Validation of Image Segmentation by Estimating Rater Bias and Variance

Simon K. Warfield[1,2], Kelly H. Zou[2], and William M. Wells[2]

[1] Computational Radiology Laboratory, Dept. Radiology, Children's Hospital
[2] Dept. Radiology, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St., Boston, MA 02115 USA
{warfield, zou, sw}@bwh.harvard.edu

Abstract. The accuracy and precision of segmentations of medical images has been difficult to quantify in the absence of a "ground truth" or reference standard segmentation for clinical data. Although physical or digital phantoms can help by providing a reference standard, they do not allow the reproduction of the full range of imaging and anatomical characteristics observed in clinical data.

An alternative assessment approach is to compare to segmentations generated by domain experts. Segmentations may be generated by raters who are trained experts or by automated image analysis algorithms. Typically these segmentations differ due to intra-rater and inter-rater variability. The most appropriate way to compare such segmentations has been unclear.

We present here a new algorithm to enable the estimation of performance characteristics, and a true labeling, from observations of segmentations of imaging data where segmentation labels may be ordered or continuous measures. This approach may be used with, amongst others, surface, distance transform or level set representations of segmentations, and can be used to assess whether or not a rater consistently over-estimates or under-estimates the position of a boundary.

1 Introduction

Previous work for estimating a reference standard from segmentations has utilized the Expectation-Maximization (EM) algorithm to estimate performance characteristics and the hidden "true" segmentation from a collection of independent binary segmentations indicating presence or absence of a structure in a given image and from a collection of segmentations with multi-category labellings, such as gray matter, white matter and cerebrospinal fluid [1]. The method has been used to characterize image segmentations [2] and to infer labellings from repeated registrations [3]. These approaches are not appropriate when an ordering is present in the segmentation labels, such as when the segmentation boundary is represented by a distance transform or level set.

The objective of this work was to generalize the existing methodology in simultaneous ground truth estimation and performance level estimation to segmentations with ordered labels. We propose here a model in which we summarize the quality of a rater-generated segmentation by the mean and variance of the distance between the segmentation boundary and a reference standard boundary. We propose an Expectation-Maximization algorithm to estimate the reference standard boundary, and the bias and variance of each rater. This enables interpretation of the quality of a segmentation generator, be it an expert human or an algorithm. A good quality rater will have a small

R. Larsen, M. Nielsen, and J. Sporring (Eds.): MICCAI 2006, LNCS 4191, pp. 839–847, 2006.

average distance from the true boundary (low bias) and high precision (small variance of distance from the true boundary).

We applied the proposed algorithm to the assessment of brain tumor segmentations. We evaluated the algorithm by testing its operation on phantoms to confirm its operation with a known reference standard, and we assessed the significance of the method by comparing and contrasting to other means of measuring segmentation similarities with the Dice [4] overlap measure and with STAPLE [1].

2 Method

2.1 Notations and Assumptions

We observe a set of segmentations of an image, created by some number of human expert or algorithmic raters. It is our goal to estimate the true labeling of the image, and to estimate performance characteristics of each rater in comparison. The true labeling is unknown (hidden), but if it was not hidden, then it would be straightforward to compute bias and variance parameters for each rater characterizing the way in which the rater labeling differs from the true labeling. Since the true labeling is hidden, we describe here an Expectation-Maximization algorithm [5] in order to estimate the true labeling and rater performance parameters. Our EM algorithm proceeds iteratively, first estimating the complete data log-likelihood function, and then identifying the parameters that maximize the estimated complete data log-likelihood function. The estimation of the complete data log-likelihood involves computing the expected value of this function conditioned upon the observed rater labelings and previous estimates of the rater parameters. Computation of this expectation requires estimation of the conditional probability of the true score given the observed scores and rater performance parameters.

Let $i = 1,\ldots,I$ index the raters generating a segmentation of an image and $j = 1,\ldots,J$ index the voxels in the image. The label of each voxel is a continuous scalar, such as may be obtained from a distance transform or level set representation of a segmentation (but is not restricted to those sources) and is referred to as a score. The score s_{ij} assigned by rater i for voxel j has the following composition:

$$s_{ij} = \tau_j + \beta_i + \varepsilon_{ij}, \tag{1}$$

where τ_j is the underlying true score for this voxel, and β_i is the bias of rater i. The error term is assumed to have an uncorrelated normal distribution, i.e., $\varepsilon_{ij} \sim N(0,\sigma_i^2)$ for each rater i. That is, we characterize the rater performance by a bias β_i and a variance σ_i^2. We assume that, given an image to be scored, the raters score the image independently.

We wish to estimate the the bias and variance that characterizes the performance of each rater. If the true score was known, we could estimate the rater parameters $\theta = (\sigma,\beta)$ by solving the *complete data* log likelihood problem:

$$\hat{\theta} = \arg\max_{\theta} \log \Pr(\mathbf{s},\tau|\sigma,\beta) \tag{2}$$

Since the true score is unknown, we instead estimate the complete data log likelihood using Expectation-Maximization [6] by computing its expected value under a conditional probability density function for the true scores given the rater scores and previous estimates of the rater parameters. That is, we identify the rater parameters by solving:

$$\theta^{(k)} = \arg\max_{\theta} E\left[\log \Pr(\mathbf{s}, \tau|\sigma, \beta)\middle|\mathbf{s}, \theta^{(k-1)}\right] \tag{3}$$

In order to compute this expectation, we need a conditional probability density for the true scores given the rater scores and rater parameter estimates. We now derive the required densities and estimators.

In the absence of spatial correlations between voxels, the joint distribution of the scores of all the voxels conditional upon the true scores and rater parameters, is assumed to have the form :

$$\Pr(\mathbf{s}|\tau, \sigma, \beta) = \prod_{j=1}^{J}\prod_{i=1}^{I} \phi\left\{\frac{s_{ij} - (\tau_j + \beta_i)}{\sigma_i}\right\}, \tag{4}$$

where $\phi(\cdot)$ is the probability density function (pdf) of the standard normal distribution, $N(0,1)$. For notational simplicity, we write the pdf $N(\mu, \sigma^2)$ as $\phi_\sigma(\mu)$.

2.2 A Conditional Probability Density Function for the True Scores

Bayes' theorem is applied as follows in order to derive the posterior distribution $\Pr(\tau|\mathbf{s}, \sigma, \beta)$ from the distribution of the observed scores $\Pr(\mathbf{s}|\tau, \sigma, \beta)$.

Since the true score is independent of the rater bias and variance, the posterior distribution is

$$\Pr(\tau|\mathbf{s}, \sigma, \beta) = \Pr(\mathbf{s}|\tau, \sigma, \beta) \cdot \frac{\Pr(\tau, \sigma, \beta)}{\Pr(\mathbf{s}, \sigma, \beta)} \tag{5}$$

$$= \Pr(\mathbf{s}|\tau, \sigma, \beta) \cdot \frac{\Pr(\tau)\Pr(\sigma, \beta)}{\Pr(\mathbf{s}, \sigma, \beta)}, \tag{6}$$

$$= \Pr(\mathbf{s}|\tau, \sigma, \beta) \cdot \frac{\Pr(\tau)}{\Pr(\mathbf{s}|\sigma, \beta)}. \tag{7}$$

Upon integrating the expression of Equation 7 over τ, since the marginal distribution integrates to 1, we have

$$\int_\tau \Pr(\tau|\mathbf{s}, \sigma, \beta) d\tau = \int_\tau \Pr(\mathbf{s}|\tau, \sigma, \beta) \cdot \frac{\Pr(\tau)}{\Pr(\mathbf{s}|\sigma, \beta)} d\tau, \tag{8}$$

$$1 = \frac{1}{\Pr(\mathbf{s}|\sigma, \beta)} \int_\tau \Pr(\mathbf{s}|\tau, \sigma, \beta)\Pr(\tau), \tag{9}$$

$$\Pr(\tau|\mathbf{s}, \sigma, \beta) = \frac{\Pr(\mathbf{s}|\tau, \sigma, \beta)\Pr(\tau)}{\int_\tau \Pr(\mathbf{s}|\tau, \sigma, \beta)\Pr(\tau) d\tau}. \tag{10}$$

We have considerable freedom in choosing the prior distribution on τ, $\Pr(\tau) = \prod_{j=1}^{J} \Pr(\tau_j)$. A multivariate Gaussian distribution is a natural choice, and in the absence of other information we may choose a multivariate uniform distribution, with scale parameter h,

$$\Pr(\tau) = \prod_{j=1}^{J} U_h(\tau_j) = \frac{1}{h^J}, \tag{11}$$

which simplifies the form of the posterior:

$$\Pr(\tau|\mathbf{s},\sigma,\beta) = \frac{\Pr(\mathbf{s}|\tau,\sigma,\beta)\frac{1}{h^J}}{\frac{1}{h^J}\int_\tau \Pr(\mathbf{s}|\tau,\sigma,\beta)d\tau}, \tag{12}$$

$$= \frac{\Pr(\mathbf{s}|\tau,\sigma,\beta)}{\int_\tau \Pr(\mathbf{s}|\tau,\sigma,\beta)d\tau}. \tag{13}$$

Let the bias-adjusted score be denoted $\mu_{ij} = \mathbf{s}_{ij} - \beta_i$. From Equation 4 and Equation 10, we have

$$\Pr(\tau|\sigma,\mu) = \frac{1}{Z}\prod_{j=1}^{J}\Pr(\tau_j)\prod_{i=1}^{I}\phi_{\sigma_i}(\mu_{ij}-\tau_j), \tag{14}$$

where the normalizing constant is

$$Z = \int_{\tau_J}\dots\int_{\tau_1}\prod_{j=1}^{J}\Pr(\tau_j)\prod_{i=1}^{I}\phi_{\sigma_i}(\mu_{ij}-\tau_j)d\tau_1\dots d\tau_J. \tag{15}$$

Therefore we can write for each voxel,

$$\begin{aligned}\Pr(\tau_j|\mu,\sigma) &= \frac{1}{Z_j}\Pr(\tau_j)\prod_{i=1}^{I}\phi_{\sigma_i}(\mu_{ij}-\tau_j)\\ &= \frac{1}{Z_j}\Pr(\tau_j)\prod_{i=1}^{I}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{(\mu_{ij}-\tau_j)^2}{2\sigma_i^2}\right\}\\ &= \frac{1}{Z_j}\left\{\prod_{i=1}^{I}(2\pi\sigma_i^2)^{-1/2}\right\}\exp(w_{ij})\Pr(\tau_j).\end{aligned} \tag{16}$$

which is a product of normal distributions with mean μ_{ij} and variance σ_i^2 and where

$$\begin{aligned}w_{ij} &= -\frac{1}{2}\sum_{i=1}^{I}\frac{(\mu_{ij}-\tau_j)^2}{\sigma_i^2}\\ &= -\frac{1}{2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}(\mu_{ij}^2 - 2\mu_{ij}\tau_j + \tau_j^2)\\ &= -\frac{1}{2}\left[\sum_{i=1}^{I}\frac{\mu_{ij}^2}{\sigma_i^2} - 2\tau_j\sum_{i=1}^{I}\frac{\mu_{ij}}{\sigma_i^2} + \tau_j^2\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\right]\end{aligned} \tag{17}$$

If $\Pr(\tau_j)$ of Equation 16 is uniform, then completing the square and identifying the terms in a Gaussian probability density function yields the following variance and mean terms

$$\frac{1}{\sigma^2} = \sum_{i=1}^{I}\frac{1}{\sigma_i^2}, \tag{18}$$

$$\mu_j = \frac{\sum_{i=1}^{I}\frac{\mathbf{s}_{ij}-\beta_i}{\sigma_i^2}}{\frac{1}{\sigma^2}}. \tag{19}$$

If $\Pr(\tau_j)$ of Equation 16 is distributed $N(\mu_{\tau_j},\sigma_{\tau_j}^2)$ then it acts analogously to another rater with the specified mean and variance.

Observe that the mean score for voxel j indicated by this distribution is an inverse rater variance weighted sum of the difference between the score of the rater and the rater bias, and that the variance of the distribution is the harmonic mean of the rater variances.

2.3 Estimating the Complete Data Log-Likelihood

The parameters maximizing the conditional expectation of the complete data log-likelihood can be found from

$$\theta^{(k)} = \arg\max_{\theta} E\left[\log \Pr(\mathbf{s},\tau|\sigma,\beta)\big|\mathbf{s},\theta^{(k-1)}\right] \tag{20}$$

$$= \arg\max_{\theta} E\left[\log \Pr(\mathbf{s}|\tau,\sigma,\beta)\big|\mathbf{s},\theta^{(k-1)}\right] \tag{21}$$

$$= \arg\max_{\sigma,\beta} E\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\log\left(\frac{1}{\sqrt{(2\pi\sigma_i^2)}}\exp(-\frac{1}{2\sigma_i^2}(\mu_{ij}-\tau_j)^2)\right)\big|\mathbf{s},\theta^{(k-1)}\right] \tag{22}$$

$$= \arg\max_{\sigma,\beta} \sum_{i=1}^{I}\sum_{j=1}^{J}\left[-\log\sigma_i - \frac{1}{2\sigma_i^2}(\mu_{ij}^2 - 2\mu_{ij}E(\tau_j) + E(\tau_j^2))\right] \tag{23}$$

Now, given the distribution $Pr(\tau|\mathbf{s},\sigma,\beta)$ above, we have

$$E(\tau_j) = \mu_j^{(k-1)} \tag{24}$$

$$E(\tau_j^2) = \mathrm{var}(\tau_j) + E(\tau_j)^2 \tag{25}$$

$$= (\sigma^2)^{(k-1)} + (\mu_j^2)^{(k-1)} \tag{26}$$

Hence,

$$\theta^{(k)} = \arg\max_{\sigma,\beta} \sum_{i=1}^{I}\sum_{j=1}^{J}\left[-\log\sigma_i - \frac{1}{2\sigma_i^2}(\mu_{ij}^2 - 2\mu_{ij}\mu_j^{(k-1)} + (\sigma^2)^{(k-1)} + (\mu_j^2)^{(k-1)})\right] \tag{27}$$

On differentiating Equation 27 with respect to the parameters β,σ and solving for a maximum we find the following estimators for the rater performance parameters:

$$\beta_i^{(k)} = \frac{1}{J}\sum_{j=1}^{J}(\mathbf{s}_{ij} - \tau_j^{(k-1)}), \tag{28}$$

$$(\sigma_i^2)^{(k)} = \frac{1}{J}\sum_{j=1}^{J}\left(\mathbf{s}_{ij} - \beta_i^{(k)} - \tau_j^{(k-1)}\right)^2 + (\sigma^2)^{(k-1)} \tag{29}$$

The rater bias estimator is the average observed difference between rater score and the previously estimated 'true' score over all the voxels, and the rater variance estimator is a sum of a term describing a natural empirical variance estimator (given estimates of the rater bias and the true score) and a term that is the harmonic mean of the previous estimates of the rater variances over all raters. Thus, the estimated variance for a rater cannot be smaller than the variance associated with the true score estimate.

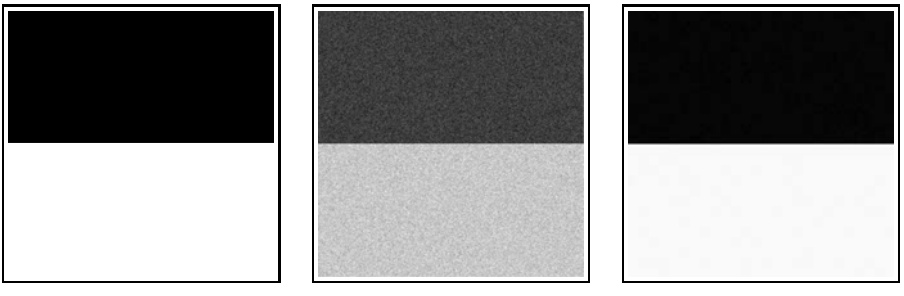
3 Results

We applied the proposed estimation scheme to a set of digital phantoms, generated by synthetic raters with pre-specified bias and variance parameters, in order to determine if the estimation scheme would correctly identify the bias and variance of segmentation generators for which we knew the true parameter values.

We applied the proposed estimation scheme to MRI scans of four different brain tumors. We compared the estimated true contour to that obtained from averaging the segmentations. We compared the segmentations identified as best and as worst by the algorithm to the original MRI scan visually, and by computing a spatial overlap measure between the reference standard and each segmentation.

3.1 Estimation of Parameters of Synthetic Raters Using a Digital Phantom

Segmentations were generated by randomly perturbing an image consisting of a rectangular region with intensity 100 and a rectangular region of intensity 200. Ten random segmentations were generated, drawing from five synthetic raters with a bias of +10 units and a variance of 100 units and five synthetic raters with a bias of -10 units and a variance of 50 units. The estimation scheme was executed and estimates of the true image and of the performance characteristics of each rater were obtained. This is illustrated in Figure 1. The estimated bias was 9.9968 ± 0.003 and -9.9968 ± 0.0002 which is a tight estimate around the true value. The estimated variance was 100.16 ± 0.28 and 50.112 ± 0.101 which again is very close to the specified values of the rater variances.



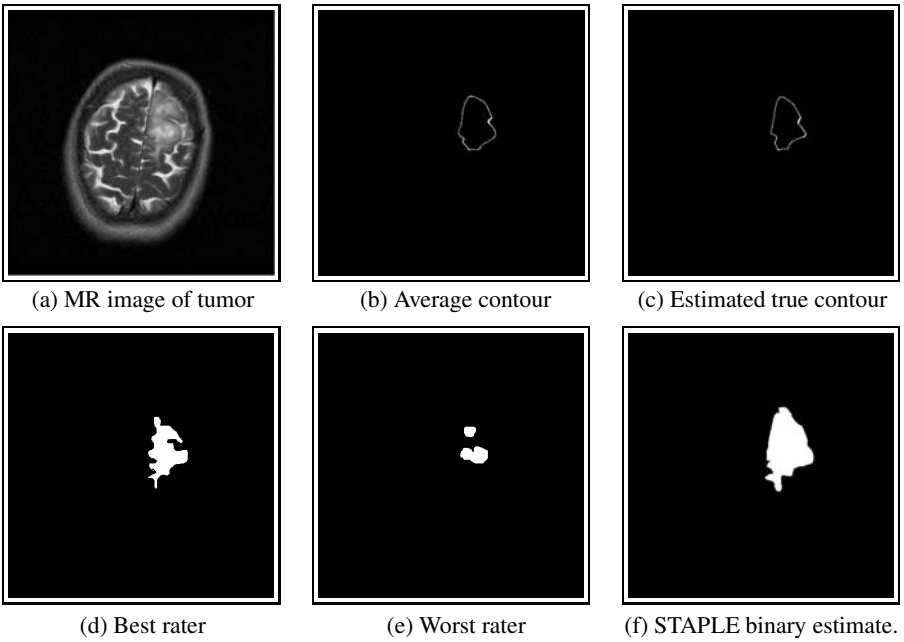
(a) Synthetic True Labeling (b) Rater segmentation with bias of +10, variance of 100 (c) Estimated true labelling

Fig. 1. Estimated true labels from synthetic raters. The specified labeled data was created with two regions of intensity equal to 100 and 200 respectively. Five raters with a bias of 10 and a variance of 100, and five raters with a bias of -10 and a variance of 50 were simulated to create synthetic segmentations. The estimation scheme was used to estimate each rater bias and variance, and to estimate the hidden true labeling. As can be seen in (c), despite the noisy and limited observations such as that shown in (b), the estimated reference standard appears very similar to the specified data of (a). The closeness of the estimated parameters for each of the raters to the specified values confirms the estimate is effective.

3.2 Segmentations of Brain Tumor MRI by Human Raters

Each brain tumor was segmented up to three times by each of nine raters for a maximum of twenty seven segmentations. Each rater segmentation consisted of a closed contour delineating the extent of the tumor that the rater perceived in the MRI. A signed distance transform of each contour was computed in order to obtain a segmentation with each voxel representing the shortest distance to the boundary.

The estimation scheme was executed to find the best overall reference standard utilizing all of the segmentations. The scheme was run for 100 iterations, resulting in the sum of total estimated true scores changing by less than 0.01 at the final iteration, which required less than 90 seconds of computation time on a PowerBook G4 in each case. One illustrative brain tumor is shown in Figure 2, together with the average contour, the estimated true contour and the rater segmentations the algorithm indicated to be most like and most different from the true contour.



(a) MR image of tumor (b) Average contour (c) Estimated true contour

(d) Best rater (e) Worst rater (f) STAPLE binary estimate.

Fig. 2. Estimated true segmentation from 21 human rater segmentations. The estimated true contour is consistent with the human raters, and a better reflection of the true tumor boundary than the average. The best and worst rater segmentations as determined by the algorithm is displayed. For comparison, a binary estimate obtained with STAPLE is shown in (f).

We sought to summarize the spatial overlap between the rater segmentations, the average segmentation enclosed by the contour obtained by averaging each of the signed distance transforms, and the segmentation defined by the zero level set of the estimation

procedure. Different raters perceived the tumor boundary differently, and we measured the pairwise spatial overlap using the Dice measure. The coefficient of variation of the Dice overlap measures between the estimated reference segmentation and each of the rater segmentations was 0.15, which was smaller than the coefficient of variation of the Dice overlap measures between the average segmentation and the raters (0.17) and the coefficient of variation of the Dice overlap measures between each rater and the other raters (which ranged from the smallest of 0.16 to 0.36). This indicates that there was considerable variability between the raters, but that the estimated segmentation was more consistent with the raters than both the average of the segmentations and the other rater segmentations, and so is a good summary of the segmentations.

4 Discussion and Conclusion

Validation of image segmentations using an estimated reference standard is a valuable strategy for clinical imaging data where the true segmentation is unknown. Previous algorithms were designed for binary or unordered multi-category labels. Such methods strongly penalize segmentations that may differ by small mis-localizations.

The novel approach developed here is suitable for segmentations represented by a surface or boundary from which a distance may be computed, or for boundaries represented directly by a level set. This may be especially well suited to complicated structures where the true boundary is challenging to estimate, such as in MRI of brain tumors, or where spatial mislocalization of the segmentation is expected and should be tolerated. An example of the latter situation would be in the estimation of center lines of blood vessels, or the colon, or in the analysis of spicules in mammography.

We demonstrated that the estimation scheme was able to recover the true parameters of synthetic raters and from this form a good estimate of the true image structure using digital phantoms. We demonstrated that from a collection of human segmentations of brain tumors, the estimation scheme was able to identify a reference standard that was closer to the rater segmentations than the raters were to each other. We compared the reference standard to an estimate obtained by averaging and the results demonstrate that the average segmentation is less representative of the true anatomy. We demonstrated that the estimation scheme was able to rank the human raters, identifying the best and worst segmentations and providing valuable estimates of the rater bias and variance.

Future work will examine the possibility of incorporating a model for spatial correlation in the true labeling, which will enable us to relax the assumption of voxelwise independence. It will also be interesting to examine alternatives for the prior probability of each label, such as may be derived from an anatomical atlas.

Acknowledgements. This investigation was supported by NSF ITR 0426558, a research grant from CIMIT, a research grant from the Whitaker Foundation, grant RG 347A2/2 from the NMSS, and by NIH grants R21 MH67054, R01 RR021885, P41 RR013218, and U41 RR019703.

References

1. 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 Trans Med Imag*, vol. 23, pp. 903–921, 2004.
2. T. Rohlfing, D. B. Russakoff, and C. R. Maurer, "Expectation maximization strategies for multi-atlas multi-label segmentation," in *Proceedings of International Conference of Information Processing in Medical Imaging*, pp. 210–221, 2003.
3. T. Rohlfing, D. B. Russakoff, and C. R. Maurer, "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation.," *IEEE Transactions On Medical Imaging*, vol. 23, pp. 983–994, August 2004.
4. L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
5. A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Ser. B.*, vol. 39, pp. 34–37, 1977.
6. G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York, New York: Wiley-Interscience, 1996.