Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Evaluation of Segmentation algorithms for Medical Imaging

Aaron Fenster PhD, Bernard Chiu
Imaging Research Laboratories, Robarts Research Institute, London, ON, CANADA
Graduate Program in Biomedical Engineering
The University of Western Ontario, London, Canada, N6A 5K8

Abstract—**This paper describes an approach to be used for medical image segmentation evaluation. The process for segmenting organs and structures from medical images is gaining increased importance in the diagnosis of diseases and in guiding minimally invasive surgical and therapeutic procedures. While investigators are continuing to develop novel new segmentation approaches, little attention has been given to the development of a uniform and common framework for and performance metrics to be used in comparing different algorithms, in optimizing algorithms and in evaluating their performance. Choosing an appropriate effectiveness measure of object segmentation is a difficult task and weighting the importance of different possible performance metrics requires matching the metrics to the segmentation objectives. However, in all tasks, it is now believed that three types of metrics must be measured and reported: accuracy, precision and efficiency. In this paper, we review some of these metrics.**

I. Introduction

In the analysis of medical images it is often essential that objects/organs/structures be distinguished or segmented from their background. Over the past decade, segmentation techniques have gained importance in the quantitative analysis of medical images and in image guided interventional procedures. Many approaches have now been investigated and some have been implemented in commercial imaging and analysis systems. The literature on segmentation techniques is rich with innovative ideas and algorithms. Because image segmentation is often the first step in the analysis of the information, an appropriate, accurate, precise and efficient approach must be used to minimize erroneous or inappropriate results. In choosing a segmentation technique for a particular task, it is important to understand that: there is no universally applicable segmentation technique that will work for all types of medical images and all organs, and, that no segmentation technique is perfect.

This work was supported in part by Canadian Institute of Health Research and the Ontario R & D Challenge Fund. A. F. holds a Canada Research Chair in Biomedical Engineering, and acknowledges the support of the Canada Research Chair Program. B. C. acknowledges the support by the Ontario Graduate Scholarship.

B. C. is with the Imaging Research Laboratories, Robarts Research Institute, London, Ontario N6A 5K8, Canada and is a PhD graduate student in the Graduate Program in Biomedical Engineering at The University of Western Ontario (e-mail: bchiu@ imaging.robarts.ca).

A. F. is the director of the Imaging Research Laboratories, Robarts Research Institute, London, Ontario N6A 5K8, Canada (Phone: (519) 663-3834; fax: (519) 663-3900; e-mail: afenster@ imaging.robarts.ca).

While a great deal of effort has been spent on the development of segmentation algorithms, a lesser emphasis has been placed on appropriate evaluation of segmentation algorithms. Investigators continue to develop yet more segmentation algorithms but often fail to evaluate them using uniform or standard performance measurement metrics. While most developers of segmentation algorithms do provide some information on evaluation of their algorithm, a framework and consistent approach by which segmentation algorithms performance can be compare is lacking. Although unified theories have been proposed, the problem is not yet resolved [1-6] As a result, finding the appropriate segmentation algorithm for a particular task as well as choosing the optimal parameters internal to the segmentation algorithm is still a problem. Thus, it is necessary to identify appropriate tools for evaluating and comparing segmentation algorithms. In this paper we review and describe such tools as an aid for developers in optimizing and evaluating their algorithms.

II. SEGMENTATION EVALUATION METRICS

Designing or choosing an appropriate effectiveness measure of object segmentation is a difficult task. The evaluation metric should provide information relevant to the task, whether it is diagnostic or interventional. For example, some tasks require real-time operation, such as those used in surgical and interventional procedures, while segmentation tasks for diagnostic procedures can be performed of-line. In these situations, weighting the importance of different performance metrics in choosing the optimal segmentation approach may differ. However, in all tasks, it is now believed that three types of metrics must be measured and reported: accuracy, precision and efficiency.

1 Accuracy

Accuracy of a segmentation technique refers to the degree to which the segmentation results agree with the true segmentation. In situations in which the true segmentations are known, e.g., the true segmentation is known when specially constructed test phantoms are used. However, when dealing with images of patients or research animal models, the true segmentation is not known. In those situations a surrogate measure of "truth" is used. Surrogates for "truth" may be obtained with manual segmentation, a different imaging system, or a segmentation algorithm that is known to produce accurate results. A measure of accuracy should ideally reflect the amount of disagreement between

the "true" and test segmentation, and should not depend on the dimensions of the image [13,14].

1.1 Distance-based metrics

For some segmentation tasks, the delineation of the boundary is critical and is the objective of the segmentation. In these situations, distance-based metrics are important and used to measure the distance between the segmentation generated boundary (test boundary) and the "true" boundary. If the test boundary, *B*, and "true" boundary, *T*, are respectively defined by two sets of vertices $B = \{ b_i : i = 1 \cdots K \}$ and $T = \{ t_n : i = 1 \cdots N \}$, the distance between b_i and *T* is defined by:

$$d(b_i, T) = \min_n \| b_i - t_n \|$$

For each image *j*, 3 parameters can be computed: (a) MAD_j, the mean absolute difference, a measure of the mean error in segmentation, (b) $MAXD_j$, the maximum difference, a measure of the maximum error in segmentation, (c) PC_j, the percentage of the vertices (c_i) from which the distances to the "true" contour *T*, $d(b_i, T)$, are less than *p* pixels. This is used to evaluate the fraction of points that can be classified as "very close" to the "true" contour, or a difference from the "true" contour that is not medically or biologically significant.

$$MAD_j = \frac{1}{K} \sum_{i=1}^{K} d(b_i, T)$$

$$MAXD_j = \max_{i \in [1,K]} \{ d(b_i, T) \}$$

$$PC_j = \frac{no.\ of\ elements\ in \{ b_i \in B : d(b_i, T) < 5\ pixels \}}{K}$$

1.2 Area- or Volume-based metrics

Some segmentation tasks are use to measure the area or volume of the object, e.g., the size of the tumor for staging or the size of the prostate for treatment planning. In these situations, area-based, or more common, volume-based metrics are used to compare the object enclosed by a segmentation boundary and the "true" boundary.

Two approaches may be used to evaluate the segmentation algorithm. In the first, the continuous variable of area or volume can be compared to the "true" value and a percent error can be determined. Since the "true" value is often subject to user or machine variability and the algorithm measured quantity is often result of user interaction with the algorithm [15,16], the mean and variance of the "true" and segmented quantities should be determined and compared using standard statistical methods such as two-way analysis of variance and the t-test [11, 12]

Another set of performance measures borrows the methodology from the object detection literature [7,8,9]. If V_S and V_T represent the regions enclosed by the segmented boundary and the "true" boundary respectively, we define the true positive volume (*TP*) as the volume enclosed by both the "true" and algorithm segmented boundaries i.e., $V_{TP} = V_S \cap V_T$, the false positive (*FP*) volume is $V_{FP} = V_S - V_T$, the false negative (*FN*) volume is $V_{FN} = V_T - V_S$, and the true negative (*TN*) volume is $V_{TN} = SCENE - V_S - V_T$. With these quantities, accuracy metrics can be defined thus:

$$TPF\ \ (True\ \ Positive\ \ Fraction) = \frac{V_{TP}}{V_T} = Sensitivity$$

$$FNF\ \ (False\ \ Negative\ \ Fraction) = \frac{V_{FN}}{V_T} = Specificity$$

$$FPF\ \ (False\ \ Positive\ \ Fraction) = \frac{V_{FP}}{SCENE - V_T}$$

$$= 1 - Specificity$$

$$TNF\ \ (True\ \ Negative\ \ Fraction) = \frac{V_{TN}}{SCENE - V_T}$$

$$= 1 - Sensitivity$$

where SCENE is the region encompassing all possible segmented regions, *TPF* is the volume fraction in the "true" segmented boundary that is also enclosed by the algorithm segmented boundary; the *FNF* is volume fraction enclosed by the "true" boundary that was missed by the segmentation algorithm, the *FPF* is the volume fraction enclosed by the algorithm segmented boundary that was not enclosed by the "true" boundary, and the *TNF* is the volume fraction in the background scene that was not enclosed by the "true" boundary and was not enclosed by the algorithm boundary.

It is important to note that *FPF* and *TNF* can be made arbitrarily small by choosing a large region for *SCENE*. In this case, these values cannot be meaningfully assessed. However, with appropriate choice of *SCENE*, a segmentation algorithm can be optimized by plotting the *TPF* (Sensitivity) versus the *FPF* (1-specificity) as is done in ROC analysis. By varying the segmentation parameters and calculating ROC metrics, the optimal choice may be found [7,8,9].

Since FNF and FPF may be impractical to calculate for reasons discussed above, another metric has been used to assess segmentation accuracy. In this metric, the fraction of false regions segmented by the algorithm is measured, thus:

$$FF\ \ (False\ \ Fraction) = 1 - \frac{V_{FP} + V_{FN}}{V_T}$$

1.3 Metrics for evaluating the performance on the entire set of images

In evaluating the segmentation algorithm, it is important that a sufficient number of images be used in the evaluation. Demonstration of performance and comparison to other approaches using only a few is not sufficient. Since statistical tests should be used to compare performance, the number of images needed may be estimated using estimates of variance and effect size [10].

2. Precision

Precision of a segmentation algorithm provides information on the repeatability of the technique when used to segment a type of image. Sources of variability can be due to subjective observer interactions with the algorithm, e.g., manual initialization, object choice, and interaction of user and object. In evaluation the precision of an algorithm, the appropriate metric (distance, area and volume) is determined for a set of images using both the algorithm and repeated estimates of "truth", e.g., repeated manual segmentations. In this way, repeated segmentations will provide an estimate of the segmentation variance [17,18,19].

Using estimates of variance of the algorithm and of the surrogate for "truth" (e.g., manual segmentation), the algorithm variability may be compared to the variability in determining the "true" segmentation. In addition, comparison of the variability of the algorithm to another algorithm or to the same algorithm but with a different set of parameters can also be evaluated. The comparisons should involve appropriate statistical tests such as analysis of variance (ANOVA) and F statistics [10, 11]. Care must be used in using a sufficient number of segmentation tests to allow meaningful comparisons.

3. Efficiency

Efficiency of the segmentation provides information on the practical use of the algorithm. Often, this is measured as segmentation time, but should include measures of all aspects of user interaction and whether the approach is suitable for all images. Thus, in addition to segmentation algorithm execution time, the time for initialization, editing and inspection should also be documented as well as the failure rate.

Figure 1 A cross-sectional profile of five repeated manual segmentations of the carotid artery.

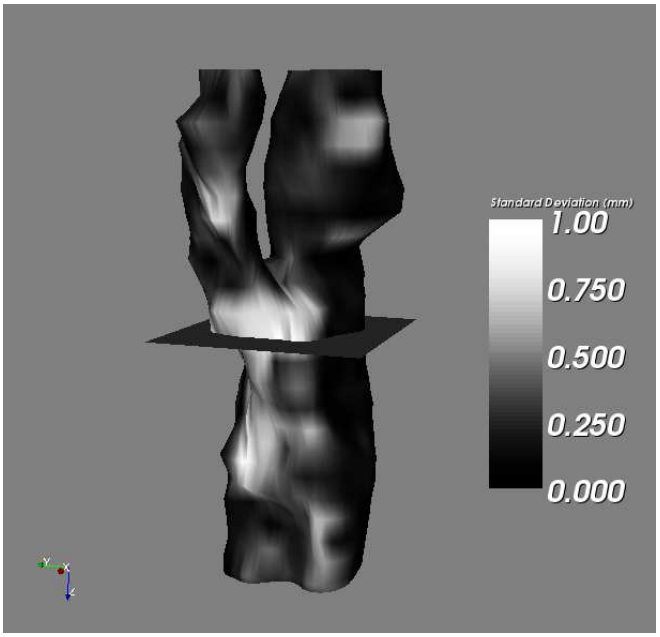

Figure 2 The mean surface of the carotid arteries with the standard deviations of the repeated segmentations superimposed on top. The plane represents the position at which the cross-sectional profile shown in Figure 1 is obtained.

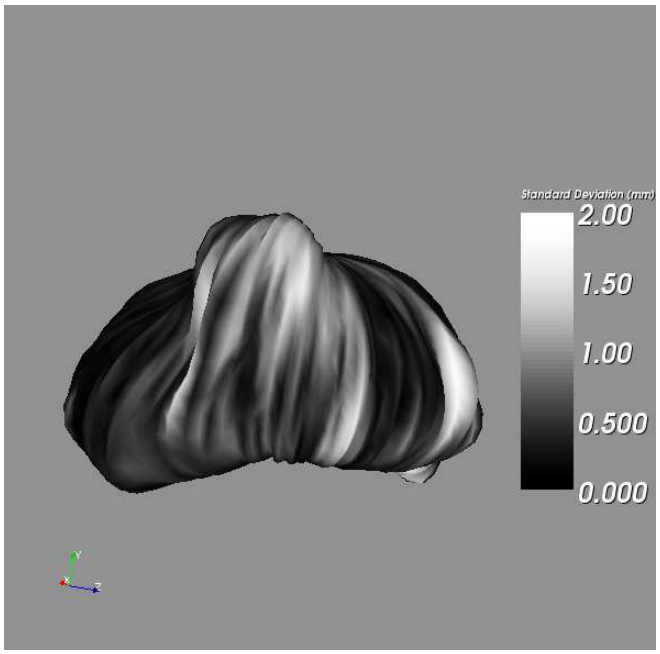

Figure 3 The mean surface of five repeated manual segmentations of the prostate with the standard deviation map superimposed on top

4. Comparison of algorithms

Since a segmentation algorithm will suffer from variability due to variation in the objects to be segmented and due to user interaction, any comparison of the accuracy metrics discussed above must be carried out with the appropriate statistical tests. Thus, an improvement in the segmentation as measured by any metrics discussed above may not be statistically significant because the improvement is small relative to the segmentation variation and/or an insufficient number of cases were used in the evaluation. [10]

ACKNOWLEDGMENT

The authors gratefully acknowledge for software assistance from L. Gardi, and I. Gyacskov, as well as the support from the technical staff at the Imaging Research Laboratories, Robarts Research Institute.

REFERENCES

[1] Y. Zhang, "A survey on evaluation methods for image segmentation," Pattern Recognition, 29:1335-1346, 1996.

[2] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images" IEEE Transactions on Medical Imaging, 16:642-652, 1997.

[3] H. Christensen and W. Foerstner. Special issue on performance evaluation. Machine Vision and Applications 9(5), 1997.

[4] K.W. Bowyer and P.J. Phillips. *Empirical Evaluation Techniques in Computer Vision*. IEEE Computer Society Press, 1998.

[5] M.A. Viergever, H.S. Stiehl, and R. Klette. *Performance Characterization band Evaluation of Computer Vision Algorithms.* Kiuwer Academic Publishing, 2000.

[6] K.W. Bowyer, "Validation of medical image analysis techniques," in: *The Handbook of Medical Imaging*, Eds. J. Beutel, H. Kundel and R. van Metter, SPIE: Bellingham Washington, 2000.

[7] V. Chalana and Y. Kim. "A methodology for evaluation of boundary detection algorithms on medical images". IEEE Transactions on Medical Imaging 16:642-652, 1997.

[8] C.E. Metz. "ROC methodology in radiologic imaging". Investigative Radiology 21:720-733, 1986.

[9] B.J. McNeil and J.A. Hanley. "Statistical approaches to analysis of receiver operating characteristic ROC curves". Medical Decision Making 14:137-150, 1984.

[10] J.P. Egan. Signal Detection Theory and ROC Analysis. Academic Press. New York, NY 1975.

[11] M. Bland. *An Introduction to Medical Statistics*. Oxford University Press, 1995.

[12] J. H. Zar. *Biostatistical Analysis*. Prentice Hall Inc. 1984.

[13] P. Saha, J. Udupa, D. Odhner, "Scale-based fuzzy connected image segmentation: Theory, algorithms and validation," Computer Vision and Image Understanding 77:145-174, 2000.

[14] M. Kamber, R. Shinghal, D. Collins, G. Francis, and A. Evans, "Model-based 3D segmentation of multiple sclerosis lesions in magnetic resonance brain images," IEEE Transactions on Medical Imaging, 14:442-452.

[15] Mitchell JR, Jones C, Karlik SJ, Kennedy K, Lee D, Rutt B, Fenster A. "Magnetic Resonance Multispectral Analysis of Multiple Sclerosis Lesions," Journal of Magnetic Resonance Imaging 7(3): 499-511, 1997.

[16] Mitchell JR, Karlik S, Lee DH, Ekiaszim M, Rice GP, Fenster A. *The Variability of Manual and Computer Assisted Quantification of Multiple Sclerosis Lesion Volumes*. Medical Physics 23(1): 85-97, 1996.

[17] Chiu B, Freeman GH, Salama MMA, Fenster A. *Prostate segmentation algorithm using dyadic wavelet transform and discrete dynamic contour*. Physics in Medicine and Biology. 49: 4943-4960, Oct 2004.

[18] Wang Y, Cardinal N, Downey D, Fenster A, *Semi-automatic 3D segmentation of the prostate using 2D ultrasound images.* Medical Physics. 30(5): 887-97, 2003.

[19] Ladak H, Wang Y, Downey D, Fenster A, Testing and Optimization of a Semi-Automatic Prostate Boundary Segmentation Algorithm using Virtual Operators. Medical Physics. 30(7): 1637-47, 2003.