2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

MODEL-DEPENDENT UNCERTAINTY ESTIMATION OF MEDICAL IMAGE SEGMENTATION

Tsachi Hershkovitch and Tammy Riklin-Raviv

Electrical and Computer Engineering Department
Zlotowski Center for Neuroscience
Ben-Gurion University, Beer-Sheva, Israel

ABSTRACT

Segmentation is a prevalent research area in medical imaging analysis. Nevertheless, estimation of the uncertainty margins of the extracted anatomical structure or pathology boundaries is seldom considered. This paper studies the concept of segmentation uncertainty of clinical images, acknowledging its great importance to patient follow up, user-interaction guidance, and morphology-based population studies.

We propose a novel approach for model-dependent uncertainty estimation for image segmentation. The key contribution is an alternating, iterative algorithm for the generation of an image-specific *uncertainty map*. This is accomplished by defining a *consistency-based measure* and applying it to segmentation samples to estimate the uncertainty margins as well as the *midline* segmentation. We utilize the stochastic active contour framework as our segmentation generator, yet any sampling method can be applied. The method is validated on synthetic data for well-defined objects blurred with known Gaussian kernels. Further assessment of the method is provided by an application of the proposed consistency-based algorithm to ensembles of stochastic segmentations of brain hemorrhage in CT scans.

Index Terms— Segmentation uncertainty, stochastic active contours, brain hemorrhage in CT scans

1. INTRODUCTION

Medical image segmentation is of great significance for clinical research as well as therapy planning and guidance. Nevertheless, weak contrast, low signal to noise ratio (SNR) and imaging artifacts often turn the task of region-of-interest (ROI) extraction into a difficult one, even for a human rater. While inter-rater variability is a well-discussed issue, automatic segmentation algorithms are often considered (unjustifiably) "objective" and repeatable. Indeed, a specific deterministic algorithm would produce the same output for the same input. However, in general, segmentation results are not only model dependent, but are also sensitive to changes in hyperparameters, initialization and stopping criteria.

This paper addresses the question of segmentation uncertainty in the context of automatic segmentation algorithms. Specifically, we propose a method that allows local and quantitative measure of the uncertainty margins of a given ROI. While we note that the estimated uncertainty margins are tightly related to the chosen generative segmentation model, their potential influence on a variety of medical-related applications should not be ignored. Potential applications include cross-sectional studies (e.g., comparison of a specific anatomical structure across a population); patient follow-up (e.g., testing tumor's growth or reaction to treatment) and diagnosis (e.g., malignancy-benignancy classification when done based on the diffusivity and roughness of the ROI's boundaries). Moreover, the ability to define the uncertainty margins can leverage semi-automatic, user-interaction platforms. The marked areas can be used to refer the user to provide input in regions with possible ambiguities.

Segmentation sampling plays a critical role in segmentation uncertainty estimation. Pioneered by Fan et al. [1], numerous algorithms use Markov Chain Monte Carlo (MCMC) methods to sample segmentations from a proposal distribution that is based on the presumed image model. To guarantee that the proposal distribution asymptotically converges to the true, unknown, distribution, the Metropolis-Hastings acceptance probability is used [2]. In [3, 4] a random maximum a posteriory (MAP) perturbation model for sampling from Gibbs distribution is presented. Recent work [5] exploits these frameworks for segmentation uncertainty estimation by using samples generated by perturbing the energy functional of a conditional random field (CRF), followed by MAP inference. In [6], segmentations are sampled using a Gaussian process, based on expert's manual annotation. In [7] a generative model is used for generating a probabilistic label map from a set of binary images. While most sampling approaches aim to estimate the underlying distribution for finding the MAP segmentation, fusion of the samples for the construction of an uncertainty map usually gets less attention.

Given a set of segmentations, the simplest approach

This study was partially supported by the Israel Science Foundation (1638/16 T.R.R), IDF Medical Corps (T.R.R.) and the Israel Ministry of Science and Technology fund for Engineering (T.R.R.)

would be to average them, yet at the price of being vulnerable to outliers. The STAPLE algorithm presents a plausible alternative [8]. Segmentations are weighted based on their specificity and sensitivity with respect to an iteratively constructed soft 'consensus' segmentation. Yet, as was shown in [9], the consensus segmentation may converge to a binary segmentation, thus providing little or no information on the variability of the generating samples. We note that other segmentation weighting methods, such as those used in the atlas fusion approaches [10], are not applicable for our purposes, as the weights are based on the associated gray-level images, while here we consider multiple segmentations of a single given image. In this sense, segmentation uncertainty can be viewed as an image-specific probabilistic-atlas.

We, hereby, introduce a novel approach for model-dependent uncertainty estimation for image segmentation. The key contribution is a consistency-based algorithm for the generation of an image-specific *uncertainty map*. While we use the stochastic active contour framework [11] as a segmentation generator, any segmentation sampling algorithm can be applied. The foundation of the proposed method is an alternating, iterative process in which the consistency score (weight) of each segmentation is repeatedly re-calculated, following re-estimation of the segmentation uncertainty map, constructed from a weighted sum of all the segmentations. The consistency score of a given segmentation is defined by its cosine similarity to the constructed uncertainty map. This measure, as opposed to the Dice coefficients [12], for example, enables using soft, rather than binary, segmentation values. Outliers, therefore have negligible consistency values, while highly consistent segmentations get higher weights. This also allows a representation of multiple distinct clusters of segmentations without having one overriding the others. Having an uncertainty map, one can generate the *midline* segmentation as well as its uncertainty margins; therefore, providing quantitative information on the contour's point-wise reliability as well as on regions with possibly high ambiguities.

As 'ground truth' uncertainty margins do not exist, we test our method using synthetic data for well-defined objects blurred with known Gaussian kernels. We then show that the width of the uncertainty margins is directly proportional to the widths of the blurring kernel. Validation is further provided by an application of the proposed consistency-based algorithm to ensembles of stochastic segmentations of cerebral hemorrhage (CH) in CT scans. Specifically, we show that the compatibility (Dice scores) between the midline segmentations and the expert's manual annotations is inversely proportional to the entropy of the respective consistency scores. In addition, compatibility between the point-wise uncertainties along the mid-line segmentation and the image contrast is visually demonstrated.

2. METHODS

The proposed segmentation uncertainty estimation is composed of two main stages. Given an image, we first use a stochastic active contour segmentation to produce a large set of plausible segmentations. We then use a consistency-based measure to score the segmentations and construct their weighted sum to allow uncertainty margins extraction.

2.1. Problem definition

Let $I\colon \Omega \to \mathbb{R}$ define a gray level image, where Ω is the image domain. Let $\omega^+ \in \Omega$ define a ROI in I. We note that the 'true' ROI boundaries, $\partial\omega^+$, are not well defined. However, we assume that we can sample segmentations $\Gamma\colon \Omega \to \{0,1\}$ from a posterior probability distribution $p(\Gamma|I,\theta)$, where θ denotes some pre-defined model parameters. In contrast to most segmentation approaches, we do not aim to find the MAP segmentation. Instead, given $\{\Gamma_m \in \{0,1\}\}_{m=1,\ldots,M}$ and two scalars: $\tau_H \in (0.5,1)$ and $\tau_L \in (0,0.5)$, our goal is to find a soft uncertainty map $S\colon \Omega \to [0,1]$ such that $\omega_u = \{\mathbf{x} \in \Omega | \tau_L < S(\mathbf{x}) < \tau_H\}$ denote 'plausible' uncertainty margins of ω^+.

2.2. Level-set formulation

Let ϕ_m denote a signed distance function (SDF) associated with the binary segmentation Γ_m, i.e., $\phi_m(\mathbf{x}) > 0$ for each and only each $\mathbf{x} \in \Omega$ such that $\Gamma(\mathbf{x}) = 1$. Adapting the log-odds formulation, as in [13], we use the logistic-regression function $\mathcal{H}(z) = (1+\exp(-z))^{-1}$ to obtain a soft segmentation $\tilde{\Gamma}_m \in [0,1]$, where $\tilde{\Gamma}_m(\mathbf{x}) = \mathcal{H}(\phi(\mathbf{x}))$. Following [14] we define a cost functional $\mathcal{E}(\phi) \propto p(\phi > 0|I,\theta)$ as follows:

$$\mathcal{E}(\phi) = \mathcal{E}_{IL}(\phi, I) + \mathcal{E}_{REG}(\phi), \tag{1}$$

where, $\mathcal{E}_{IL}(\phi, I)$ is an image-likelihood term and $\mathcal{E}_{REG}(\phi)$ a regularization term. Specifically, we use a mixture of Gaussians to model the intensity distribution, extending the classical Chan-Vese level-set framework [15]. Nevertheless, any other active contour formulation can be applied. Commonly, the optimal ϕ is estimated by minimizing $\mathcal{E}(\phi)$ through a series of GD steps, with step size Δt: $\phi_t = \phi_{t-1} + \Delta t \frac{\partial \phi_t}{\partial t}$, where $\frac{\partial \phi_t}{\partial t}$ is derived using the first variation of the functional in Eq. (1). Here, we introduce a zero-mean normally distributed 'noise' to the GD process, as suggested in the stochastic active contour framework of [11]:

$$\phi_t = \phi_{t-1} + \Delta t \frac{\partial \phi_t}{\partial t} + \sqrt{\Delta t}\mathcal{N}_t\left(0, \sigma^2\right). \tag{2}$$

The non-deterministic process in Eq. (2) converges to a different level-set function ϕ_m for every run. The resulting set of soft segmentations $\{\tilde{\Gamma}_m = \mathcal{H}(\phi_m)\}_{m=1}^M$ depends on the functional (Eq. (1)) attractors and their basins of attraction and, to some extent, on the hyper-parameters in Eq. (2).

2.3. Consistency measure and uncertainty map

We next define an iterative, alternating process. In every iteration k, we calculate the cosine similarity measure C_m^k of each segmentation $\tilde{\Gamma}_m$ with respect to the current estimate of the soft uncertainty map S^k. Let N define the number of image pixels (or voxels), then $\forall m \in 1, \ldots, M$:

$$S^{k+1}(\mathbf{x}_n) = \frac{\sum_{m=1,\ldots,M} C_m^k \tilde{\Gamma}_m(\mathbf{x}_n)}{\sum_{m=1,\ldots,M} C_m^k}, \quad \forall \mathbf{x}_n \in \Omega, \tag{3}$$

$$C_m^k = \frac{\sum_{n=1}^{N} S^k(\mathbf{x}_n)\tilde{\Gamma}_m(\mathbf{x}_n)}{\sqrt{\sum_{n=1}^{N}(S^k(\mathbf{x}_n))^2}\sqrt{\sum_{n=1}^{N}\left(\tilde{\Gamma}_m(\mathbf{x}_n)\right)^2}}. \tag{4}$$

The cosine similarity ranges between -1 and 1. Yet, as the nominator is always positive the consistency value C_m^k is a score between 0 and 1. Considering $\tilde{\Gamma}_m$ and S^k as two non-zero vectors in $\mathbb{R}^N$, zero consistency value indicates orthogonality while one indicates complete alignment. We set $C_m^0 = \frac{1}{M}$. The algorithm converges when $\sum_{n=1}^{N} \left|S^k(\mathbf{x}_n) - S^{k-1}(\mathbf{x}_n)\right|^2 < \epsilon$ where, ϵ is set to a small positive scalar.

3. EXPERIMENTS

We assess the proposed uncertainty estimation method using synthetic images blurred with different Gaussian kernels and real brain CT data of patients with CHs.

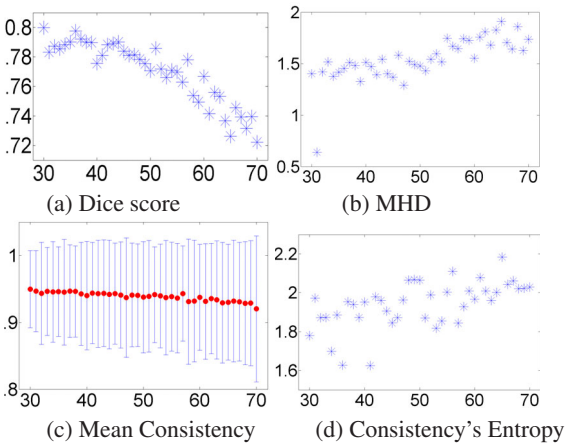
Synthetic data: We synthesized 40 images by sampling foreground (fg) and background (bg) intensities from two different normal distributions: $\mathcal{N}_{fg}(160, 60)$, $\mathcal{N}_{bg}(100, 60)$. Each image was then blurred with a different Gaussian kernel. For each image, 2000 segmentations were generated using stochastic active contour, and were used for the construction of the respective uncertainty map and uncertainty margins $\omega_u = \{\tau_L = 0.1 < S < \tau_H = 0.9\}$.

To quantitatively assess our method, we define:

$$S_H = \begin{cases} 1 & \text{if } S > \tau_H \\ 0 & \text{otherwise} \end{cases} \qquad S_L = \begin{cases} 1 & \text{if } S > \tau_L \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Dice scores and the Modified Hausdorf Distance (MHD) [16] scores were measured between S_L and S_H (Eq. 5). We also constructed a consistency score histogram (based on the 2000 segmentations) and calculated the mean, and the entropy of the consistency scores. Fig. 1 presents the Dice scores (a), the MHD (b), the mean (red) and STD (blue bars) of the consistency scores (c) and the entropy (d) of each image with respect to its blur kernel width. Note that the MHD measurements, STDs of the consistency scores, and entropy values are approximately directly proportional to the width of the blur kernels, while the means of the consistency scores and the Dice coefficients present the opposite trend - as expected.

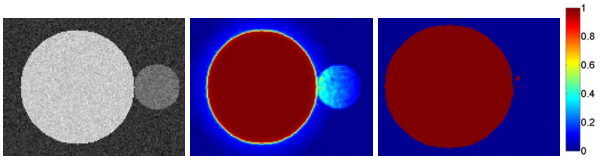
Fig. 2 presents a visual comparison between the proposed uncertainty map (middle) and the conditional probability of the true segmentation obtained by the STAPLE algorithm (right) with respect to the synthetic image shown on the right. When the number of segmentations is high, the STAPLE algorithm often convergences into a binary map rather than a voxelwise probability map, as could be expected. The reader is referred to the STAPLE's equations to verify it mathematically, for particular, yet, quite frequent, cases [8].



(a) Dice score (b) MHD

(c) Mean Consistency (d) Consistency's Entropy

Fig. 1. **Quantitative measurements based on the synthetic data**. Dice (a) and MHD (b) measures between the S_H and S_L. (c) Means and STDs of the consistency scores. (d) Entropy values of the consistency scores. All measurements were performed for each synthetic image and were plotted with respect to its blur kernel width.



Original Image Proposed STAPLE's result

Fig. 2. A visual comparison of the proposed uncertainty map (middle) corresponding to the synthetic image on the left to the conditional probability of the true segmentation obtained by the STAPLE algorithm (right).

Brain CT scans: We next present the performances of the proposed algorithm for CH segmentation. Scans of 19 stroke patients were acquired with Philips Brilliance CT 64 system without radiocontrast agents injection. The data resolution is $512\times512\times[90-100]$ with voxel size of $0.48\text{mm}\times0.48\text{mm}\times 3\text{mm}$, with $1.5\,[mm]$ overlap in the axial direction. For each scan, 2000 stochastic active contour segmentations were generated and were used for the construction of the scan's uncertainty map, uncertainty margins, and midline segmentation.

Fig. 3 presents four different examples of uncertainty estimation of CH segmentation in CT scans. Rows 1-2 present full size images. Rows 3-4 present zoom-in views in the regions marked by yellow rectangles in the full-size images. Rows 2 and 4 show the images along with the calculated un-

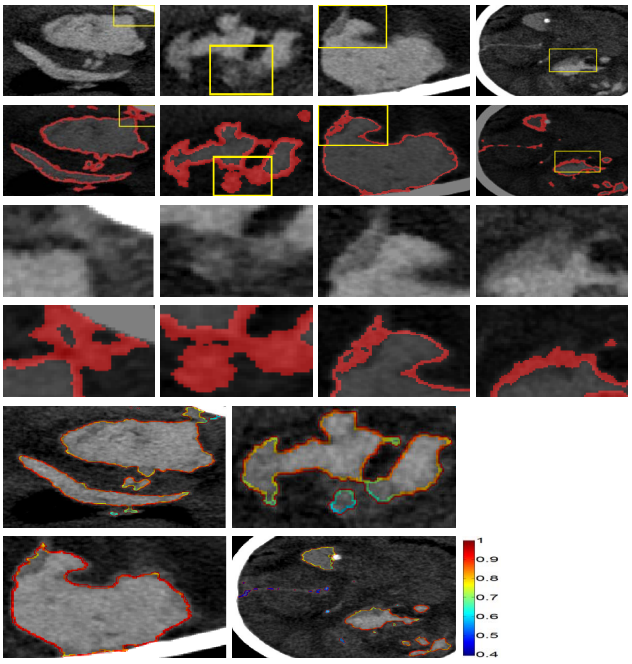
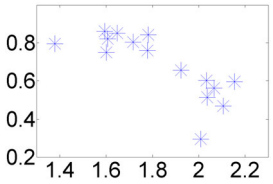


Fig. 3. **CH segmentation uncertainty.** Each column presents a different CT scan. Row 1: original images. Row 2: error margins. Rows 3-4: zoom-in of the yellow square region of the first rows, respectively. Rows 5-6: color-coded contour uncertainty.

Fig. 4. Dice scores for each of the 19 mid-line segmentations (based on expert's annotations) as a function of the corresponding consistency scores' entropy.



certainty margins. The fifth row presents the midline segmentation contour of each image color-coded. Hot and cold colors indicate high and low certainty, respectively. The point-wise certainty values are inversely proportional to the average minimal Euclidean distance between each point along the midline contour and the contours of S_H and S_L. Note that low certainty regions on the midline contour are usually located in low contrast image regions and vice versa. Fig. 4 presents the Dice scores, with respect to expert's annotations, of the 19 midline segmentations (each is associated with a different CT image) as a function of the respective consistency scores' entropy. The Dice scores are approximately inverse proportional to the entropy values, supporting the proposed measure.

4. CONCLUSIONS

We presented a consistency-based formulation for segmentation uncertainty estimation. Given a set of segmentation samples, the proposed algorithm provides the uncertainty margins and the midline segmentation of the associated image. We assessed our method using both synthetic data and CT scans of stroke patients. Specifically, we demonstrated compatibility between the entropy of the consistency scores with respect to the image blurriness and contrast, measured by blur kernel widths (for the synthetic data) and Dice scores (for the CT scans). The proposed tool is shown to provide meaningful fusion of soft segmentations and segmentation model assessment. We believe that the concept of segmentation uncertainty margins, highlighted here, has great importance for medical imaging analysis.

5. REFERENCES

[1] Fan et al., "MCMC curve sampling for image segmentation," in *MICCAI*. Springer, 2007, pp. 477–485.

[2] W.K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[3] Hazan et al., "On sampling from the gibbs distribution with random maximum a-posteriori perturbations," in *NIPS*, 2013, pp. 1268–1276.

[4] G. Papandreou and A. Yuille, "Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models," in *ICCV*, 2011, pp. 193–200.

[5] Alberts et al., "Uncertainty quantification in brain tumor segmentation using CRFs and random perturbation models," in *ISBI*. IEEE, 2016, pp. 428–431.

[6] Le et al., "Sampling image segmentations for uncertainty quantification," *Medical image analysis*, vol. 34, pp. 42–51, 2016.

[7] Elhabian et al., "Optimal parameter map estimation for shape representation: A generative approach," in *ISBI*. IEEE, 2016, pp. 660–663.

[8] S.K. Warfield, K.H. Zou, and W.M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE TMI*, vol. 23, no. 7, pp. 903–921, 2004.

[9] K. Van Leemput and M. R. Sabuncu, "A cautionary analysis of staple using direct inference of segmentation truth," in *MICCAI*, 2014, pp. 398–406.

[10] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Medical image analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[11] O. Juan, R. Keriven, and G. Postelnicu, "Stochastic motion and the level set method in computer vision: Stochastic active contours," *IJCV*, vol. 69, no. 1, pp. 7–25, 2006.

[12] L. Dice, "Measure of the amount of ecological association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[13] Pohl et al., "Using the logarithm of odds to define a vector space on probabilistic atlases," *Medical Image Analysis*, vol. 11, no. 6, pp. 465–477, 2007.

[14] N. Paragios and R. Deriche, "Geodesic active regions: A new framework to deal with frame partition problems in computer vision," *JVCIR*, vol. 13, pp. 249–268, 2002.

[15] T.F. Chan and L.A. Vese, "Active contours without edges," *IEEE TIP*, vol. 10, no. 2, pp. 266–277, 2001.

[16] M-P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *IAPR*, 1994, vol. 1, pp. 566–568.