Validation of Image Segmentation and Expert Quality with an Expectation-Maximization Algorithm

Simon K. Warfield, Kelly H. Zou, and William M. Wells

Computational Radiology Laboratory and Surgical Planning Laboratory, Harvard Medical School and Brigham and Women's Hospital, 75 Francis St., Boston, MA 02115 USA
{warfield,zou,sw}@bwh.harvard.edu

Abstract. Characterizing the performance of image segmentation approaches has been a persistent challenge. Performance analysis is important since segmentation algorithms often have limited accuracy and precision. Interactive drawing of the desired segmentation by domain experts has often been the only acceptable approach, and yet suffers from intra-expert and inter-expert variability. Automated algorithms have been sought in order to remove the variability introduced by experts, but no single methodology for the assessment and validation of such algorithms has yet been widely adopted. The accuracy of segmentations of medical images has been difficult to quantify in the absence of a "ground truth" segmentation for clinical data. Although physical or digital phantoms can help, they have so far been unable to reproduce the full range of imaging and anatomical characteristics observed in clinical data. An attractive alternative is comparison to a collection of segmentations by experts, but the most appropriate way to compare segmentations has been unclear.

We present here an Expectation-Maximization algorithm for computing a probabilistic estimate of the "ground truth" segmentation from a group of expert segmentations, and a simultaneous measure of the quality of each expert. This approach readily enables the assessment of an automated image segmentation algorithm, and direct comparison of expert and algorithm performance.

1 Introduction

Medical image segmentation has long been recognized as a challenging problem. Many different approaches have been proposed, and different approaches are often suitable for different clinical applications.

Characterizing the performance of image segmentation approaches has also been a persistent challenge. Interactive drawing of the desired segmentation by domain experts has often been the only acceptable approach, and yet suffers from intra-expert and inter-expert variability. Automated image segmentation algorithms have been sought in order to remove the variability introduced by experts.

The accuracy of automated image segmentation of medical images has been difficult to quantify in the absence of a "ground truth" segmentation for clinical data. Although physical or digital phantoms can provide a level of known "ground truth" [1,2], they have so far been unable to reproduce the full range of imaging characteristics (partial volume artifact, intensity inhomogeneity artifact, noise) and normal and abnormal anatomical variability observed in clinical data. A common alternative to phantom studies, has been

T. Dohi and R. Kikinis (Eds.): MICCAI 2002, LNCS 2488, pp. 298–306, 2002.

a behavioural comparison: an automated algorithm is compared to the segmentations generated by a group of experts, and if the algorithm generates segmentations sufficiently similar to the experts it is regarded an acceptable substitute for the experts. Typically, good automated segmentation algorithms will also require less time to apply, and have better reproducibility, than interactive segmentation by an expert.

The most appropriate way to carry out the comparison of an automated segmentation to a group of expert segmentations is so far unclear. A number of metrics have been proposed to compare segmentations, including volume measures, spatial overlap measures (such as Dice [3] and Jaccard similarities [4]) and boundary measures (such as the Hausdorff measure). Agreement measures between different experts have also been explored for this purpose [5]. Studies of rules to combine segmentations to form an estimate of the underlying "true" segmentation have as yet not demonstrated any one scheme to be much favourable to another. Per-voxel voting schemes have been used in practice [6,7].

We present here a new Expectation-Maximization (EM) algorithm for estimating simultaneously the "ground truth" segmentation from a group of expert segmentations, and a measure of the quality of each expert. Our algorithm is formulated as an instance of the Expectation-Maximization algorithm [8]. In our algorithm, the expert segmentation decision at each voxel is directly observable, the hidden ground truth is a binary variable for each voxel, and the quality of each expert is represented by sensitivity and specificity parameters. The complete data consists of the expert decisions, which are known, and the ground truth, which is not known. If we also knew the ground truth it would be straightforward to estimate the expert quality parameters. Since the complete data is not available, we replace the ground truth (hidden) variables with their expected values under the assumption of the previous estimate of the expert quality parameters. We can then re-estimate the expert quality parameters. We iterate this sequence of estimation of the quality parameters and ground truth variables until convergence is reached.

This approach readily enables the assessment of an automated image segmentation algorithm, and direct comparison of expert and algorithm performance.

We applied the estimation algorithm described here to several **digital phantoms** for which the ground truth is known. These experiments indicate the approach is robust to small parameter changes, and indicate how our algorithm resolves ambiguities between different experts in a natural way. We applied the method to the assessment of a previously published automated image segmentation algorithm [9] for the segmentation of **brain tumors** from magnetic resonance images [10]. We compared the performance of three experts to that of the segmentation algorithm on ten cases. Illustrative results are presented. We assessed multiple segmentations by a single expert for the task of identifying the **prostate peripheral zone** from magnetic resonance images.

2 Method

We describe an EM algorithm for estimating the hidden ground truth and expert segmentation quality parameters from a collection of segmentations by experts.

Let $\mathbf{T}$ be the hidden binary ground truth segmentation, $(\mathbf{p}, \mathbf{q})$ be the sensitivity and specificity parameters characterising the performance of each expert, and $\mathbf{D}$ be the

segmentation decisions made by each expert. Hence, the probability mass function of the complete data is $f(\mathbf{D}, \mathbf{T}|\mathbf{p}, \mathbf{q})$.

We want to estimate the quality parameters of the experts as the parameters that maximize the log likelihood function

$$\hat{\mathbf{p}}, \hat{\mathbf{q}} = \arg\max_{\mathbf{p},\mathbf{q}} \ln f(\mathbf{D}, \mathbf{T}|\mathbf{p}, \mathbf{q}). \tag{1}$$

If we knew the ground truth segmentation, we could construct a 2×2 conditional probability table, by comparing the decision D_{ij} of expert j (for $j = 1, ..., K$) as to presence or absence of a structure at voxel i with the ground truth. Let q_j represent the 'true negative fraction' or specificity (i.e. relative frequency of $D_{ij} = 0$ when $T_i = 0$) and p_j represent the 'true positive fraction' or sensitivity (relative frequency of $D_{ij} = 1$ when $T_i = 1$). These parameters $\{p_j; q_j\} \in [0, 1]$ are assumed to depend upon the expert, and may be equal but in general are not.

We assume that the experts decisions are all conditionally independent given the ground truth and the expert quality parameters, that is $(D_{ij}|T_i, p_j, q_j) \perp (D_{ij'}|T_i, p_{j'}, q_{j'}), \forall j \neq j'$.

Since we don't know the ground truth $\mathbf{T}$, we treat it as a random variable and instead solve for the parameters which maximize the function $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$, where

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \boldsymbol{\theta}_2 \ldots \boldsymbol{\theta}_k], \qquad \boldsymbol{\theta}_j = (p_j, q_j)^T \quad \forall j \in [1, \ldots, K] \tag{2}$$

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = E_{g(\mathbf{T}|\mathbf{D},\hat{\boldsymbol{\theta}})} \left[\ln f(\mathbf{D}, \mathbf{T}|\boldsymbol{\theta})\right]. \tag{3}$$

This can be written as

$$\hat{\mathbf{p}}, \hat{\mathbf{q}} = \arg\max_{\mathbf{p},\mathbf{q}} E_{g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^o,\hat{\mathbf{q}}^o)} \left[\ln f(\mathbf{D}, \mathbf{T}|\mathbf{p}, \mathbf{q})\right] \tag{4}$$

$$= \arg\max_{\mathbf{p},\mathbf{q}} E_{g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^o,\hat{\mathbf{q}}^o)} \left[\ln \frac{f(\mathbf{D}, \mathbf{T}, \mathbf{p}, \mathbf{q})}{f(\mathbf{p}, \mathbf{q})}\right] \tag{5}$$

$$= \arg\max_{\mathbf{p},\mathbf{q}} E_{g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^o,\hat{\mathbf{q}}^o)} \left[\ln \frac{f(\mathbf{D}, \mathbf{T}, \mathbf{p}, \mathbf{q}) f(\mathbf{T}, \mathbf{p}, \mathbf{q})}{f(\mathbf{T}, \mathbf{p}, \mathbf{q}) f(\mathbf{p}, \mathbf{q})}\right] \tag{6}$$

$$= \arg\max_{\mathbf{p},\mathbf{q}} E_{g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^o,\hat{\mathbf{q}}^o)} \left[\ln f(\mathbf{D}|\mathbf{T}, \mathbf{p}, \mathbf{q}) f(\mathbf{T})\right] \tag{7}$$

where $\hat{\mathbf{p}}^o, \hat{\mathbf{q}}^o$ are the previous estimates of the expert quality parameters, and the last follows under the assumption that $\mathbf{T}$ is independent of the expert quality parameters so that $f(\mathbf{T}, \mathbf{p}, \mathbf{q}) = f(\mathbf{T}) f(\mathbf{p}, \mathbf{q})$.

The process to identify the expert quality parameters and ground truth consists of iterating between 1) estimation of the hidden ground truth given a previous estimate of the expert quality parameters, and 2) estimation of the expert quality parameters based on how they performed given the new estimate of the ground truth. This algorithm can be recognized as an EM algorithm, in which the parameters that maximize the log likelihood function are estimated based upon the expected value of the hidden ground truth. The process can be initialized by assuming values for the expert specific sensitivity and specificity parameters, or by assuming an initial ground truth estimate. In most of our experiments we have initialized the algorithm by assuming that the experts are each equally good and have high sensitivity and specificity, but are not infallible. This is

equivalent to initializing the algorithm by estimating an initial ground truth as an equal weight combination of each of the expert segmentations.

2.1 Estimation of Ground Truth Given Expert Parameters

In this section, the estimator for the hidden ground truth is derived.

$$g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^{\mathbf{o}},\hat{\mathbf{q}}^{\mathbf{o}}) = \frac{g(\mathbf{D}|\mathbf{T},\hat{\mathbf{p}}^{\mathbf{o}},\hat{\mathbf{q}}^{\mathbf{o}})g(\mathbf{T})}{\sum_{\mathbf{T}} g(\mathbf{D}|\mathbf{T},\hat{\mathbf{p}}^{\mathbf{o}},\hat{\mathbf{q}}^{\mathbf{o}})g(\mathbf{T})} \tag{8}$$

$$= \frac{\prod_i \left[\prod_j g(D_{ij}|T_i,\hat{p_j}^o,\hat{q_j}^o)g(T_i)\right]}{\sum_{T_1}\sum_{T_2}\cdots\sum_{T_N}\prod_i \left[\prod_j g(D_{ij}|T_i,\hat{p_j}^o,\hat{q_j}^o)g(T_i)\right]} \tag{9}$$

$$= \frac{\prod_i \left[\prod_j g(D_{ij}|T_i,\hat{p_j}^o,\hat{q_j}^o)g(T_i)\right]}{\prod_i \left[\sum_{T_i}\prod_j g(D_{ij}|T_i,\hat{p_j}^o,\hat{q_j}^o)g(T_i)\right]}. \tag{10}$$

$$g(T_i|\mathbf{D_i},\hat{\mathbf{p}}^{\mathbf{o}},\hat{\mathbf{q}}^{\mathbf{o}}) = \frac{\prod_j g(D_{ij}|T_i,\hat{p_j}^o,\hat{q_j}^o)g(T_i)}{\sum_{T_i}\prod_j g(D_{ij}|T_i,\hat{p_j}^o,\hat{q_j}^o)g(T_i)} \tag{11}$$

where $g(T_i)$ is the a priori probability of T_i, and a voxelwise independence assumption has been made.

We store for each voxel the estimate of the probability that the ground truth at each voxel is $T_i = 1$. Since the ground truth is treated as a binary random variable, the probability that $T_i = 0$ is simply $1 - g(T_i = 1|\mathbf{D_i},\hat{\mathbf{p}}^{\mathbf{o}},\hat{\mathbf{q}}^{\mathbf{o}})$.

Let $\alpha = \prod_{j:D_{ij}=1}\hat{p_j}^o\prod_{j:D_{ij}=0}(1-\hat{p_j}^o)$ and $\beta = \prod_{j:D_{ij}=0}\hat{q_j}^o\prod_{j:D_{ij}=1}(1-\hat{q_j}^o)$ where $j : D_{ij} = 1$ denotes the values of the index j for which the expert decision at voxel i (i.e. D_{ij}) has the value 1. Using the notation common for EM algorithms, let W_i be the weight variable.

$$W_i \equiv g(T_i = 1|\mathbf{D_i},\hat{\mathbf{p}}^{\mathbf{o}},\hat{\mathbf{q}}^{\mathbf{o}}) \tag{12}$$

$$= \frac{g(T_i = 1)\alpha}{g(T_i = 1)\alpha + (1 - g(T_1 = 1))\beta}. \tag{13}$$

The weight W_i indicates the probability of the ground truth at voxel i being equal to one. It is a normalized product of the prior probability of $T_i = 1$, the sensitivity of each of the experts that decided ground truth was one and the product of (1 - sensitivity) of each of the experts that decided the ground truth was zero.

2.2 Estimation of the Quality Parameters of the Experts

Given the estimate of the value of the ground truth derived above, we can find the values of the expert quality parameters that maximize the expectation of the log likelihood function.

$$\hat{\mathbf{p}},\hat{\mathbf{q}} = \arg\max_{\mathbf{p},\mathbf{q}} E_{g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^{\mathbf{o}},\hat{\mathbf{q}}^{\mathbf{o}})}\left[\ln\left(f(\mathbf{D}|\mathbf{T},\mathbf{p},\mathbf{q})f(\mathbf{T})\right)\right] \tag{14}$$

$$= \arg\max_{\mathbf{p},\mathbf{q}} E_{g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^{\mathbf{o}},\hat{\mathbf{q}}^{\mathbf{o}})}\left[\ln\prod_{ij} f(D_{ij}|T_i,p_j,q_j) + \ln\prod_i f(T_i)\right] \tag{15}$$

Since $\ln \prod_i f(T_i)$ is not a function of $\mathbf{p}, \mathbf{q}$, it follows that

$$\hat{\mathbf{p}}, \hat{\mathbf{q}} = \arg\max_{\mathbf{p},\mathbf{q}} \sum_j \sum_i E_{g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^o,\hat{\mathbf{q}}^o)} [\ln f(D_{ij}|T_i, p_j, q_j)] \tag{16}$$

$$\begin{aligned} \hat{p_j}, \hat{q_j} &= \arg\max_{p_j,q_j} \sum_i E_{g(\mathbf{T}|\mathbf{D},\hat{\mathbf{p}}^o,\hat{\mathbf{q}}^o)} [\ln f(D_{ij}|T_i, p_j, q_j)] \qquad (17) \\ &= \arg\max_{p_j,q_j} \sum_i \Big[W_i \ln f(D_{ij}|T_i = 1, p_j, q_j) \\ &\qquad + (1 - W_i) \ln f(D_{ij}|T_i = 0, p_j, q_j) \Big] \\ &= \arg\max_{p_j,q_j} \sum_{i:D_{ij}=1} W_i \ln p_j + \sum_{i:D_{ij}=1} (1 - W_i) \ln(1 - q_j) \\ &\qquad + \sum_{i:D_{ij}=0} W_i \ln(1 - p_j) + \sum_{i:D_{ij}=0} (1 - W_i) \ln q_j \qquad (18) \end{aligned}$$

A necessary condition at a maximum of the above with respect to parameter p_j is that the first derivative equal zero. On differentiating $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ with respect to parameter p_j and solving for 0, we find (similarly for $\hat{q_j}$)

$$\hat{p_j} = \frac{\sum_{i:D_{ij}=1} W_i}{\sum_{i:D_{ij}=1} W_i + \sum_{i:D_{ij}=0} W_i}, \tag{19}$$

$$\hat{q_j} = \frac{\sum_{i:D_{ij}=0} (1 - W_i)}{\sum_{i:D_{ij}=1} (1 - W_i) + \sum_{i:D_{ij}=0} (1 - W_i)}. \tag{20}$$

We can interpret the weight estimate W_i as the strength of belief in the underlying ground truth being equal to 1. In the case of perfect knowledge about the ground truth, i.e. $W_i \in \{0.0, 1.0\}$, the estimator of sensitivity given by Equation 19 corresponds to the usual definition of sensitivity as the true positive fraction. When the ground truth is a continuous parameter, i.e. $W_i \in [0, 1]$ as considered here, the estimator can be interpreted as the ratio of the number of true positive detections to the total amount of ground truth $T_i = 1$ voxels believed to be in the data, with each voxel detection weighted by the strength of belief in $T_i = 1$. Similarly, the specificity estimator of Equation 20 is a natural formulation of an estimator for the specificity given a degree of belief in the underlying $T_i = 0$ state.

3 Results

We describe experiments to characterize our algorithm running on synthetic data for which the ground truth is known, for assessment of an algorithm and experts segmenting a brain tumor from MRI, and for assessing segmentations from MRI of the peripheral zone of the prostate.

Table 1. Digital phantom consisting of class 1 set to cover half the image, with each expert segmentation identical to the ground truth. In each case, thresholding the converged ground truth estimate at 0.5 recovered the known ground truth exactly. In each of the experiments with two experts, each generating a segmentation identical to the known ground truth but with imperfect initial estimates of expert quality, the final ground truth weights are identical to the known ground truth and the algorithm has discovered the experts are operating perfectly, even when the a priori probability for ground truth was varied.

# experts	initial p_j, q_j	$Pr(T_i = 1)$	final $\hat{p_j}, \hat{q_j}$	# iterations	$W_i \vert T_i = 1$, $W_i \vert T_i = 0$
1	{0.9, 0.9}	0.5	{0.9, 0.9}	1	0.9, 0.1
1	{0.9, 0.9}	0.4	{0.95, 0.80}	22	0.76, 0.04
1	{0.9, 0.9}	0.6	{0.80, 0.95}	22	0.96, 0.24
2	{0.9, 0.9}, {0.9, 0.9}	0.5	{1.0, 1.0}, {1.0, 1.0}	4	1.0, 0.0
2	{0.9, 0.9}, {0.9, 0.9}	0.4	{1.0, 1.0}, {1.0, 1.0}	4	1.0, 0.0
2	{0.9, 0.9}, {0.9, 0.9}	0.6	{1.0, 1.0}, {1.0, 1.0}	5	1.0, 0.0

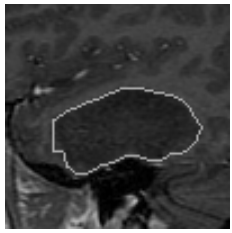
Table 2. Ground truth class 1 was set to be a small square occupying roughly 11% of the 256x256 pixel image. In each experiment, thresholding the converged ground truth weights at 0.5 recovers the known ground truth exactly. In each of the multi-expert experiments, the ground truth weights converged to $W_i = 1$ at the class 1 voxels and $W_i = 0$ at the class 0 voxels. In the final experiment, one of the experts segmentation was set equal to the ground truth, and one was set equal to the ground truth shifted left 10 columns, and one was set equal to the ground truth shifted right 10 columns. The algorithm was able to discover one of the experts was generating a segmentation identical to the ground truth, and that the other two experts were slightly incorrect. The correct ground truth was indicated by the final ground truth weights and the quality estimates for each expert accurately reflected the segmentations.

# experts	initial $p_j = q_j$	$Pr(T_i = 1)$	final $\hat{p_j}, \hat{q_j}$	# itns	$W_i \vert T_i = 1$, $W_i \vert T_i = 0$
1	0.9	0.5	{0.217, 0.997}	9	0.98, 0.44
1	0.9	0.12	{0.696, 0.973}	21	0.78, 0.04
2	0.9, 0.9	0.5	{1.0, 1.0},{1.0, 1.0}	8	1.0,0.0
2	0.9, 0.9	0.12	{1.0, 1.0},{1.0, 1.0}	4	1.0,0.0
3	0.9, 0.9, 0.9	0.12	{0.88,0.99},{1.0,1.0},{0.88,0.99}	11	1.0, 0.0

Results of Experiments with Digital Phantoms with Known Ground Truth Table 1 illustrates experiments with known ground truth and each class occupying half the image. Table 2 illustrates experiments with known ground truth class 1 set to be a small square occupying roughly 11% of the 256x256 pixel image.

Validation of Brain Tumor Segmentation from MRI Figure 1 illustrates our algorithm applied to the analysis of expert segmentations of a brain tumor [10]. The analysis indicates the program is generating segmentations similar to that of the experts, with higher sensitivity than one of the experts, but with lower sensitivity than two other experts.

(a) border of ground truth estimate.



(b) sum of manual segmentations.

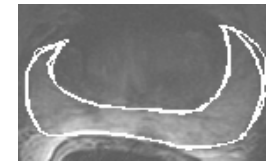
(c) expert quality assessment.

expert	$\hat{p}_j$	$\hat{q}_j$
1	0.8951	0.9999
2	0.9993	0.9857
3	0.9986	0.9982
program	0.9063	0.9990

Fig. 1. Part of MRI with a brain tumor visible and the border of the estimated ground truth overlayed, sum of expert segmentations, and quality assessment of three experts and a program. The results indicate the program is comparable to the experts, performing better than one expert and not as well as two other experts.

(a) ground truth estimate

(b) sum of manual segmentations



(c) overlay of (a) on MRI

Fig. 2. Ground truth segmentation estimate, sum of manual segmentations, and an overlay of borders of regions from the ground truth estimate on the MRI. The voxel intensity in (a) is proportional to the probability of the ground truth being 1 at each voxel. The overlayed borders of regions of different ground truth probability are readily appreciated in (c). The lowest probability is in regions where the expert segmentation is most difficult. The ground truth estimates rapidly reaches a near final configuration. The final result does not depend strongly upon the expert parameter initialization or the ground truth prior probability. **A binary estimate of the ground truth can be made by thresholding the ground truth estimate at 0.50, and in this case is equivalent to taking all the voxels indicated as prostate peripheral zone by at least three of the five segmentations, and then all the voxels indicated by two of the high quality segmentations but not those voxels indicated by only two of the lower quality segmentations.** Recall that our algorithm simultaneously estimates the quality of each of the segmentations together with the probabilistic ground truth estimate. This result cannot be achieved by a simple voting rule such as selecting voxels indicated by three out of five segmentations.

Validation of Prostate Peripheral Zone Segmentation from MRI Figure 2 illustrates our algorithm applied to the analysis of five segmentations by one expert of the peripheral zone of a prostate as seen in a conventional MRI scan (T2w acquisition, 0.468750 x 0.468750 x 3.0 mm^3). The ground truth estimates rapidly reaches a near final configuration, and more slowly refines the last few significant figures of the ground truth and expert parameter estimates. The final result does not depend strongly upon the expert parameter initialization or the ground truth prior probability. The final expert parameter estimates found are record in Table 3.

Table 3. Quality estimates for the five prostate segmentations generated by one expert, and Dice similarity coefficient (DSC), $DSC \equiv \frac{2|A \cap B|}{|A|+|B|}$ with $|A|$ the area of region A, comparing the expert segmentation with the ground truth estimate with $T_i \geq 0.5$. Results are shown for two different assumptions of the prior probability of prostate peripheral zone being present at each voxel and indicate the final estimates do not depend strongly on the ground truth prior probability assumption. The rank order of the experts is the same in each case. Note that our algorithm provides more information than DSC, for example, DSC of expert 2 and 4 is similar, but they have quite different sensitivities.

Expert segmentations	1	2	3	4	5
$g(T_i = 1) = 0.10$					
final $\hat{p_j}$	0.87509	0.987198	0.921549	0.907344	0.880789
final $\hat{q_j}$	0.999163	0.994918	0.999435	0.999739	0.999446
DSC	0.927660	0.957527	0.954471	0.949058	0.932096
$g(T_i = 1) = 0.02$					
final $\hat{p_j}$	0.878533	0.991261	0.936831	0.918336	0.894861
final $\hat{q_j}$	0.998328	0.993993	0.99932	0.999359	0.999301
DSC	0.913083	0.951027	0.967157	0.954827	0.944756

4 Discussion and Conclusion

We have presented an algorithm for simultaneously constructing an estimate of the "ground truth" segmentation from a collection of segmentations, and an estimate of the quality of each segmentation generator. This can be used to assess new segmentations by direct comparison to the ground truth estimate.

At least one digital brain phantom [2] was constructed from segmentations obtained from high signal-to-noise images by manual correction of the output of an automated segmentation algorithm. The approach described here provides a straightforward and principled way to combine manual segmentations to provide a "ground truth" estimate for the construction of such phantoms.

Acknowledgements

This investigation was supported by NIH grants P41 RR13218, P01 CA67165, R01 RR11747, R01 CA86879, R33 CA99015, R21 CA89449, by a research grant from the Whitaker Foundation, and by a New Concept Award from the Center for Integration of Medicine and Innovative Technology. We gratefully acknowledge the contributions of the experts who generated the segmentations of the clinical data, and of Dr. Clare Tempany for helpful discussions regarding the detection of the prostate peripheral zone, and its clinical significance.

References

1. M. Styner, C. Brechbühler, G. Székely, and G. Gerig, "Parametric estimate of intensity inhomogeneities applied to MRI," *IEEE Transactions On Medical Imaging*, vol. 19, pp. 153–165, March 2000.

2. D. Collins, A. Zijdenbos, V. Kollokian, J. Sled, N. Kabani, C. Holmes, and A. Evans, "Design and Construction of a Realistic Digital Brain Phantom," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 463–468, June 1998.
3. L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
4. P. Jaccard, "The distribution of flora in the alpine zone," *New Phytologist*, vol. 11, pp. 37–50, 1912.
5. A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation," *IEEE Transactions on Medical Imaging*, vol. 13, pp. 716–724, December 1994.
6. S. Warfield, J. Dengler, J. Zaers, C. R. Guttmann, W. M. Wells III, G. J. Ettinger, J. Hiller, and R. Kikinis, "Automatic Identification of Grey Matter Structures from MRI to Improve the Segmentation of White Matter Lesions," *J Image Guid Surg*, vol. 1, no. 6, pp. 326–338, 1995.
7. S. K. Warfield, R. V. Mulkern, C. S. Winalski, F. A. Jolesz, and R. Kikinis, "An Image Processing Strategy for the Quantification and Visualization of Exercise Induced Muscle MRI Signal Enhancement," *J Magn Reson Imaging*, vol. 11, pp. 525–531, May 2000.
8. 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.
9. S. K. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis, "Adaptive, Template Moderated, Spatially Varying Statistical Classification," *Med Image Anal*, vol. 4, pp. 43–55, Mar 2000.
10. M. Kaus, S. K. Warfield, A. Nabavi, P. M. Black, F. A. Jolesz, and R. Kikinis, "Automated Segmentation of MRI of Brain Tumors," *Radiology*, vol. 218, pp. 586–591, Feb 2001.