Web-based Multi-observer Segmentation Evaluation Tool

Yaoyao Zhu [a], Xiaolei Huang [a], Daniel Lopresti [a]
L. Rodney Long [b], Sameer Antani [b], Zhiyun Xue [b], George Thoma [b]

[a] *Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015, USA*
{yaz304,xih206,dal9}@lehigh.edu

[b] *National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA*
{rlong,santani,xuez,gthoma}@mail.nih.gov

Abstract

Multi-observer segmentation evaluation is useful in the imaging community. We have developed a web-based software application for automatic performance evaluation of multiple image segmentations, based on the Baysian Decision framework. It computes a probabilistic estimate of the true segmentation (ground truth map) and performance measures for the individual segmentations (sensitivity and specificity). The strength of the tool is that it integrates two kinds of prior knowledge about segmentations: the truth prior (the prior probability of different segmentations) and the observer prior (the performance measures of observers), which can generate more accurate evaluations.

1. Introduction

Segmentation is a fundamental problem in many pattern recognition and image processing applications. Multiple-observer segmentation evaluation is helpful in many scenarios such as evaluating the performance of multiple observers' segmentations simultaneously [4] or measuring segmentation complexity [3] and so on. In order to get more accurate evaluation, there are two kinds of commonly-used prior knowledge that can be integrated into multi-observer segmentation evaluation. One is the truth prior, *i.e.*, the prior probability, which is defined as the probability of a pixel inside the segmentation (foreground) in a binary classification. The other is the observer prior, *i.e.*, the performance measures of observers. Based on different scenarios where there are different evaluation needs and different prior knowledge available, our tool uses Bayesian Decision theory and the MAP optimization principle for multi-observer segmentation evaluation. One method of combining the segmentations from multiple observers is Majority Vote Rule [2]. However, unlike our tool, it does not take into consideration the variability in quality or performance among the voters and also does not incorporate any prior knowledge regarding segmentations. Our tool is also different from the STAPLE algorithm [4], which is a well-known multi-observer segmentation evaluation algorithm. Our tool is more flexible and can handle more scenarios than the STAPLE algorithm. For example, in the STAPLE algorithm the truth prior is dominating so that in certain scenarios when the observer prior is available, this information can not be effectively used to positively influence the results.

Our tool has been used to evaluate multi-observer segmentations for medical images such as those in the NCI/NLM medical repository of digital cervicographic images (cervigrams). In the database, multiple observers have marked several important regions on cervigrams that are of anatomical or clinical interest. Our tool can combine multiple observers' segmentations to generate a more accurate ground truth map for each image and to produce a performance level estimate of each segmentation.

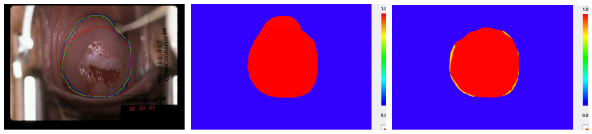
2. Methods and Experimental Results

We choose sensitivity p and specificity q [4] to measure the performance level of each binary segmentation. The result after probabilistically combining multiple segmentations is usually presented as a ground truth map. In the map, each pixel is represented by a color indicating the probability that it belongs inside the ground truth segmentation.

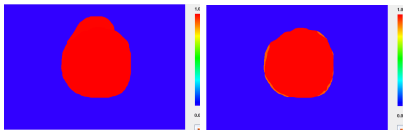
In our method, we explicitly take into account two kinds of prior knowledge: the truth prior ($f(T_i = 1)$) for pixel i and the observer performance-level prior (p, q) values. If a certain prior is unknown, it can be initialized with a uniform distribution or initialized based on observers' segmentation data. Then the Bayesian Decision Theory [1] is used to make a decision based on the posterior probability distribution $f(T|D)$ (D is a matrix describing the binary decisions made for each segmentation). The maximum a posteriori

Exp.	γ		Observer1 (red)	Observer2 (green)	Observer3 (blue)	Result
1	0.5	p q	0.9999 0.9999	0.9999 0.9999	0.9999 0.9999	Figure1(b)
2	0.5	p q	0.9999 0.9999	0.9999 0.7	0.9999 0.7	Figure1(c)
3	0.710	p q	0.9999 0.9999	0.9999 0.9999	0.9999 0.9999	Figure1(d)
4	0.710	p q	0.9999 0.9999	0.9999 0.7	0.9999 0.7	Figure1(e)

Table 1. Initializing the prior probability with $\gamma = 0.5$ and by data



(a) Original image (b) Result for Exp. 1 (c) Result for Exp. 2



(d) Result for Exp. 3 (e) Result for Exp. 4

Figure 1. Estimated ground truth maps with the setups in Table 1

(MAP) estimator can be applied to selecting the most probable ground truth T. Therefore, we have, for any pixel i, let

$$A_i = f(D_{ij}|T_i = 1) = (\prod_{j:D_{ij}=1} p_j \prod_{j:D_{ij}=0} (1 - p_j))f(T_i = 1) \quad (1)$$

$$B_i = f(D_{ij}|T_i = 0) = (\prod_{j:D_{ij}=0} q_j \prod_{j:D_{ij}=1} (1 - q_j))f(T_i = 0) \quad (2)$$

Then

$$f(T_i = 1|D) = \frac{f(D|T_i = 1)f(T_i = 1)}{\sum_T f(D|T_i)f(T_i)} = \frac{A_i}{A_i + B_i} \quad (3)$$

where $f(T_i = 1|D)$ indicates the posterior probability of the true segmentation at pixel i being equal to one. Thus the MAP estimator will assign the class label of pixel i to be 1 (*i.e.*, foreground pixel, $T_i = 1$) if $f(T_i = 1|D) > 0.5$, or assign the label 0 (i.e. background pixel, $T_i = 0$) if $f(T_i = 1|D) < 0.5$.

Next we discuss several scenarios with different prior knowledge available and show experimental results.

2.1. Scenario one: truth prior probability $f(T_i = 1)$ is unknown

In this scenario, we do not know the truth prior probability. We follow the Bayesian Decision framework and calculate directly $f(T_i = 1|D)$ using Equations (1), (2) and (3); the unknown truth prior probability is modeled through one of two ways:

1. We assume there is no prior available about the ground truth map and initialize with a uniform distribution (*i.e.*, $f(T_i = 1) = f(T_i = 0) = 0.5$). This case is illustrated in Experiment 1 and 2 in Table 1. On the example image, two observers (in green and blue lines) give similar segmentations in the cervigram while the other (in red line) is different from the two. In Experiment 1, each observer has equal (p, q) values while in Experiment 2, Observer 1 is an expert with higher (p, q) values and Observers 2 and 3 are non-experts. The results are consistent with the (p, q) values set for each observer. In Experiment 2, the result leans toward the segmentation by Observer 1, who is an expert (Figure 1(c)).
2. We assume the observers' segmentation data reflect the prior distribution of the true segmentation and thus initialize the prior probability using the data. (The STAPLE algorithm also adopts this initialization scheme in the absence of the truth prior). More specifically, we can either initialize with a single global (homogeneous) prior as the sample mean of the relative proportion of a label in the multiple observers' segmentations [4]:

$$\gamma = \frac{1}{RN} \sum_{j=1}^{R} \sum_{i=1}^{N} D_{ij} \quad (4)$$

or with a spatially-varying prior map as the sample mean of all observers' labels:

$$f(T_i = 1) = \frac{1}{R} \sum_{j=1}^{R} D_{ij} \quad (5)$$

This case is illustrated in Experiments 3 and 4 in Table 1. Similar results were obtained using observers' segmentation data and using a global prior $\gamma = 0.5$ to initialize the truth prior probability.

2.2. Scenario two: observer prior (p, q) values are unknown

In this scenario, the known truth prior is directly applied in Equation (3), while the missing (p, q) values of each observer can be set in two ways:

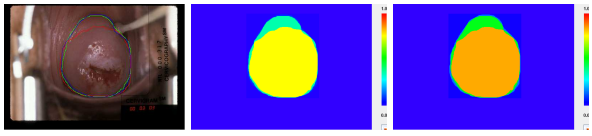
1. We assume everyone has the same performance level thus the same (p, q) values, i.e., $p_i = q_i = t(0 < t < 1)$. Whenever this value changes, the estimated ground

truth probability map changes accordingly, which reflects the changing confidence in the observers. This case is illustrated in Experiments 1, 2 and 3 in Table 2. We clearly see the effect of the truth prior probability.

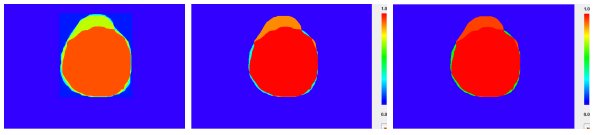
2. We can initialize the (p, q) values of each observer based on the multiple observers' segmentation data. In this case, the sample mean map (Equation (5)) is taken as the prior estimate of the ground truth and a threshold of 0.5 is applied to the probability map to obtain a binary map. Then the initial (p, q) values are calculated. This case is illustrated in Experiments 4 and 5 in Table 2. Each observer has (p, q) values initialized from the segmentation data. We clearly see the effect on the estimated ground truth probability map given changes in the truth prior probability.

Exp.	γ		Observer1 (red)	Observer2 (green)	Observer3 (blue)	Result
1	0.2	Initial p	0.7	0.7	0.7	Figure2(b)
		Initial q	0.7	0.7	0.7	
		Final p	0.9999	0.9999	0.9999	
		Final q	0.899	0.739	0.731	
2	0.3	Initial p	0.7	0.7	0.7	Figure2(c)
		Initial q	0.7	0.7	0.7	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	
3	0.5	Initial p	0.7	0.7	0.7	Figure2(d)
		Initial q	0.7	0.7	0.7	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	
4	0.3	Initial p	0.978	0.990	0.988	Figure2(e)
		Initial q	0.763	0.954	0.961	
		Final p	0.944	0.975	0.972	
		Final q	0.763	0.9999	0.9999	
5	0.5	Initial p	0.978	0.990	0.988	Figure2(f)
		Initial q	0.763	0.954	0.961	
		Final p	0.978	0.99	0.989	
		Final q	0.763	0.954	0.962	

Table 2. Initializing (p, q) values with $t = 0.7$ and by data



(a) Original image (b) Result for Exp. 1 (c) Result for Exp. 2



(d) Result for Exp. 3 (e) Result for Exp. 4 (f) Result for Exp. 5

Figure 2. Estimated ground truth maps with the setups in Table 2

3. Software Design

Our web-based multi-observer segmentation evaluation tool is developed in Java and the architecture of the software is shown in Figure 3.

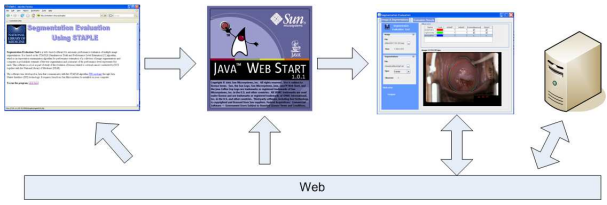


Figure 3. Architecture of the web-based tool

The system consists of three components: the web browser, the application and the server. The web browser is accessible to users by which they download and evoke the Java application. It is made possible by the Java Web Start technology.

The Java application has a user-friendly interface, which has the following features: (1) Loading and viewing the image and segmentation information; (2) Communicating with the server and displaying results. A user may select among the different scenarios implemented in our framework and the options to initialize the missing priors. The application submits the image, multiple-observer segmentations and prior information to the server and receives evaluation results from the server. The estimated ground truth map is shown along with the original image on the application. The position and probability of a pixel can also be shown; (3) Exporting the final results including the posterior ground truth map and the (p, q) values to files in a selected local directory; (4) Quick-start guide. The help documentation for a quick start is developed with JavaHelp 2.0.

The software on the server side includes a Java servlet and algorithms. The Java servlet communicates with the application. It receives the image, observer segmentations, and prior information from the application and sends the results back to the application after the algorithms finish computing.

References

[1] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[2] J.Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[3] S. Lotenberg, H. Greenspan, H. Greenspan, S. Gordon, L. Long, J. Jeronimo, and S. Antani. Automatic evaluation of uterine cervix segmentations. In *Proceedings of the SPIE, Medical Imaging 2007*, volume 6515, March 2007.

[4] S. Warfield, K. Zou, and W. Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, July 2004.