Comparative Validation of Graphical Models for Learning Tumor Segmentations From Noisy Manual Annotations

Frederik O. Kaster[1,2], Bjoern H. Menze[3,4], and Marc-André Weber[5] and Fred A. Hamprecht[1]

[1] Heidelberg Collaboratory for Image Processing, University of Heidelberg, Germany
frederik.kaster@iwr.uni-heidelberg.de
fred.hamprecht@iwr.uni-heidelberg.de
[2] German Cancer Research Center, Heidelberg, Germany
[3] CSAIL, Massachusetts Institute of Technology, Cambridge MA, USA
menze@csail.mit.edu
[4] INRIA Sophia-Antipolis Méditerrannée, France
[5] Department of Diagnostic Radiology, University of Heidelberg, Germany
MarcAndre.Weber@med.uni-heidelberg.de

Abstract. Classification-based approaches for segmenting medical images commonly suffer from missing ground truth: often one has to resort to manual labelings by human experts, which may show considerable intra-rater and inter-rater variability. We experimentally evaluate several latent class and latent score models for tumor classification based on manual segmentations of different quality, using approximate variational techniques for inference. For the first time, we also study models that make use of image feature information on this specific task. Additionally, we analyze the outcome of hybrid techniques formed by combining aspects of different models. Benchmarking results on simulated MR images of brain tumors are presented: while simple baseline techniques already gave very competitive performance, significant improvements could be made by explicitly accounting for rater quality. Furthermore, we point out the transfer of these models to the task of fusing manual tumor segmentations derived from different imaging modalities on real-world data.

1 Introduction and related work

The use of machine learning methods for computer-assisted radiological diagnostics faces a common problem: In most situations, it is impossible to obtain reliable ground-truth information for e.g. the location of a tumor in the images. Instead one has to resort to manual segmentations by human labelers, which are necessarily imperfect due to two reasons. Firstly, humans make labeling mistakes due to insufficient knowledge or lack of time. Secondly, the medical images upon which they base their judgment may not have sufficient contrast to discriminate between tumor and non-tumor tissue. In general, this causes both a systematic

bias (tumor outlines are consistently too large or small) and a stochastic fluctuation of the manual segmentations, both of which depend on the specific labeler and the specific imaging modality.

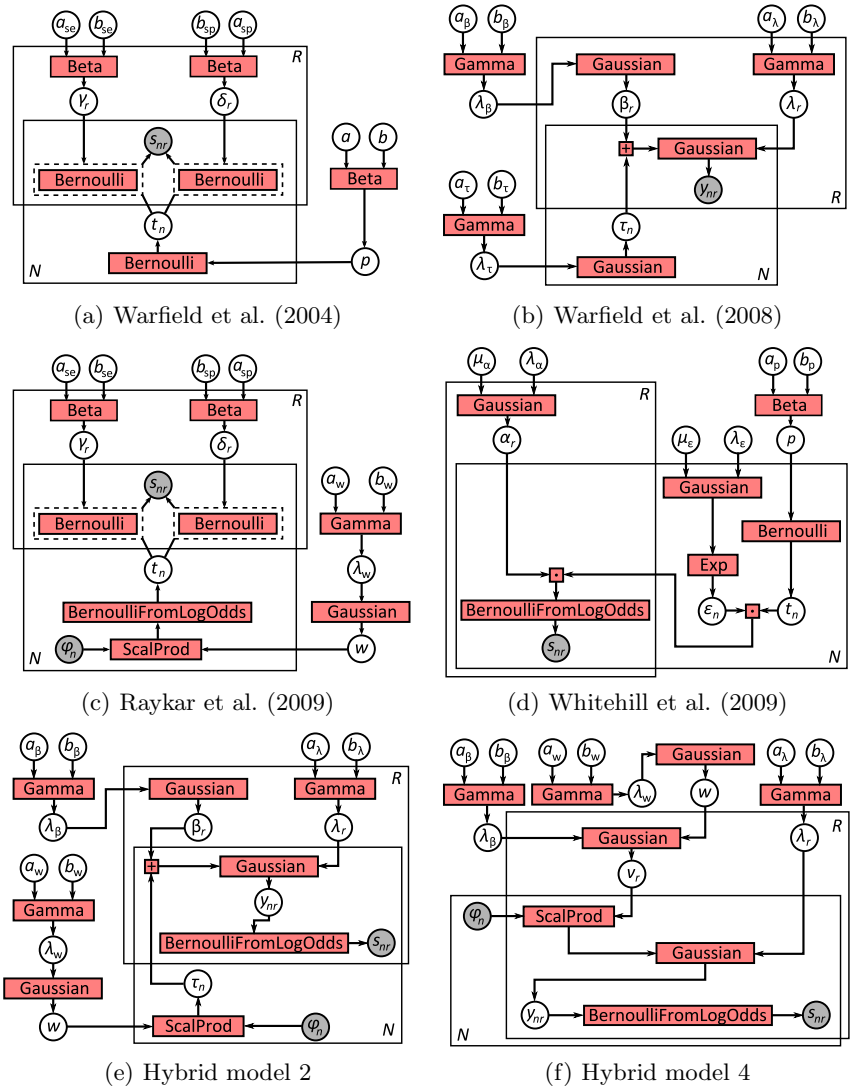
One can alleviate this problem by explicitly modelling the decision process of the human raters: in medical image analysis, this line of research started with the STAPLE algorithm (Warfield et al., 2004) and its extensions (Warfield et al., 2008), while in the field of general computer vision, it can already be traced back to the work of Smyth et al. (1995). Similar models were developed in other application areas of machine learning (Raykar et al., 2009; Whitehill et al., 2009; Rogers et al., 2010): some of them make also use of image information and produce a classifier, which may be applied to images for which no annotations are available. The effect of the different imaging modalities on the segmentation has not yet found as much attention.

In this paper, we systematically evaluated these competing methods as well as novel hybrid models for the task of computer-assisted tumor segmentation in radiological images: we used the same machinery on annotations provided by multiple human labelers with different quality and on annotations based on multiple imaging modalities. While traditionally these methods have been tackled by expectation maximization (EM; Dempster et al., 1977), we formulate the underlying inference problems as probabilistic graphical models (Koller and Friedman, 2009) and thereby render them amenable to generic inference methods (see Fig. 1). This facilitates the inference process and makes it easier to study the effect of modifications on the final inference results.

2 Theory and Modelling

Previous models In the following we detail on earlier and novel probabilistic models studied in the present work. In the STAPLE model proposed by Warfield et al. (2004, Fig. 1(a)), the discrete observations $s_{nr} \in \{0, 1\}$ are noisy views on the true scores $t_n \in \{0, 1\}$, with $n \in \{1, \ldots, N\}$ indexing the image pixels and $r \in \{1, \ldots, R\}$ indexing the raters. The r-th rater is characterized by the sensitivity γ_r and the specificity $1-\delta_r$, and the observation model is $s_{nr} \sim t_n \mathrm{Ber}(\gamma_r) + (1 - t_n)\mathrm{Ber}(\delta_r)$, with "Ber" denoting a Bernoulli distribution. A Bernoulli prior is given for the true class: $t_n \sim \mathrm{Ber}(p)$. While the original formulation fixes $p = 0.5$ and uses uniform priors for γ_r and δ_r, we modify the priors to fulfil the conjugacy requirements for the chosen variational inference techniques: hence we impose beta priors on $\gamma_r \sim \mathrm{Beta}(a_{\mathrm{se}}, b_{\mathrm{se}})$, $\delta_r \sim \mathrm{Beta}(b_{\mathrm{sp}}, a_{\mathrm{sp}})$ and $p \sim \mathrm{Beta}(a_{\mathrm{p}}, b_{\mathrm{p}})$. The latter distribution is introduced in order to learn the share of tumor tissue among all voxels from the data.

The model by Raykar et al. (2009, Fig. 1(c)) is the same as (Warfield et al., 2004) except for the prior on t_n: the authors now assume that a feature vector ϕ_n is observed at the n-th pixel and that $t_n \sim \mathrm{Ber}\big(\{1+\exp(-w^\top \phi_n)\}^{-1}\big)$ follows a logistic regression model. A Gaussian prior is imposed on $w \sim \mathcal{N}(0, \lambda_{\mathrm{w}}^{-1} I)$. In contrast to (Warfield et al., 2004), they obtain a classifier that can be used to predict the tumor probability on unseen test images, for which one has access



(a) Warfield et al. (2004)

(b) Warfield et al. (2008)

(c) Raykar et al. (2009)

(d) Whitehill et al. (2009)

(e) Hybrid model 2

(f) Hybrid model 4

Fig. 1. Graphical model representations. Red boxes correspond to factors, circles correspond to observed (gray) and unobserved (white) variables. Solid black rectangles are plates indicating an indexed array of variables (Buntine, 1994). The dashed rectangles are *"gates"* denoting a mixture model with a hidden selector variable (Minka and Winn, 2009).

to the features ϕ_n but not to the annotations s_{nr}. One may hypothesize that the additional information of the features ϕ_n can help to resolve conflicts: in a two-rater scenario, one can decide that the rater has less noise who labels pixels

with similar ϕ_n more consistently. In our graphical model formulation, we add a gamma prior for the weight precision $\lambda_w \sim \text{Gam}(a_w, b_w)$.

Whitehill et al. (2009, Fig. 1(d)) propose a model, in which the misclassification probability depends on both the pixel and the rater: $s_{nr} \sim \text{Ber}\big(\{1 + \exp(-t_n \alpha_r \epsilon_n)\}^{-1}\big)$ with the rater accuracy $\alpha_r \sim \mathcal{N}(\mu_\alpha, \lambda_\alpha^{-1})$ and the pixel difficulty ϵ_n with $\log(\epsilon_n) \sim \mathcal{N}(\mu_\epsilon, \lambda_\epsilon^{-1})$ (this parameterization is chosen to constrain ϵ_n to be positive).

In the continuous variant of STAPLE by Warfield et al. (2008, Fig. 1(b)), the observations y_{nr} are continuous views on a continuous latent score τ_n. The r-th rater can be characterized by a bias β_r and a noise precision λ_r: $y_{nr} \sim \mathcal{N}(\tau_n + \beta_r, \lambda_r^{-1})$, with a Gaussian prior on the true scores: $\tau_n \sim \mathcal{N}(0, \lambda_\tau^{-1})$. In contrast to the original formulation, we add Gaussian priors on the biases, i.e. $\beta_r \sim \mathcal{N}(0, \lambda_\beta^{-1})$. For the precisions of the Gaussians, we use gamma priors: $\lambda_\tau \sim \text{Gam}(a_\tau, b_\tau)$, $\lambda_\beta \sim \text{Gam}(a_\beta, b_\beta)$ and $\lambda_r \sim \text{Gam}(a_\lambda, b_\lambda)$. Note that when thresholding the continuous scores, the tumor boundary may shift because of the noise, but misclassifications far away from the boundary are unlikely: this is an alternative to (Whitehill et al., 2009) for achieving a non-uniform noise model.

Novel hybrid models We also study four novel hybrid models, which incorporate all aspects of the previous proposals simultaneously: while they provide a classifier as in (Raykar et al., 2009), they do not assume misclassifications to occur everywhere equally likely. In the simplest variant (hybrid model 1), we modify the model from (Warfield et al., 2008) by a linear regression model for $\tau_n \sim \mathcal{N}(w^\top \phi_n, \lambda_\tau^{-1})$ with $w \sim \mathcal{N}(0, \lambda_w^{-1})$. Note that this model predicts a (noisy) linear relationship between the distance transform values y_{nr} and the features ϕ_n, while experimentally the local image appearance saturates in the interior of the tumor or the healthy tissue. To alleviate this concern (hybrid model 2, Fig. 1(e)), one can interpret y_{nr} as an unobserved malignancy score, which influences the (observed) binary segmentations s_{nr} via $s_{nr} \sim \text{Ber}\big(\{1 + \exp(-y_{nr})\}^{-1}\big)$. This is a simplified version of the procedure presented in Rogers et al. (2010), with a linear regression model for the latent score instead of a Gaussian process regression. Alternatively one can model the raters as using a biased weight vector rather than having a biased view on an ideal score, i.e. $y_{rn} \sim \mathcal{N}(v_r^\top \phi_n, \lambda_r^{-1})$ with $v_r \sim \mathcal{N}(w, \lambda_\beta^{-1} I)$. Again the score y_{nr} may be observed directly as a distance transform (hybrid model 3) or indirectly via s_{nr} (hybrid model 4, Fig. 1(f)).

Inference For the graphical models considered here, exact inference by the junction tree algorithm is infeasible owing to the high number of variables and the high number of V structures, which lead to a nearly complete graph after moralization (Koller and Friedman, 2009). However, one can perform approximate inference using e.g. variational message passing (Winn and Bishop, 2005): the true posterior for the latent variables is approximated by the closest factorizing distribution (as measured by the Kullback-Leibler distance), for which inference is tractable. As a prerequisite, all priors must be conjugate; this holds for all models discussed above except (Whitehill et al., 2009). Here we cannot apply

the generic variational message passing scheme to this model, and show the results from the EM inference algorithm provided by the authors instead.

We employed the INFER.NET 2.3 Beta implementation for variational message passing (Minka et al., 2009) to perform inference on the algorithms by Warfield et al. (2004), Warfield et al. (2008), Raykar et al. (2009) and the four hybrid models. The default value of 50 iteration steps was found to be sufficient for convergence, since doubling the number of steps led to virtually indistinguishable results. For the algorithm by Whitehill et al. (2009), we used the GLAD 1.0.2 reference implementation.[6] Alternative choices for the generic inference method would have been expectation propagation (Minka, 2001) and Gibbs sampling (Gelfand and Smith, 1990). We experimentally found out that expectation propagation had considerably higher memory requirements than variational message passing for our problems, which prevented its use for our problems on the available hardware. Gibbs sampling was not employed since some of the factors incorporated in our models (namely gates and factor arrays) are not supported by the current INFER.NET implementation.

We also compared against three baseline procedures: majority voting, training a logistic regression classifier from the segmentations of every single rater and averaging the classifier predictions (ALR), and training a logistic regression classifier on soft labels (LRS): if S out of R raters voted for tumor in a certain pixel, it was assigned the soft label $S/R \in [0, 1]$.

3 Experiments

We performed two experiments in order to study the influences of labeler quality and imaging modality separately. In the first experiment, we collected and fused multiple human annotations of varying quality based on one single imaging modality: here we used simulated brain tumor measurements for which ground truth information about the true tumor extent was available, so that the results could be evaluated quantitatively. In the second experiment, we collected and fused multiple human annotations, which were all of high quality but had been derived from different imaging modalities showing similar physical changes caused by glioma infiltration with different sensitivity.

Human raters Simulated brain tumor MR images were generated by means of the TumorSim 1.0 software (Prastawa et al., 2009).[7] The advantage of these simulations was the existence of ground truth about the true tumor extent (in form of probability maps for the distribution of white matter, gray matter, cerebrospinal fluid, tumor and edema). Our task was to discriminate between "pathological tissue" (tumor and edema) and "healthy tissue" (the rest). We used nine volumes: three for each tumor class that can be simulated by this software (ring-enhancing, uniformly enhancing and non-enhancing). Each volumetric images contained $256 \times 256 \times 181$ voxels and the three different imaging modalities

[6] `http://mplab.ucsd.edu/~jake/OptimalLabelingRelease1.0.2.tar.gz`
[7] `http://www.sci.utah.edu/releases/tumorsim_v1.0/TumorSim_1.0_linux64.zip`

(T_1-weighted with and without gadolinium enhancement and T_2-weighted) were considered perfectly registered with respect to each other. The feature vectors ϕ_i consisted of four features for each modality: gray value, gradient magnitude and the responses of a minimum and maximum filter within a 3×3 neighborhood. A row with the constant value 1 was added to learn a constant offset for the linear or logistic models (since there was no reason to assume that features values at the tumor boundary are orthogonal to the final weight vector).

The image volumes were segmented manually based on hypointensities in the T_1-weighted images, using the manual segmentation functionality of the ITK-SNAP 2.0 software.[8] In order to control the rater precision, time limits of 60, 90, 120 and 180 seconds for labeling a 3D volume were imposed and five segmentations were created for each limit: we expect the segmentations to be precise for generous time limits, and to be noisy when the rater had to label very fast. The set of raters was the same for the different time constraints, and the other experimental conditions were also kept constant across the different time constraints. This was statistically validated: the area under curve value of the receiver operating characteristic of the ground-truth probability maps compared against the manual segmentations showed a significant positive trend with respect to the available time ($p = 1.8 \times 10^{-4}$, F test for a linear regression model). Since tight time constraints are typical for the clinical routine, we consider this setting as realistic, although it does not account for rater bias.

We extracted the slices with the highest amount of tumor lesion, and partitioned them into nine data subsets in order to estimate the variance of segmentation quality measures, with each subset containing one third of the slices extracted from three different tumor datasets (one for each enhancement type). For memory reasons, the pixels labeled as "background" by all raters were randomly subsampled to reduce the sample size. A cross-validation scheme was used to test the linear and log-linear classifiers (all except Warfield et al. (2004, 2008); Whitehill et al. (2009)) on features ϕ_n not seen during the training process: we repeated the training and testing nine times and chose each of the data subsets in turn as the training dataset (and two different subsets as the test data).

The following default values for the hyperparameters were used: $a_{\mathrm{Se}} = 10$, $b_{\mathrm{Se}} = 2$, $a_{\mathrm{Sp}} = 10$, $b_{\mathrm{Sp}} = 2$, $a_{\mathrm{w}} = 2$, $b_{\mathrm{w}} = 1$, $a_{\mathrm{p}} = 2$, $b_{\mathrm{p}} = 2$, $a_\tau = 2$, $b_\tau = 1$, $a_\beta = 2$, $b_\beta = 1$, $a_\lambda = 2$, $b_\lambda = 1$. We confirmed in additional experiments that inference results changed only negligibly when these hyperparameters were varied over the range of a decade. In order to check the effect of the additional priors that we introduced into the models of Warfield et al. (2004), Warfield et al. (2008) and Raykar et al. (2009), we also ran experiments with exactly the same models as in the original papers (by fixing the corresponding variables or using uniform priors). However, this led to uniformly worse inference results than in our model formulations.

Multiple modalities For evaluation on real-world measurements, we used a set of twelve multimodal MR volumes acquired from glioma patients (T_1-, T_2-,

[8] `http://www.itksnap.org/pmwiki/pmwiki.php?n=Main.Downloads`

FLAIR- and post-gadolinium T_1-weighting), which had been affinely registered to the FLAIR volume: we used a automated multi-resolution mutual information registration procedure as included in the MedINRIA[9] software. Manual segmentations of pathological tissue (tumor and edema) were provided separately for every modality on 60 slices extracted from these volumes (20 axial, sagittal and coronal slices each of which intersecting with the tumor center). In these experiments, we propose to use the described models to infer a single probability map summarizing all tumor-induced changes in the different imaging modalities. In particular, we identify every modality as a separate "rater" with a specific and consistent bias with respect to the joint probability map inferred.

4 Results

	Specificity	Sensitivity	CCR	AUC	Dice
Majority vote	.987(007)	.882(051)	.910(032)	.972(008)	.827(020)
ALR	.953(018)	*.920(036)*	*.931(025)*	.981(005)	*.855(031)*
LRS	.953(019)	.919(037)	*.931(025)*	.981(005)	*.855(030)*
Warfield et al. (2004)	.987(007)	.882(051)	.910(032)	.972(008)	.827(020)
Warfield et al. (2008)	***1.000(001)***	.617(130)	.692(139)	**.989(003)**	.584(211)
Raykar et al. (2009)	.988(006)	.886(045)	.913(028)	***.993(003)***	.830(024)
Whitehill et al. (2009)	.988(004)	.913(016)	*.931(008)*	.980(003)	.845(063)
Hybrid model 1	.940(078)	.692(060)	.751(070)	.902(117)	.603(191)
Hybrid model 2	.972(019)	.716(048)	.770(057)	.953(015)	.628(163)

Table 1. Evaluation statistics for the training data (i.e. the manual annotations of the raters were used for inference), under the 120/120/90 scenario. The first three rows show the outcome of the three baseline techniques. The best result in each column is marked *in italics*, while **bold figures** indicate a significant improvement over the best baseline technique ($P < .05$, rank-sum test with multiple-comparison adjustment). Estimated standard deviations are given in parentheses. The outcome of the other scenarios was qualitatively similar (especially concerning the relative ranking between different inference methods). ALR = Averaged logistic regression. LRS = Logistic regression with soft labels. CCR = Correct classification rate (percentage of correctly classified pixels). AUC = Area Under Curve of the receiver operating characteristics curve obtained when thresholding the ground-truth probability map at 0.5. Dice = Dice coefficient of the segmentations obtained when thresholding both the inferred and the ground-truth probability map at 0.5.

Multiple raters We studied several scenarios, i.e. several compositions of the rating committee. Here we exemplarily report the results for two of them: one with a majority of good raters (120/120/90, i.e. two raters with a 120 sec constraint

[9] `https://gforge.inria.fr/projects/medinria`

	Sensitivity	Specificity	CCR	AUC	Dice
ALR	.937(017)	.924(038)	.928(029)	*.978(009)*	.837(065)
LRS	.936(017)	.925(038)	.928(029)	*.978(009)*	.837(066)
Raykar et al. (2009)	.927(019)	***.937(031)***	*.936(025)*	.977(013)	*.853(038)*
Hybrid model 1	.851(152)	.735(181)	.760(167)	.852(172)	.619(142)
Hybrid model 2	***.973(013)***	.727(174)	.786(116)	.952(026)	.667(084)

Table 2. Evaluation statistics for the test data (i.e. the manual annotations of the raters were not used for inference), under the 120/120/90 scenario. Note that one can only employ the inference methods which make use of the image features ϕ_n and estimate a weight vector w: the unobserved test data labels are then treated as missing values and are marginalized over. All methods which only use the manual annotations (majority voting, Warfield et al. (2004) and Warfield et al. (2008)) cannot be applied to these examples. The results for the other scenarios were qualitatively similar (especially concerning the relative ranking between different inference methods). Cf. the caption of table 1 for further details.

and one rater with a 90 sec constraint) and one with a majority of poor raters (60/60/60/180/180, i.e. three raters with a 60 sec constraint, and two raters with a 180 sec constraint). Tables 1 and 2 show the results of various evaluation statistics both for training data (for which the human annotations were used) and test data. Sensitivity, specificity, correct classification rate (CCR) and Dice coefficient are computed from the binary images that are obtained by thresholding both the ground-truth probability map and the inferred posterior probability map at 0.5. If n_{fb} denotes the number of pixels that are thereby classified as foreground (tumor) in the ground truth and as background in the posterior probability map (and n_{bb}, n_{bf} and n_{ff} are defined likewise), these statistics are computed as follows:

$$\text{Sensitivity} = \frac{n_{\mathrm{ff}}}{n_{\mathrm{fb}} + n_{\mathrm{ff}}}, \qquad \text{Specificity} = \frac{n_{\mathrm{bb}}}{n_{\mathrm{fb}} + n_{\mathrm{bb}}},$$
$$\text{CCR} = \frac{n_{\mathrm{ff}} + n_{\mathrm{bb}}}{n_{\mathrm{ff}} + n_{\mathrm{bb}} + n_{\mathrm{bf}} + n_{\mathrm{fb}}}, \qquad \text{Dice} = \frac{2n_{\mathrm{ff}}}{2n_{\mathrm{ff}} + n_{\mathrm{bf}} + n_{\mathrm{fb}}}$$

Additionally we report the Area Under Curve (AUC) value for the receiver operating curve obtained by binarizing the ground-truth probabilities with a fixed threshold of 0.5 and plotting sensitivity against $1 -$ specificity while the threshold for the posterior probability map is swept from 0 to 1. Most methods achieved Dice coefficients in the range of 0.8–0.85, except for the models operating on a continuous score (the hybrid models and the model by Warfield et al. (2008)). Since our features were highly discriminative, even simple label fusion schemes such as majority voting gave highly competitive results. Qualitatively, there is little difference between these two scenarios (and the other ones under study). While some graphical models perform better than the baseline methods on the training data (namely (Raykar et al., 2009) and (Warfield et al., 2008)), they bring no improvement on the test data.

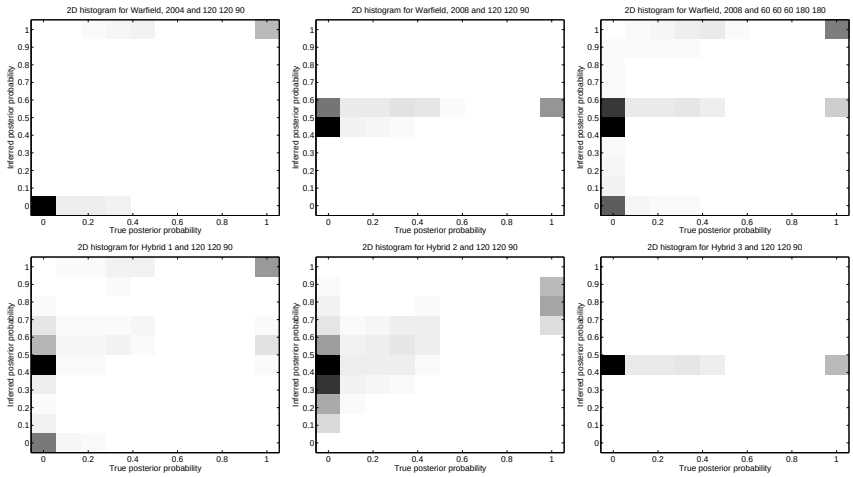


Fig. 2. Comparison of ground-truth (abscissa) and inferred posterior (ordinate) tumor probabilities, visualized as normalized 2D histograms. All histograms are normalized such that empty bins are white, and the most populated bin is drawn black. We show the inference results of (Warfield et al., 2004), (Warfield et al., 2008), and the hybrid models 1–3. The results of hybrid model 4 were similar to hybrid model 3, and the results of (Raykar et al., 2009) and (Whitehill et al., 2009) were similar to (Warfield et al., 2004). Mostly the two scenarios 120/120/90 and 60/60/60/180/180 gave similar results so that we show only the results for the former, with the exception of (Warfield et al., 2008) (top middle and top right). For the ideal inference method, all bins outside the main diagonal would be white; Warfield et al. (2004) comes closest.

Unexpectedly, the hybrid models perform worse and with lesser stability than the simple graphical models, and for hybrid models 3 and 4, the inference converges to a noninformative posterior probability of 0.5 everywhere. It should be noted that the posterior estimates of the rater properties did not differ considerably between corresponding algorithms such as (Warfield et al., 2008) and (Raykar et al., 2009), hence the usage of image features does not allow one to distinguish between better and poorer raters more robustly.

In order to account for partial volume effects and blurred boundaries between tumor and healthy tissue, it is preferable to visualize the tumors as soft probability maps rather than as crisp segmentations. In Fig. 2, we compare the ground-truth tumor probabilities with the posterior probabilities following from the different models. The models assuming a latent binary class label ((Warfield et al., 2004), (Raykar et al., 2009) and (Whitehill et al., 2009)) tend to sharpen the boundaries between tumor and healthy tissue overly, while the latent score models (all others) smooth them. One can again note that the true and inferred probabilities are completely uncorrelated for hybrid model 3 (and 4).

Multiple modalities The optimal delineation of tumor borders in multi-modal image sequences and obtaining ground truth remains difficult. So, in the present

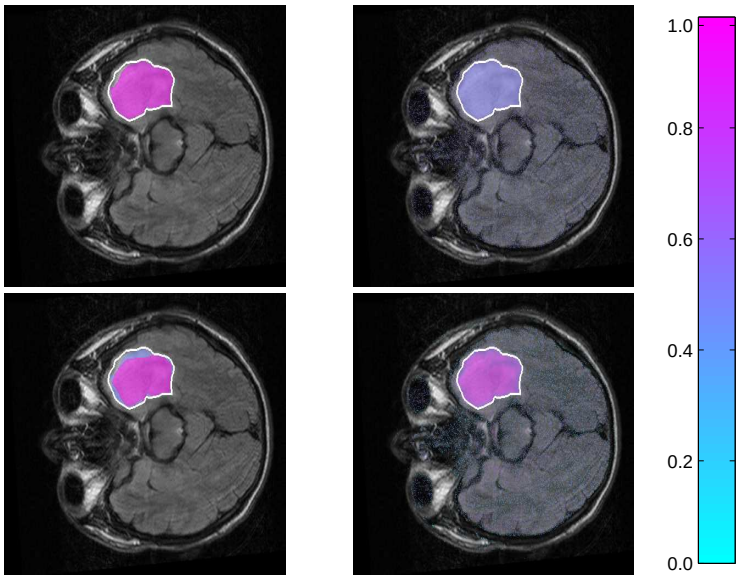


Fig. 3. Example of a FLAIR slice with manual segmentation of tumor drawn on the same FLAIR image (white contour), and inferred mean posterior tumor probability maps for (Warfield et al., 2004) (top left), Warfield et al. (2008) (top right), (Whitehill et al., 2009) (bottom left) and hybrid model 2 (bottom right). The results of hybrid model 3 and 4 were nearly identical to (Warfield et al., 2008), the results of hybrid model 1 to model 2, and the results of (Raykar et al., 2009) to (Whitehill et al., 2009). Pixels with $P_{\text{tumor}} < .01$ are masked out. The "noisy" appearance of the two right-hand side maps is due to the random subsampling of background pixels. We recommend to view this figure in the colored online version.

study we confine ourselves to a first, qualitative comparison of the different models. Fig. 3 shows the posterior probability maps for a real-world brain image example. The results of (Warfield et al., 2004) and (Warfield et al., 2008) can be regarded as extreme cases: the former yields a crisp segmentation without accounting for uncertainty near the tumor borders, while the latter assigns a probability near 0.5 to all pixels and is hence inappropriate for this task. Hybrid model 1 (or 2) and (Whitehill et al., 2009) or (Raykar et al., 2009) are better suited for the visualization of uncertainties.

5 Discussion and Outlook

In this study, we introduced graphical model formulations to the task of fusing noisy manual segmentations: e.g. the model by Raykar et al. (2009) had not been previously employed in this context, and it was found to improve upon simple logistic regression on the training data. However, these graphical models do not always have an advantage over simple baseline techniques: compare the

results of (Warfield et al., 2004) to majority voting. Hybrid models combining the aspects of several models did not fare better than simple models. This ran contrary to our initial expectations, which were based on two assumptions: that different pixels have a different probability of being mislabeled, and that it is possible to detect these pixels based on the visual content (these pixels would be assigned high scores far away from the decision boundary). This may be an artifact of our time-constrained labeling experiment: if misclassifications can be attributed mostly to chance or carelessness rather than to ignorance or visual ambiguity, these assumptions obviously do not hold, and a uniform noise model as in (Warfield et al., 2004) or (Raykar et al., 2009) should be used instead. It is furthermore not yet understood why the slight model change between hybrid models 1 / 2 and hybrid models 3 / 4 leads to the observed failure of inference. For the future, it should be checked if these effects arise from the use of an approximate inference engine or are inherent to these models: hence unbiased Gibbs sampling results should be obtained for comparison purposes, using e.g. the WinBUGS modelling environment (Lunn et al., 2000).

The use of simulated data for the main evaluation is the main limitation of our approach, as simulations always present a simplification of reality and cannot account for all artifacts and other causes for image ambiguity that are encountered in real-world data. However, this limitation is practically unavoidable, since we are assessing the imperfections of the currently best clinical practice for the precise delineation of brain tumors, namely manual segmentation of MR images by human experts. This assessment requires a superior gold standard by which the human annotations may be judged, and this can only be obtained from an *in silico* ground truth. For animal studies, a possible alternative lies in sacrificing the animals and delineating the tumor on histological slices which can be examined with better spatial resolution. However, these kinds of studies are costly and raise ethical concerns. Additionally, even expert pathologists often differ considerably in their assessment of histological images (Giannini et al., 2001).

Better segmentations could presumably be achieved by two extensions: More informative features could be obtained by registration of the patient images to a brain atlas, e.g. in the spirit of Schmidt et al. (2005). An explicit spatial regularization could be achieved by adding an MRF prior on the latent labels or scores, and employing a mean-field approximation (Zhang, 1992) to jointly estimate the optimum segmentation and the model parameters by EM.

References

Buntine, W.: Operations for Learning with Graphical Models. Journal of Artificial Intelligence Research 2, 159–225 (1994)

Dempster, A., Laird, N., Rubin, D., et al.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)

Gelfand, A.E., Smith, A.F.: Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association 85(410), 398–409 (1990)

Giannini, C., Scheithauer, B., Weaver, A., et al.: Oligodendrogliomas: reproducibility and prognostic value of histologic diagnosis and grading. Journal of Neuropathology & Experimental Neurology 60(3), 248 (2001)
Koller, D., Friedman, N.: Probabilistic Graphical Models – Principles and Techniques. MIT Press (2009)
Lunn, D., Thomas, A., Best, N., et al.: WinBUGS – A Bayesian modelling framework: Concepts, structure and extensibility. Statistics and Computing 10, 325–337 (2000)
Minka, T.: Expectation Propagation for approximate Bayesian inference. In: Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, 362 – 369 (2001)
Minka, T., Winn, J., Guiver, J., et al.: Infer.NET 2.3 (2009). Microsoft Research Cambridge. `http://research.microsoft.com/infernet`
Minka, T., Winn, J.: Gates. In: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (eds.), Advances in Neural Information Processing Systems 21, 1073–1080. MIT Press (2009)
Prastawa, M., Bullitt, E., Gerig, G.: Simulation of Brain Tumors in MR Images for Evaluation of Segmentation Efficacy. Medical Image Analysis 13(2), 297–311 (2009)
Raykar, V., Yu, S., Zhao, L., et al.: Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit. In: 26th International Conference on Machine Learning (ICML) (2009)
Rogers, S., Girolami, M., Polajnar, T.: Semi-parametric analysis of multi-rater data. Statistics and Computing in press (2010)
Schmidt, M., Levner, I., Greiner, R., et al.: Segmenting Brain Tumors using Alignment-Based Features. In: Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA), 215–220 (2005)
Smyth, P., Fayyad, U., Burl, M., et al.: Inferring Ground Truth From Subjective Labelling of Venus Images. In: G. Tesauro, D. Toretzy, T. Leen (eds.), Advances in Neural Information Processing Systems 7, 1085–1092. MIT Press (1995)
Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23(7), 903–921 (2004)
Warfield, S., Zou, K., Wells, W.: Validation of image segmentation by estimating rater bias and variance. Philosophical Transactions of the Royal Society A 366(1874), 2361–2375 (2008)
Whitehill, J., Ruvolo, P., fan Wu, T., et al.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In: Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, A. Culotta (eds.), Advances in Neural Information Processing Systems 22, 2035–2043. MIT Press (2009)
Winn, J., Bishop, C.: Variational Message Passing. Journal of Machine Learning Research 6, 661–694 (2005)
Zhang, J.: The mean field theory in EM procedures for Markov random fields. IEEE Transactions on Signal Processing 40(10), 2570–2583 (1992)