense lesion as segmented in T2 and FLAIR. Separately, we compare the "enhancing core" label map with the hyper-intense lesion as segmented in T1c MRI. Quantitatively, we calculate volume overlap between expert annotation A and predicted segmentation B using the Dice score $D(A,B) = 2 * \frac{A \wedge B}{A \vee B}$. We compute Dice scores for whole brain when testing performances on the BRATS data set. We also calculate Dice score within a 3cm distance from the lesion to measure local differences in lesion segmentation rather than in global detection performances.

[2]available from http://www.medicalimagecomputing.com/downloads/ems.php

B. Model Properties and Evaluation on the BRATS data set

Comparison of generative modelling approaches: We compare the proposed generative model against related generative tissue segmentations models and evaluate the relevance of individual components of our approach on the BRATS training data set. We calculate Dice scores in the area containing the lesion and the 3cm margin.

Figure 3A illustrates the benefit of the proposed multivariate tumor and tissue segmentation over a univariate segmentations that treat tumor voxels as intensity outliers similar to Van Leemput's EM segmentation approach for white matter lesion [51]. On the given data this baseline approach leads to a high number of false positives, either requiring stronger spatial regularization or a more adaptive tuning of the outlier threshold. Figure 3B reports the benefit of enforcing intensity constraints within the proposed generative model. While the benefit for the large hyper-intense regions visible in T2 and FLAIR is minor, the difference for segmenting the enhancing tumor core visible in T1c in high-grade patients is striking: the constraint disambiguates tumor-related hypo-intensities – similar to those visible in native T1, for example, from edema – from hyper-intensities induced by the contrast agent in the active rim. Figure 3C reports a comparison between our approach and Prastawa's classic tumor EM segmentation approach [22] that models lesions as an additional class with a "flat" global atlas prior. We test different values for the tumor prior α in Eq. 2, evaluating result for $\alpha_{flat} \in [.005, .01, .02, .04, .1, .2, .4]$. We find that every channel and every segmentation task has a different optimal α_{flat}. However, each optimally tuned generative model with flat prior is still outperformed by the proposed generative model.

Enforcing spatial regularity: Our model has a single parameter that has to be set which is the regularization parameter β coupling segmentations of neighbouring voxels. Based on our previous study [55], we performed all experiments reported in Figure 3 with weak spatial regularization ($\beta = .3$). To confirm these preliminary results we test different regularization settings with $\beta \in [0, 2^{-3}, 2^{-2}, \ldots, 2^{3}]$, now also evaluating channel-specific performance in the lesion area (Figure 7 in the online *Supplementary Materials*). We find a strong regularization to be optimal for the large hyper-intense lesions in FLAIR $\beta \geq .5$, suppressing small spurious structures, while little or no regularization is best for the hardly visible hypo-intense structures visible in T1 ($\beta < .5$). Both T2 and T1c are rather insensitive to regularization. We find the previous value of $\beta = .3$ to work well, but choose $\beta = .5$ for both low- and high-grade tumors in further experiments, somewhat better echoing the relevance of FLAIR.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

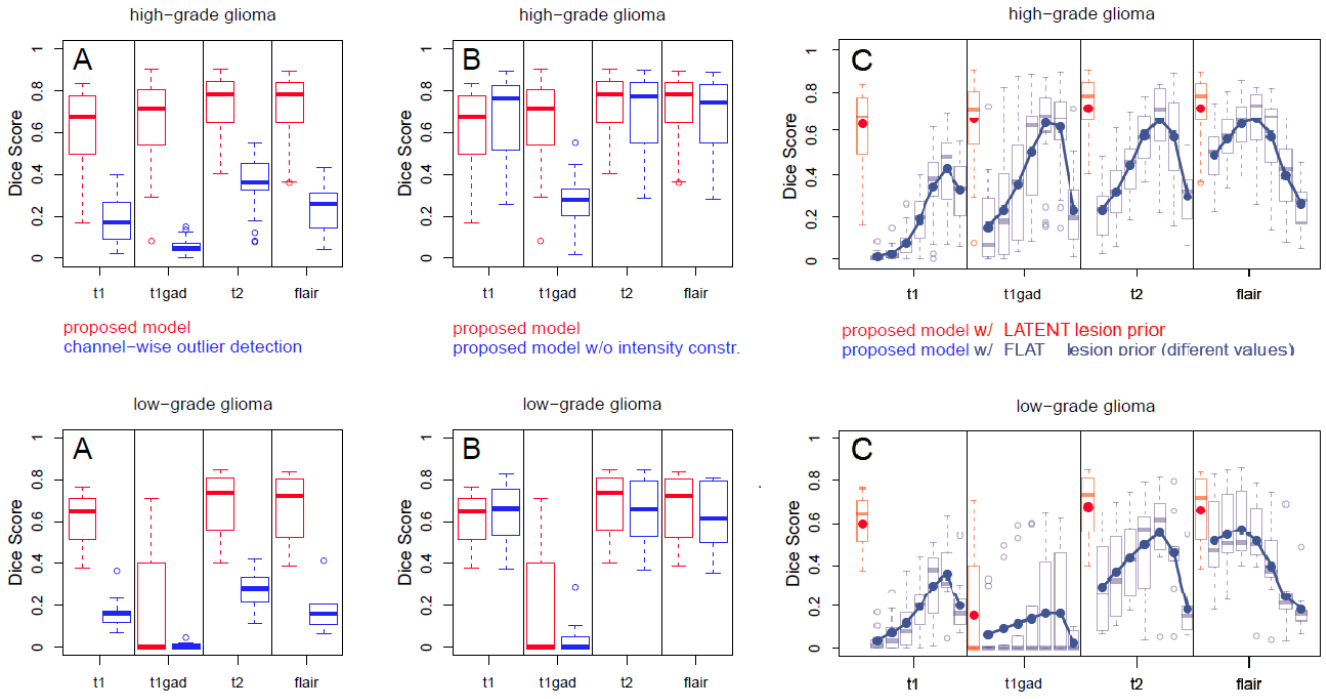

Fig. 3. Evaluation of the generative model and comparison against alternative generative modeling approaches: high-grade (top) and low-grade cases (bottom). Reported are Dice scores for channel-specific segmentations for both low- and high-grade cases in the BRATS training set calculated in the lesion area. Boxplots indicate quartiles, circles indicate outliers. Results of the proposed model are shown in red, while results for related but different generative segmentation methods are shown in blue. Figure A reports performances of univariate tumor segmentations similar to [51]. Figure B: performances of our algorithm with and without constraints on the expected tumor intensities indicating their relevance. Figure C: performance of a generative model with "flat" global tumor prior α_{flat} – i.e., the model of the standard EM segmenter – and evaluating seven different values $\alpha_{flat} \in [.005, \ldots, .4]$. Blue lines and dots in C indicate average Dice scores. The proposed model outperforms all tested alternatives.

Evaluation on the BRATS test set: We apply our segmentation algorithm to the BRATS test sets that have been used for the comparison of twenty glioma segmentation methods in the BRATS evaluation [1]. We identify the segmentations in FLAIR with the "whole tumor" region of the BRATS evaluation protocol, and the T1c enhancing regions with the "active tumor" region. We evaluate two sets of segmentations: segmentations that are obtained by thresholding the corresponding probabilities at 0.5, and the same segmentations after removing all regions that are smaller than 500mm^2 in the FLAIR volume. This latter postprocessing approach was motivated by our observation that smaller regions correspond to false positives in almost all cases. We calculate Dice scores for the whole brain.

Table I reports Dice scores for the BRATS test sets with results of about .60 for the whole tumor and about .50 for the active tumor region ('raw'). As visible from Figure 4, results are heavily affected by false positive regions that have intensity profiles similar to those of the tumor lesions. Applying the basic, size-based postprocessing rule improves results in most cases ('postproc.'). Most false positives are spatially separated from the real lesion and when calculating Dice scores from a region that contains the FLAIR lesion and a 3cm margin only, results improve drastically to average values of .78 (±.09 std.) for the whole tumor to and .55 (±.27 std) for the active region (not shown in the table) which aligns well with results obtained

Task: complete lesion (FLAIR)	mean (±std)	median (±MAD)
BRATS glioma – generative (raw)	.58 (±.22)	.67 (±.11)
BRATS glioma – gener. (postproc.)	.62 (±.21)	.72 (±.11)
BRATS glioma – gener.-discr. (region)	.69 (±.24)	.79 (±.06)
BRATS glioma – gener.-discr. (pixel)	.78 (±.13)	.83 (±.05)
INTER-RATER (4 raters)	.86 (±.06)	.87 (±.06)
Zurich stroke	.78 (±.11)	.79 (±.07)
INTER-RATER (2 raters)	.79 (±.11)	.80 (±.12)
Task: enhancing core (T1c)	mean (±std)	median (±MAD)
BRATS glioma – generative (raw)	.46 (±.26)	.60 (±.15)
BRATS glioma – gener. (postproc.)	.51 (±.27)	.64 (±.15)
BRATS glioma – gener.-discr. (region)	.53 (±.27)	.66 (±.14)
BRATS glioma – gener.-discr. (pixel)	.54 (±.29)	.66 (±.15)
INTER-RATER (4 raters)	.76 (±.10)	.78 (±.08)
Zurich stroke	.45 (±.33)	.64 (±.18)
INTER-RATER (2 raters)	.82 (±.08)	.83 (±.05)

TABLE I

DICE SCORES ON THE TEST SETS USED IN THIS STUDY, FOR THE TWO TASKS OF SEGMENTING THE WHOLE LESION (TOP) AND THE GADOLINIUM ENHANCING STRUCTURES (BOTTOM). BRATS RESULTS ARE CALCULATED ON THE WHOLE BRAIN, STROKE RESULTS IN THE LESION AREA. *Inter-rater* REPRESENTS THE OVERLAP OVER MULTIPLE SEGMENTATIONS OF THE CORRESPONDING TASK AND DATASETS DONE BY HUMAN RATERS [1]. REPORTED ARE MEAN WITH STANDARD DEVIATION AND MEDIAN WITH MEDIAN ABSOLUTE DEVIANCE.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

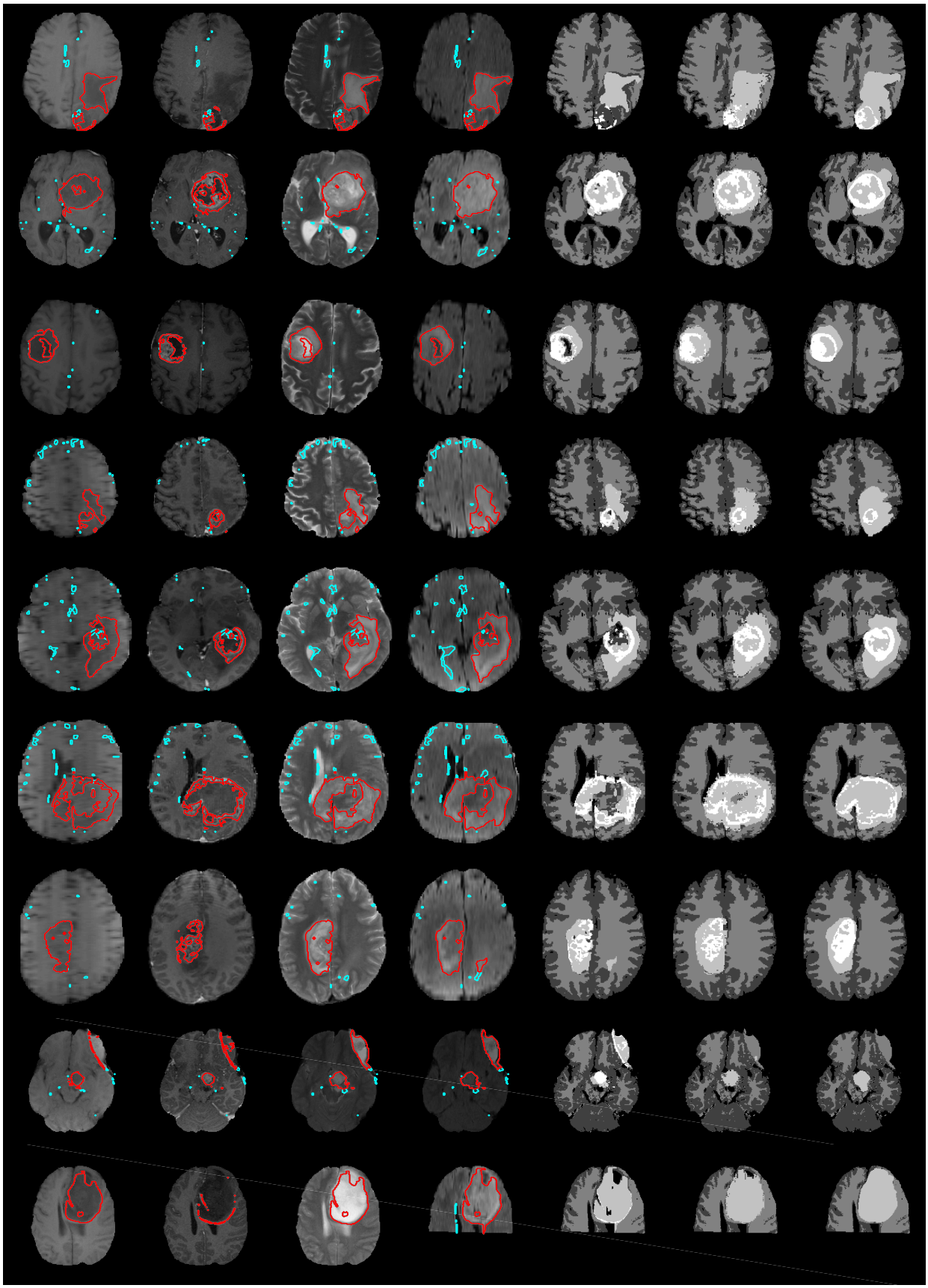

Fig. 4. Exemplary BRATS test sets, with results for generative and generative discriminative models. Shown are axial views through the tumor center for T1, T1c, T2 and FLAIR image (columns from left to right) and the segmented hypo- or hyper-intense areas (red and cyan). Regions outlined in red have been identified as "true positive" regions by the regional discriminative classifier and the resulting tumor labels are shown in column 5 with edema (bright gray) and active tumor region (white). Column 6 shows results of the voxel-wise generative-discriminative classifier, and column 7 the expert's annotation. Gray and white matter segmentations displayed in the last three columns are obtained by the generative model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

for T2 and T1c on the training set (Fig. 3).

C. Generalization Performance and Transfer to the Stroke Data Set

We test the generalization performance of the generative model by using it for delineating ischemic stroke lesions that are similar in terms of lesion size and clinical image information. We apply the generative model as optimized for the BRATS dataset to the stroke images. As a single modification we allow T1c lesions to be both inside the FLAIR and T2 enhancing area and outside, as bleeding (which leads to the T1c hyper-intensities) may not coincide with the local edema. Stroke images contain cases with both active and chronic lesions with significantly different lesion patterns.

Although datasets, imaging protocol, and even major acquisition parameters differ, we obtain results that are comparable to the tumor data. We calculate segmentations accuracies in the lesion area and observe good agreement between manual delineation and automatic segmentation in all four modalities (Fig. 5). We also observe false positives at the white matter–gray matter interfaces, similar to those we observed for the glioma tests data (Fig. 4). Most false positive regions are disconnected from the lesion and could be removed with little user interaction or postprocessing. Inter-rater differences and performance of the algorithm are comparable to those from the glioma test set, with Dice scores close to .80 for segmenting the edema and around .50-.60 for T1c enhancing structures (Table I). Results on the stroke data underline the versatility of the generative lesion segmentation model and its good generalization performance not only across different imaging sequences, but also across applications. To the best of our knowledge this is one of the first attempts to automatically segment ischemic stroke lesion in multimodal images using a generative model.

V. Experiment 2: Properties and performances of the generative-discriminative model extensions

Results of the generative model show its robustness and accuracy for delineating lesion structures. Still, it also shows to be sensitive to artifacts that cannot be recognized by evaluating the multimodal intensity pattern at the voxel level, and hypo- and hyper-intense structures can only be matched with selected tumor labels. To this end, we evaluate the two discriminative modeling strategies that are considering non-local features as input and arbitrary labels as output. We first evaluate model properties on the BRATS training set and then compare performances to results of other state-of-the art tumor segmentation algorithms on the BRATS test set.

A. Relevant Features and Information used by the Discriminative Models

The random forest classifier handles learning tasks with small sample sizes in high dimensional spaces by only relying on few "strong" variables and ignoring irrelevant features [69]. Still, in order to understand the information used when modeling the class probabilities, we can visualize the importance of the input features used. To this end we evaluate the relevance of the individual features using Breiman's feature permutation test [67] that compares the test error with the error obtained after the values of a given feature have been randomly permuted throughout all test samples. The resulting decrease in test accuracy, or increase in test error, indicates how relevant the chosen feature is for the overall classification task. Repeated for each feature of all trees in the decision forest, this measure helps to rank the features and to compare the relevance as shown in Figure 6. In our test we augment the dataset with a random feature (random samples from a Gaussian distribution with mean 0 and standard deviation of 1) to establish a lower baseline of the relevance score. For each feature we compare the distribution of changes in classification error against the changes of this random feature in a paired Cox-Wilcoxon test. We analyze feature relevance in a cross-validation on the BRATS training set.

Results for the first discriminative model acting at the regional level are shown in Fig. 6. We find plausible features to be relevant: the relative location of the region with respect to the center of the brain (indicated as $center\_x, center\_y, center\_z$ in the figure), the surface-to-volume ratio ($border2area$), the total number of lesions visible for the given patient ($num\_lesions$), the ratio of segmented voxels in T1c ($tumor_T1cN$), and some descriptors of image intensities, such as the minimum in FLAIR ($int4\_min$), the maximum, median and average of the T2 intensities ($int3\_*$), as well as the maximum in T1c ($int1\_max$) and the minimum in T1 ($int2\_min$).

For the second discriminative model acting at the pixel level we find about 80% of the features to be relevant, with with some variation across the different classification tasks. The features that rank highest in all tests are those we derived from the probability maps of the generative model: the total number of local edema or active tumor voxels, the geodesic distance to the nearest edema or active tumor pixels, but also the relative anatomical location in the MNI space, and selected image intensities and intensity differences (such as the intensity values of T1 and FLAIR for edema and T1c for active core, and local differences in the T1 image intensities).

B. Performance on the BRATS Test Set

Figure 4 displays nine exemplary image volumes of the BRATS test set. Shown are the raw probability maps of the generative model (red and cyan; columns 1-4), those regions that are selected by the regional discriminative model (cyan) and the derived tumor segmentation (column 5), as well as the output of the voxel-level tumor classifier (column 6), together with an expert annotation (column 7).

Quantitative results are reported in Table I, and we find both discriminative models to improve results over those derived from the "raw" probability maps of the generative model. With few exceptions most "false positive" artifact regions are removed (Fig. 4). The voxel-level model shows to be more accurate than the regional-level model, also correcting for "false negative" areas in the center of the tumor (rows 1, 3, 6, and 7). In addition to whole tumor and active tumor

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

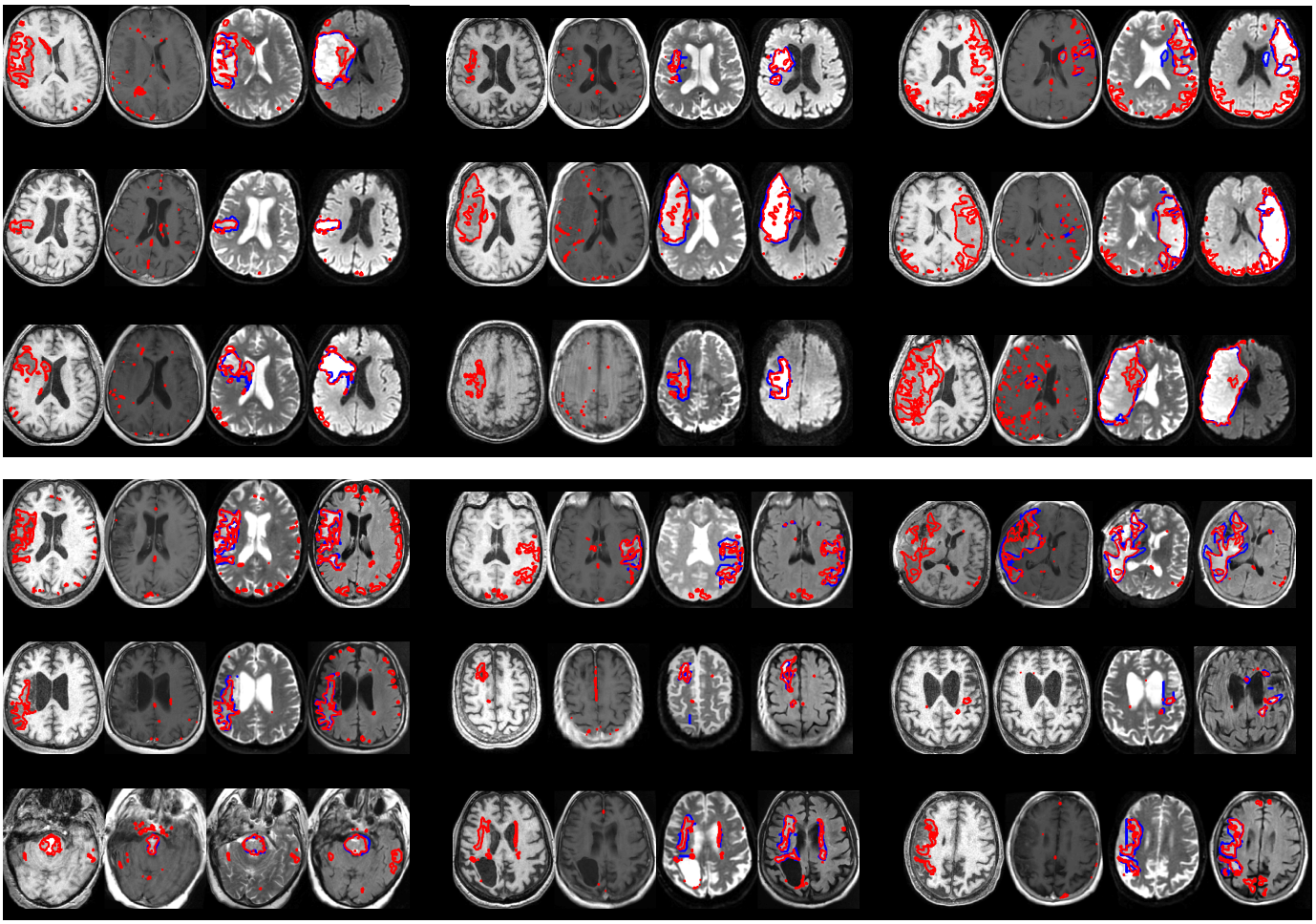

Fig. 5. Generalization to ischemic stroke cases (showing acute stroke: rows 1-3; subacute stroke: rows 4-6) with T1, T1c, FLAIR/MD, T2/MAD images of each patient (from left to right, three patients per row). Automatic segmentations are delineated in red, lesions in manually segmented volumes are shown in blue (typically beneath a red line); T1 and T1c lesions were only visible for some cases.

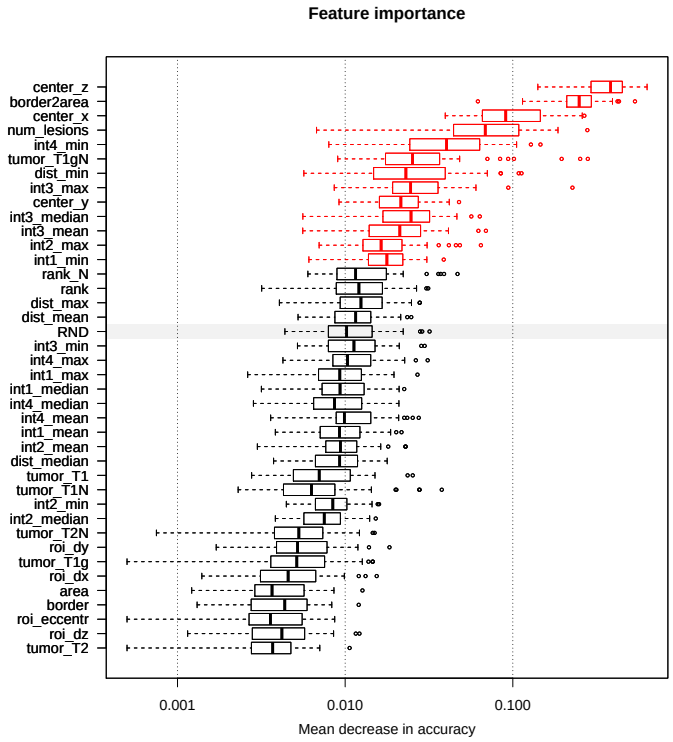

Fig. 6. Measuring feature relevance of the discriminative model. Features relevant for discriminating between false positives and true positives regions. We evaluate the permutation importance [67] of each feature extracted for the FLAIR regions (see text for details). Boxplots show the decrease in accuracy for all 255 trees of the oblique random forest (boxes representing quartiles) with high values indicating high relevance. The gray bar indicates the performance of a random feature ("RND") under this measure, features displayed in red perform significantly better (as indicated by a paired Cox-Wilcoxon test at 5% level). Location and shape of the regions are most discriminative, as well as the general number of lesions visible in the given FLAIR image, and selected image intensities.

areas, the second discriminative model is also predicting the location of necrotic and fluid filled structures, as well as the "tumor core" label (with a Dice score of .58; segmentations not shown in the figure). Sensitivities and specificities for this latter model are balanced with sensitivities of .75/.58/.63 for the three tumor regions (whole tumor/tumor core/active tumor) and a specificities of .86/.71/.56.

The BRATS challenge also allows us to compare the two generative-discriminative modeling approaches with eighteen other state of the art methods including inter-active ones, and we reproduce results of the challenge in Figure 8 in the online *Supplementary Materials* of this manuscript. The generative model with discriminative post-processing at the regional level (indicated by *Menze (G)*) performs comparable to most other approaches in terms of Dice score and robust Hausdorff distance for "whole tumor" and "active tumor". However, it cannot model the "tumor core" segmentation task as this structure does not have a direct correspondence to any of the segmented hyper- and hypo-intensity regions and, hence, does does not provide competitive results for this tumor substructure. The voxel-level generative-discriminative approach (indicated by *Menze (D)*) is able to predict "tumor core" labels. It ranks first among the twenty evaluated methods in terms of average Hausdorff distances for both "tumor core" and in "whole tumor" segmentation, and it is the second best automatic method for the "active tumor" segmentation. In the evaluation of the average Dice scores it is second best for "whole tumor", it is ranking third among the automated methods for the "tumor core" task, and its result are statistically indistinguishable from the inter-rater variation for "active tumor". Most notably, the voxel-level generative-discriminative approach is outperforming all discriminative models that are similar in terms of random forest classifier and feature design [37], [32], [2], [34], but that do not rely on the input features derived from the probability maps of the generative model.

VI. Summary and Conclusions

In this paper, we extend the atlas-based EM segmenter by a latent atlas class that represents the probability of transition from any of the "healthy" tissues to a "lesion" class. In practice, the latent atlas serves as an adaptive prior that couples the probability of observing tumor-induced intensity changes across different imaging channels for the same voxel. Using the standard brain atlas for healthy tissues together with the highly specific multi-channel information provides us with segmentations of the healthy tissues surrounding the tumor, and enables us to automatically segment the images. The proposed generative algorithm produces outlines of the tumor-induced changes for each channel which makes it independent of the multimodal imaging protocol. We complement the basic probabilistic model with a discriminative model and test two different modeling strategies, both of them addressing shortcomings of the generative model, and find the resulting discriminative-generative model to define the state of the art in tumor segmentation on the BRATS data set [1].

The proposed generative algorithm generalizes the probabilistic model of the standard EM segmenter. As such, it can be improved by combining registration and segmentation [62], or by integrating empirical or physical bias-field correction models [15], [70]. The generative segmentation algorithm that we optimized for glioma images exhibits a good level of generalization when applied to multimodal images from patients with ischemic stroke. The method is likely to also work well for traumatic brain injury with similar hypo- and hyper-intensity patterns, and it can also be adapted to multimodal segmentation tasks beyond the brain. It may be interesting to evaluate relations to multi-channels approaches that do not rely on multiple physical channels, but high-dimensional sets of features extracted from one or few physical images [71]. Analyzing feature relevance indicated that the location of a voxel or region within the MNI space helped in removing false positives, as most of them appeared at white matter–gray matter interfaces or in areas that are often subject to B-field inhomogeneities. Extensions of the generative model may use a location prior that lowers the expectation of tumor occurrences in these areas. Moreover, preliminary findings suggest that results may improve by using non-Gaussian intensity models for the lesion classes.

Some tumor structures – such as necrotic or cystic regions, or the solid tumor core – cannot easily be associated with local channel-specific intensity modifications, but are rather identified based on the wider spatial context and their relation with other tumor compartments. We addressed the segmentation of such secondary structures by combining our generative model with discriminative model extensions evaluating additional non-local features. As an alternative, relations between visible tumor structures can be enforced locally using MRF as proposed by [35], or in a non-local fashing following the hierarchical approach following [54]. Future work may also aim at integrating image segmentation with tumor growth models enforcing spatial or temporal relations as in [53], [14]. Tumor growth models – often described through partial differential equations [72] – offer a formal description of the lesion evolution, and could be used to describe the propagation of channel-specific tumor outlines in longitudinal series [73], as well as a shape and location prior for various tumor structures [23]. This could also promote a deeper integration of underlying functional models of disease progression and formation of image patterns in the modalities that are used to monitor this process [74].

To support the further use and analysis of our generative segmentation algorithm, we make an implementation available in Python from http://ibbm.in.tum.de, also illustrating its use on reference data from the BRATS challenge.

Acknowledgements

This research was supported by the The National Alliance for Medical Image Analysis (NIH NIBIB NAMIC U54-EB005149), The Neuroimaging Analysis Center (NIH NIBIB NAC P41EB015902), NIH NIBIB (R01EB013565), the Lundbeck Foundation (R141-2013-13117),the European Research Council through the ERC Advanced Grant MedYMA 2011-291080 (on Biophysical

Modeling and Analysis of Dynamic Medical Images), the Swiss NSF project Computer Aided and Image Guided Medical Interventions (NCCR CO-ME), the German Academy of Sciences Leopoldina (Fellowship Programme LPDS 2009-10), the Technische Universität München - Institute for Advanced Study (funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement n 291763), the Marie Curie COFUND program of the the European Union (Rudolf Mössbauer Tenure-Track Professorship to BHM).

REFERENCES

[1] B. Menze and el al, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, p. 33, 2014.

[2] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," Phys Med Biol, vol. 58, no. 13, pp. R97–R129, Jul. 2013.

[3] E. D. Angelini, O. Clatz, E. Mandonnet, E. Konukoglu, L. Capelle, and H. Duffau, "Glioma dynamics and computational models: A review of segmentation, registration, and in silico growth algorithms and their clinical applications," Curr Medl Imaging Rev, vol. 3, pp. 262–276, 2007.

[4] M. Kaus, S. K. Warfield, A. Nabavi, E. Chatzidakis, P. M. Black, F. A. Jolesz, and R. Kikinis, "Segmentation of meningiomas and low grade gliomas in MRI," in Proc MICCAI, 1999, pp. 1–10.

[5] Y.-F. Tsai, I.-J. Chiang, Y.-C. Lee, C.-C. Liao, and K.-L. Wang, "Automatic MRI meningioma segmentation using estimation maximization." Proc IEEE Eng Med Biol Soc, vol. 3, pp. 3074–3077, 2005.

[6] E. Konukoglu, W. M. Wells, S. Novellas, N. Ayache, R. Kikinis, and P. M. B. an K M Pohl, "Monitoring slowly EVolving tumors," in Proc ISBI, 2008, pp. 1–4.

[7] B. Bach Cuadra, C. Pollo, A. Bardera, O. Cuisenaire, and J. P. Thiran, "Atlas-based segmentation of pathological brain MR images using a model of lesion growth," IEEE T Med Imag, vol. 23, pp. 1301–14, 2004.

[8] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells, and S. Warfield, "3D segmentation in the clinic: A grand challenge ii: MS lesion segmentation," MIDAS Journal, pp. 1–5, 2008.

[9] A. Irimia, M. C. Chambers, J. R. Alger, M. Filippou, M. W. Prastawa, B. Wang, D. A. Hovda, G. Gerig, A. W. Toga, R. Kikinis, P. M. Vespa, and J. D. Van Horn, "Comparison of acute and chronic traumatic brain injury using semi-automatic multimodal segmentation of MR volumes." J Neurotrauma, vol. 28, pp. 2287–2306, 2011.

[10] M. Shenton, H. Hamoda, J. Schneiderman, S. Bouix, O. Pasternak, Y. Rathi, M.-A. Vu, M. Purohit, K. Helmer, I. Koerte et al., "A review of magnetic resonance imaging and diffusion tensor imaging findings in mild traumatic brain injury," Brain imaging and behavior, vol. 6, pp. 137–192, 2012.

[11] T. D. Farr and S. Wegener, "Use of magnetic resonance imaging to predict outcome after stroke: a review of experimental and clinical evidence," J Cerebr Blood Flow Metab, vol. 30, pp. 703–717, 2010.

[12] I. Rekik, S. Allassonnière, T. K. Carpenter, and J. M. Wardlaw, "Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. a critical appraisal," NeuroImage: Clinical, 2012.

[13] Y. Tarabalka, G. Charpiat, L. Brucker, and B. H. Menze, "Enforcing monotonous shape growth or shrinkage in video segmentation." in Proc BMCV (British Machine Vision Conference), 2013.

[14] E. Alberts, G. Charpiat, Y. Tarabalka, M. A. Weber, C. Zimmer, and M. B. H., "A nonparametric growth model for estimating tumor growth in longitudinal image sequences," in Proc MICCAI Brain Lesions Workshop (BrainLes), ser. LNCS. Springer, 2015.

[15] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain." IEEE T Med Imaging, vol. 18, pp. 885–896, 1999.

[16] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig, "A brain tumor segmentation framework based on outlier detection." Med Image Anal, vol. 8, pp. 275–283, 2004.

[17] K. M. Pohl, J. Fisher, J. J. Levitt, M. E. Shenton, R. Kikinis, W. E. L. Grimson, and W. M. Wells, "A unifying approach to registration, segmentation, and intensity correction." in LNCS 3750, Proc MICCAI, 2005, pp. 310–318.

[18] E. I. Zacharaki, D. Shen, and C. Davatzikos, "ORBIT: A multiresolution framework for deformable registration of brain tumor images," IEEE T Med Imag, vol. 27, pp. 1003–17, 2008.

[19] B. Bach Cuadra, C. Pollo, A. Bardera, O. Cuisenaire, and J. P. Thiran, "Atlas-based segmentation of pathological brain MR images using a model of lesion growth," IEEE T Med Imag, vol. 23, pp. 1301–14, 2004.

[20] D. Gering, W. Grimson, and R. Kikinis, "Recognizing deviations from normalcy for brain tumor segmentation," Lecture Notes In Computer Science, vol. 2488, pp. 388–395, Sep. 2002.

[21] N. Moon, E. Bullitt, K. Van Leemput, and G. Gerig, "Model-based brain and tumor segmentation," in Proc ICPR, 2002, pp. 528–31.

[22] M. Prastawa, E. Bullitt, N. Moon, K. V. Leemput, and G. Gerig, "Automatic brain tumor segmentation by subject specific modification of atlas priors." Acad Radiol, vol. 10, pp. 1341–48, 2003.

[23] A. Gooya, K. M. Pohl, M. Bilello, L. Cirillo, G. Biros, E. R. Melhem, and C. Davatzikos, "GLISTR: glioma image segmentation and registration." IEEE Trans. Med. Imag., vol. 31, pp. 1941–1954, 2012.

[24] S. Parisot, H. Duffau, S. Chemouny, and N. Paragios, "Joint tumor segmentation and dense deformable registration of brain MR images," in Proc MICCAI, 2012, pp. 651–658.

[25] A. Gooya, G. Biros, and C. Davatzikos, "Deformable registration of glioma images using em algorithm and diffusion reaction modeling." IEEE Trans. Med. Imag., vol. 30, pp. 375–390, 2011.

[26] M. Lorenzi, H. Menze, Bjoern, M. Niethammer, N. Ayache, and X. Pennec, "Sparse Scale-Space Decomposition of Volume Changes in Deformations Fields," in Proc MICCAI, vol. 8150, 2013, pp. 328–335.

[27] D. Cobzas, N. Birkbeck, M. Schmidt, M. Jagersand, and A. Murtha, "3D variational brain tumor segmentation using a high dimensional feature set," in Proc ICCV, 2007, pp. 1–8.

[28] A. Lefohn, J. Cates, and R. Whitaker, "Interactive, GPU-based level sets for 3D brain tumor segmentation." in Proc MICCAI, 2003, pp. 564–572.

[29] L. Gorlitz, B. H. Menze, M.-A. Weber, B. M. Kelm, and F. A. Hamprecht, "Semi-supervised tumor detection in magnetic resonance spectroscopic images using discriminative random fields," in Proc DAGM, ser. LNCS, 2007, pp. 224–233.

[30] C. Lee, S. Wang, A. Murtha, and R. Greiner, "Segmenting brain tumors using pseudo conditional random fields." in LNCS 5242, Proc MICCAI, 2008, pp. 359–66.

[31] M. Wels, G. Carneiro, A. Aplas, M. Huber, J. Hornegger, and D. Comaniciu, "A discriminative model-constrained graph cuts approach to fully automated pediatric brain tumor segmentation in 3D MRI," in LNCS 5241, Proc MICCAI, 2008, pp. 67–75.

[32] D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. M. Thomas, T. Das, R. Jena, and P. S. J., "Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR," in Proc MICCAI, 2012.

[33] E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images." Neuroimage, vol. 57, pp. 378–90, 2011.

[34] E. Geremia, B. H. Menze, and N. Ayache, "Spatially adaptive random forests." in Proc IEEE ISBI, 2013.

[35] S. Bauer, C. May, D. Dionysiou, G. S. Stamatakos, P. Büchler, and M. Reyes, "Multi-Scale Modeling for Image Analysis of Brain Tumor Studies." IEEE Trans. Bio-Med. Eng., Aug. 2011.

[36] W. Wu, A. Y. Chen, L. Zhao, and J. J. Corso, "Brain tumor detection and segmentation in a conditional random fields framework with pixel-pairwise affinity and superpixel-level features," International journal of computer assisted radiology and surgery, pp. 1–13, 2013.

[37] N. J. Tustison, K. L. Shrinidhi, M. Wintermark, C. R. Durst, B. M. Kandel, J. C. Gee, M. C. Grossman, and B. B. Avants, "Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with antsr." Neuroinformatics, vol. 13, no. 2, pp. 209–225, Apr 2015. [Online]. Available: http://dx.doi.org/10.1007/s12021-014-9245-2

[38] P. Dvorak and B. H. Menze, "Structured prediction with convolutional neural networks for multimodal brain tumor segmentation," in Proc MICCAI MCV (Medical Computer Vision Workshop), 2015.

[39] G. Urban, M. Bendszus, F. A. Hamprecht, and J. Kleesiek, "Multi-modal Brain Tumor Segmentation using Deep Convolutional Neural Networks," in Proc MICCAI BRATS (Brain Tumor Segmentation Challenge), 2014, pp. 31–35.

[40] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. Van Leemput, and B. Fischl, "Is synthesizing MRI contrast useful for inter-modality analysis?" Proc MICCAI, p. in press, 2013.

[41] S. Roy, A. Carass, and J. Prince, "A compressed sensing approach for MR tissue contrast synthesis," in Proc IPMI, 2011, pp. 371–383.

[42] S. Roy, A. Carass, N. Shiee, D. L. Pham, P. Calabresi, D. Reich, and J. L. Prince, "Longitudinal intensity normalization in the presence of multiple sclerosis lesions," in Proc ISBI, 2013.

[43] N. K. Batmanghelich, B. Taskar, and C. Davatzikos, "Generative-discriminative basis learning for medical imaging." IEEE Trans Med Imaging, vol. 31, no. 1, pp. 51–69, Jan 2012. [Online]. Available: http://dx.doi.org/10.1109/TMI.2011.2162961

[44] B. H. Menze, B. M. Kelm, P. Bachert, M.-A. Weber, and F. A. Hamprecht, "Mimicking the human expert : Pattern recognition for an automated assessment of data quality in MR spectroscopic images," Magn Reson Med, vol. 1466, pp. 1457–66, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18421692

[45] A. Criminisi, K. Juluru, and S. Pathak, "A discriminative-generative model for detecting intravenous contrast in ct images." Med Image Comput Comput Assist Interv, vol. 14, no. Pt 3, pp. 49–57, 2011.

[46] Z. Tu, K. L. Narr, P. Dollar, I. Dinov, P. M. Thompson, and A. W. Toga, "Brain anatomical structure segmentation by hybrid discriminative/generative models." IEEE Trans Med Imaging, vol. 27, no. 4, pp. 495–508, Apr 2008. [Online]. Available: http://dx.doi.org/10.1109/TMI.2007.908121

[47] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods." IEEE Trans Med Imaging, vol. 30, no. 9, pp. 1617–1634, Sep 2011. [Online]. Available: http://dx.doi.org/10.1109/TMI.2011.2138152

[48] W. Speier, J. E. Iglesias, L. El-Kara, Z. Tu, and C. Arnold, "Robust skull stripping of clinical glioblastoma multiforme data." Med Image Comput Comput Assist Interv, vol. 14, no. Pt 3, pp. 659–666, 2011.

[49] T. Riklin-Raviv, B. H. Menze, K. Van Leemput, B. Stieltjes, M. A. Weber, N. Ayache, W. M. Wells, and P. Golland, "Joint segmentation via patient-specific latent anatomy model," in Proc MICCAI-PMMIA (Workshop on Probabilistic Models for Medical Image Analysis), 2009, pp. 244–255.

[50] T. Riklin-Raviv, K. Van Leemput, B. H. Menze, W. M. Wells, 3rd, and P. Golland, "Segmentation of image ensembles via latent atlases." Med Image Anal, vol. 14, pp. 654–665, 2010.

[51] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection." IEEE T Med Imaging, vol. 20, pp. 677–688, 2001.

[52] S. Bauer, J. Tessier, O. Krieter, L. Nolte, and M. Reyes, "Integrated spatio-temporal segmentation of longitudinal brain tumor imaging studies," in Proc MICCAI-MCV, Springer LNCS, 2013.

[53] Y. Tarabalka, G. Charpiat, L. Brucker, and B. H. Menze, "Spatio-temporal video segmentation with shape growth or shrinkage constraint." IEEE Transactions in Image Processing, vol. 23, pp. 3829–3840, 2014.

[54] J. J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient multilevel brain tumor segmentation with integrated Bayesian model classification," IEEE T Med Imag, vol. 9, pp. 629–40, 2008.

[55] B. H. Menze, K. Van Leemput, D. Lashkari, M.-A. Weber, N. Ayache, and P. Golland, "A generative model for brain tumor segmentation in multi-modal images," in Proc MICCAI, ser. LNCS 751, 2010, pp. 151–159.

[56] B. H. Menze, K. Van Leemput, D. Lashkari, M.-A. Weber, N. Ayache, and P. Golland, "Segmenting glioma in multi-modal images using a generative model for brain lesion segmentation," in Proc MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge), 2012, p. 7p.

[57] B. H. Menze, E. Geremia, N. Ayache, and G. Szekely, "Segmenting glioma in multi-modal images using a generative-discriminative model for brain lesion segmentation," in Proc MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge), 2012, p. 8.

[58] W. Wells, W. Grimson, R. Kikinis, and F. Jolesz, "Adaptive segmentation of MRI data," in Proc Computer Vision, Virtual Reality and Robotics in Medicine, 1995, pp. 57–69.

[59] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," IEEE T Med Imaging, vol. 15, pp. 429–442, 1996.

[60] J. Ashburner and K. Friston, "Multimodal image coregistration and partitioning–a unified framework." Neuroimage, vol. 6, no. 3, pp. 209–217, Oct 1997. [Online]. Available: http://dx.doi.org/10.1006/nimg.1997.0290

[61] K. M. Pohl, J. Fisher, W. Grimson, R. Kikinis, and W. Wells, "A Bayesian model for joint segmentation and registration," Neuroimage, vol. 31, pp. 228–239, 2006.

[62] K. M. Pohl, S. K. Warfield, R. Kikinis, W. E. L. Grimson, and W. M. Wells, "Coupling statistical segmentation and -PCA- shape modeling," in Proc MICCAI, 2004, pp. 151–159.

[63] K. Van Leemput, "Encoding probabilistic brain atlases using Bayesian inference," IEEE TMI, vol. 28, pp. 822–837, 2009.

[64] J. E. Iglesias, M. R. Sabuncu, and K. V. Leemput, "Improved inference in Bayesian segmentation using Monte Carlo sampling: Application to hippocampal subfield volumetry," Med Image Anal, p. in press, 2013.

[65] A. Dempster and e. al., "Maximum likelihood from incomplete data via the em algorithm." Journal of the Royal Statistical Society, vol. 39, pp. 1–38, 1977.

[66] M. I. Jordan, Z. Gaharamani, T. S. Jaakola, and L. K. Saul, "An introduction to variational methods for graphical models," Machine Learning, vol. 37, pp. 183–233, 1999.

[67] L. Breiman, "Random forests," Mach Learn J, vol. 45, pp. 5–32, 2001.

[68] B. H. Menze, B. M. Kelm, D. N. Splitthoff, U. Koethe, and F. A. Hamprecht, "On oblique random forests," in Proc ECML (European Conference on Machine Learning), 2011.

[69] L. Breiman, "Consistency for a simple model of random forests," UC Berkeley, Tech. Rep. 670, 2004.

[70] C. Poynton, M. Jenkinson, and W. Wells Iii, "Atlas-based improved prediction of magnetic field inhomogeneity for distortion correction of EPI data," in Proc MICCAI, 2009, pp. 951–959.

[71] M. Toews, L. Zollei, and W. W. III., "Invariant feature-based alignment of volumetric multi-modal images." in Proc IPMI, 2013.

[72] B. H. Menze, E. Stretton, E. Konukoglu, and N. Ayache, "Image-based modeling of tumor growth in patients with glioma." in Optimal control in image processing, C. S. Garbe, R. Rannacher, U. Platt, and T. Wagner, Eds. Springer, Heidelberg/Germany, 2011.

[73] E. Konukoglu, O. Clatz, B. H. Menze, B. Stieltjes, M.-A. Weber, E. Mandonnet, H. Delingette, and N. Ayache, "Image guided personalization of reaction-diffusion type tumor growth models using modified anisotropic eikonal equations." IEEE T Med Imag, vol. 29, pp. 77–95, 2010.

[74] B. H. Menze, K. Van Leemput, A. Honkela, E. Konukoglu, M. A. Weber, N. Ayache, and P. Golland, "A generative approach for image-based modeling of tumor growth," in Proc IPMI (Inform Proc Med Imag), 2011.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

Supplementary materials for "Menze et al. *A generative probabilistic model and discriminative extensions for brain lesion segmentation - with application to tumor and stroke.* IEEE Transactions on Medical Imaging 2016."

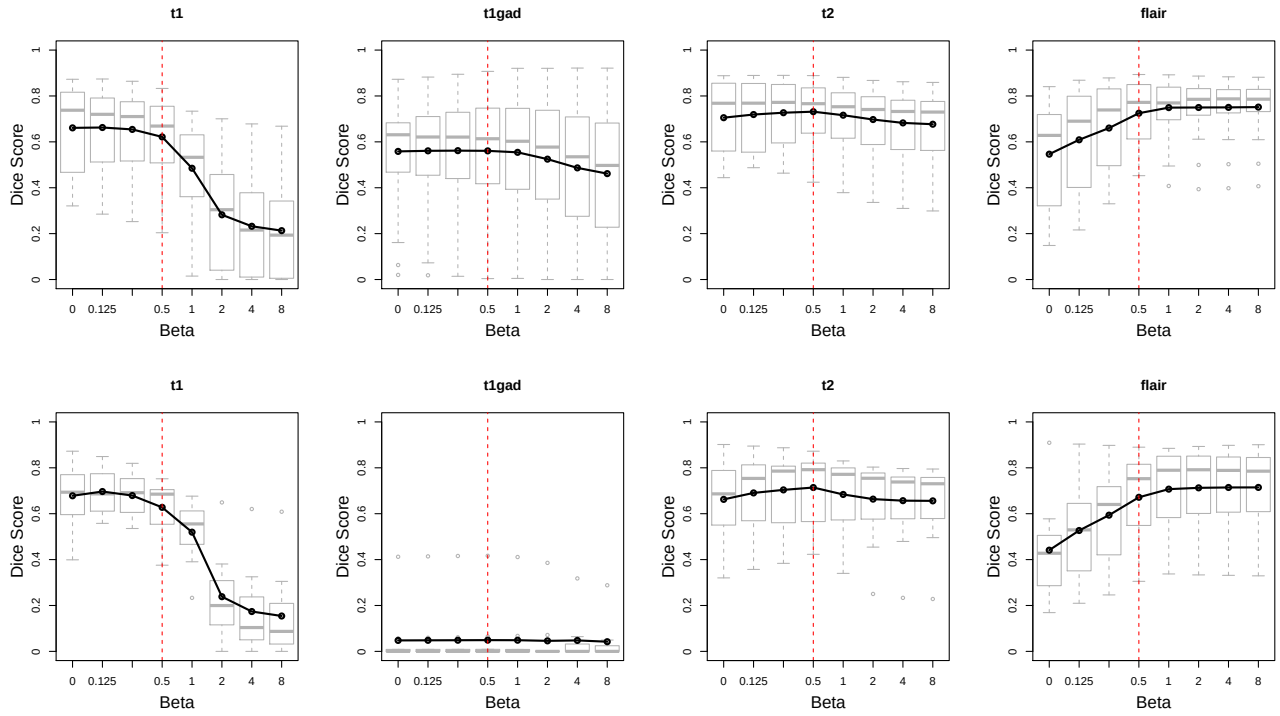

Fig. 7. Optimal spatial regularization of the tumor prior: high-grade (top row), and low-grade cases (bottom row). Reported are Dice scores for channel-specific segmentations for low- and high-grade cases of the BRATS training set, testing different values of regularization parameters $\beta \in [0, .125, ..., 8]$ in Eq. (9). Gray boxplots represent quartiles, with notches indicating outliers. Black lines and circles correspond to the mean performance. While T1c and T2 segmentations are rather insensitive to the choice of β, we choose an intermediate value of $\beta = .5$ (red dashed line) in further experiments.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2502596, IEEE Transactions on Medical Imaging

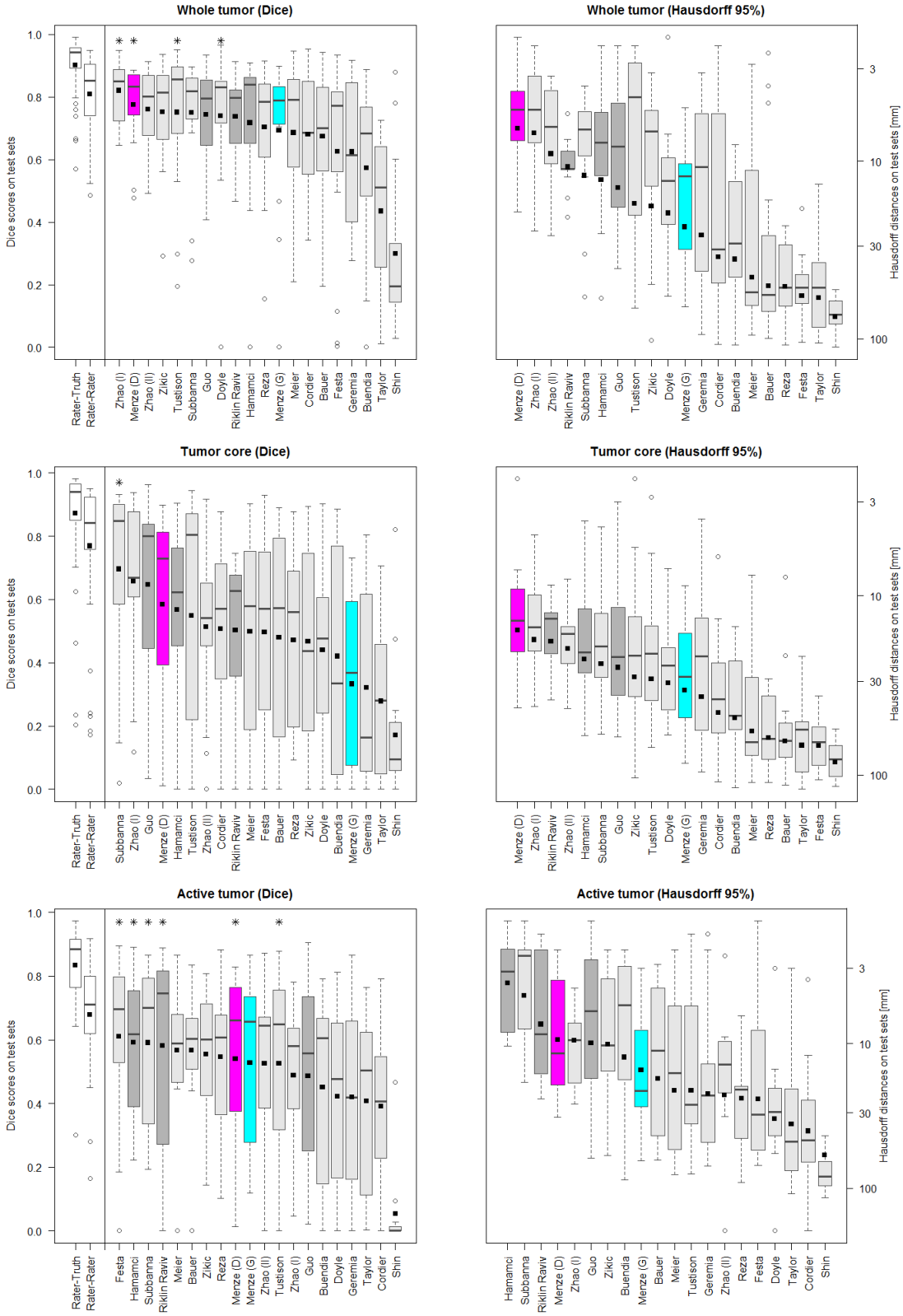

Fig. 8. Results from the BRATS evaluation reporting Dice scores and Hausdorff distances for the 'off-site' test (see Figure 8 in [1]). Methods are ranked according to the average Dice score and the robust Hausdorff distance and boxplots indicate quartiles and outliers. Results of the generative model with the two discriminative model extensions are shown in magenta and cyan, the first corresponding to results of the voxel-wise discriminative-model (*Menze (D)*) and those that have been generated by removing false positive regions from the segmentations of the generative model (*Menze (G)*). Competing inter-active segmentation methods are indicated by dark gray boxes. White boxplots report the inter-rater Dice scores, with scores calculated between individual raters (*Rater-Rater*) and between the consensus and raters (*Rater-Truth*). Stars on top of the boxes indicate methods with results that do not differ significantly from the inter-rater variation ($p<.05$). Also refer to the BRATS evaluation paper [1] for additional details (http://dx.doi.org/10.1109/TMI.2014.2377694).