Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

Composite Index for the Quantitative Evaluation of Image Segmentation Results

[1]F. Alonso, [1]M.E. Algorri, [2]F. Flores-Mangas
[1] Department of Digital Systems, Instituto Tecnológico Autónomo de México, Mexico City, México
[2]IIMAS, Universidad Nacional Autónoma de México, México City, México

Abstract—**Medical image segmentation is one of the most productive research areas in medical image processing. The goal of most new image segmentation algorithms is to achieve higher segmentation accuracy than existing algorithms. But the issue of quantitative, reproducible validation of segmentation results, and the questions: What is segmentation accuracy ?, and: What segmentation accuracy can a segmentation algorithm achieve ? remain wide open. The creation of a validation framework is relevant and necessary for consistent and realistic comparisons of existing, new and future segmentation algorithms. An important component of a reproducible and quantitative validation framework for segmentation algorithms is a composite index that will measure segmentation performance at a variety of levels. In this paper we present a prototype composite index that includes the measurement of seven metrics on segmented image sets. We explain how the composite index is a more complete and robust representation of algorithmic performance than currently used indices that rate segmentation results using a single metric. Our proposed index can be read as an averaged global metric or as a series of algorithmic ratings that will allow the user to compare how an algorithm performs under many categories.**

Keywords—**Quantitative Validation, Performance Evaluation, Error Index, Segmentation Results**

I. Introduction

Image segmentation is no doubt one of the largest research areas in medical image processing. New algorithms, algorithmic techniques, methodologies and improvements to existing methods are continuously being proposed to segment medical images into their constituent organs and tissues. For newcomers to the field, clinicians interested in practical segmentation applications and even researchers in image processing it is difficult and even confusing to choose one segmentation technique over another and to understand the benefits and disadvantages of each method. Whether image segmentation is performed using statistics[1], fuzzy logic[2], neural networks[3], active contours[4], morphology[5], mathematical models[6], texture[7], any combination of the previous methods[8] or any other technique, depends on many factors, the main one being the nature of the medical images to be segmented (MR, CT, Xray, utlrasound, etc.). Other factors for choosing an algorithmic approach are: 1) available technology, 2) degree of segmentation automatization, 3) prior information available on the images, and 4) number of organs and tissues to be segmented. Whatever the method, one of the main objectives of every segmentation is to achieve a high segmentation accuracy. But to be able to measure the accuracy of any image segmentation method, the true segmentation of the images must be known. In the last decade, manual segmentations by clinical specialists were used as gold standard against which to compare automatic segmentations. However, manual segmentations are subject to error and cannot be precisely duplicated. For this reason, the McGill group [9]-[11] have developed an MRI simulator called Brainweb and have made available simulated MR head images where the true segmentation is known and so a quantitative validation of segmentation methods can be obtained. In [12], a Harvard group has also made available 20 MR head image sets along with their segmentations to be used as standard test sets, however, in the Harvard images, the segmentations were performed manually. But even with the Brainweb and Harvard images available, it has proved hard to rate one image segmentation method against another, precisely because a unique definition of segmentation accuracy does not exist. Some of the many definitions of segmentation accuracy that can be established to rate an experimentally segmented image set against the true segmentation are: 1) number of coincident and/or non coincident segmented pixels, 2) number of coincident and/or non coincident segmented pixels on the boundaries of tissues or organs, 3) preservation of area, 4) preservation of perimeter, 5) preservation of the smoothnes or degree of curvature of the segmented boundary, 6) preservation of the statistics, 7) preservation of the entropy, 8) preservation of the center of mass, 9) preservation of the topology of the segmented regions, etc. In image segmentation literature we usually find simple metrics to quantitatively evaluate segmentation results: the Tanimoto Index[13], the Overlap Metric[14] and the Misclassification Rate[15], are a few of the error indices that are used. However, because these indices consist of a single metric to measure the segmentation accuracy or segmentation error, they do not represent a complete performance metric of segmentation algorithms. For example a number of different segmentation results can lead to the same Tanimoto Index, and a measure of the Misclassification Rate is hardly a precise indicator of segmentation accuracy. Our experience in medical image processing has led us to believe that a quantitative evaluation index must be proposed that truly captures the complex information contained in segmentation results, and that this evaluation index must be able to give the user a realistic idea of algorithmic performance at different levels. Without such an index, new segmentation algorithms will not be able to realistically compare or measure themselves against existing work, and the question of what

0-7803-8439-3/04/$20.00©2004 IEEE

segmentation accuracy a segmentation algorithm can reach will continue under determined.

In this paper we propose a prototype composite index for the quantitative evaluation of image segmentation results. This prototype index is composed of seven metrics that are obtained from experimentally segmented image sets measured against the true segmentation of the same image sets. The seven metrics that are measured over the segmented regions in the images are: preservation of mean and standard deviation, preservation of perimeter and area, Tanimoto index, Overlap metric and Misclassification rate.

II. Methodology

We use one Brainweb simulated MR image set and one Harvard MR image set to quantitatively evaluate two segmentation algorithms. We will call the Brainweb image set, test set 1, and the Harvard image set, test set 2. For both test sets the greyscale MR images and the true segmentations were available. We programmed two segmentation algorithms, one that segments regions in an image using image statistics, which we call segmenter 1 and another one that segments regions using a combination of statistics and fuzzy logic, which we call segmenter 2. Our current objective is to program more segmentation algorithms to develop a framework for the composite quantitative assessment of segmentation performance for the most widely used algorithmic techniques.

A. Segmentation and Division in Image Regions

We first segmented the white matter in both test sets with both segmenters. We then divided the resulting segmented images and the true segmentation images with a regular grid into 16 square regions. To add precision to our quantitative evaluation, we decided, in our current prototype framework, to measure the composite index 16 times on every image, that is, we measured the composite index on every square region in the images. This distributed approach gives a much more precise error measurement, since errors throughout the images will not cancel out, and also, this allows us to take into account intra-image variabilities. We measured the 7 metrics of the composite index on each of the 16 regions in 43 central images of each segmented image set.

B. Metrics 1-2. Preservation of the Mean and Standard Deviation in the Image Regions

For test set 1 we have three segmentations: the true segmentation, the segmentation provided by segmenter 1 (statistical) and the segmentation provided by segmenter 2 (statistical plus fuzzy logic). The images in the 3 segmented sets are also divided into 16 regions each. Region by region we compare the mean and standard deviation in the true segmentation vs the mean and standard deviation of the image regions in the two experimental segmentations. For each region in the experimental segmentations we obtain 2 measures: the absolute difference between the mean of the white matter and the mean of the true white matter (the white matter in the true segmentation), and also the absolute difference between their standard deviations. We repeat the measurement of absolute difference in mean and standard deviation for the regions of test set 2 and its two experimental segmentations .When the value of mean and standard deviation metrics is cero, perfect preservation of region statistics is achieved.

C. Metrics 3-4 Preservation of Area and Perimeter in the Image Regions

We repeated the same steps as for Metrics 1-2, this time measuring, region-wise, the absolute difference in area and perimeter between both test sets and their two segmentations. A measurement of cero indicates perfect area and perimeter preservation of segmented white matter.

D. Metrics 5-7 Measurement of Common Error Indices

In order to maintain a comparative measure with many of the existing segmentation validation results, we add the Tanimoto Index, the Overlap Metric and the Misclassification Rate to our composite index. These indices, as we did with metrics 1-4 are measured 16 times on every image for all experimentally segmented image sets. Values of the Tanimoto Index and the Overlap Metric of one indicate perfect agreement between the experimental and true segmentations[13][14]. An MCR of cero indicates a perfect segmentation[15].

III. Results

A. Figures and Tables

Fig. 1 shows one segmentation example for each test set, Figs. 1a and 1e show the original Brainweb (test set 1) and Harvard (test set 2) images to be segmented, 1b and 1f show the true segmentations also provided by the McGill and Harvard sites. Figs. 1c and 1g show the white matter segmentation obtained with our statistical segmentation algorithm (segmenter 1), finally Figs 1d and 1h show the white matter segmentation obtained by the statistical/fuzzy segmentation algorithm (segmenter 2). All images in Fig.1 show the grid we use to subdivide the images in 16 regions.

B. Region-wise Measurement of the Seven Metrics

Figs 2-3 show the composite quantitative evaluation of the two experimental segmentations of region 6 of test set 1. Fig 2 shows the normalized values of the seven metrics measured on region 6 over the 43 images in the segmented set produced by segmenter 1. Fig 3 shows the same results

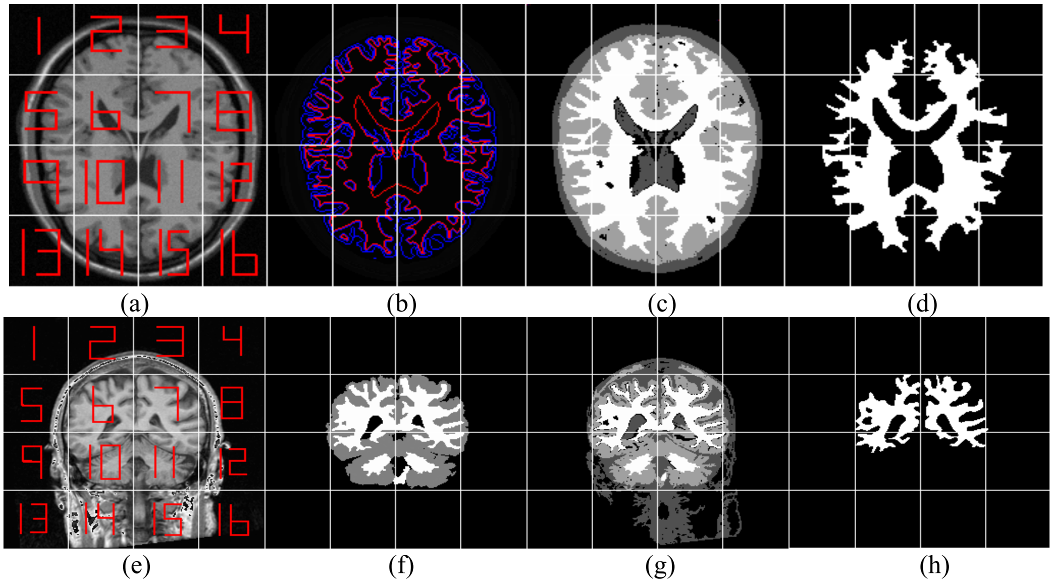

Fig 1. a), e) Original images from test set 1 and test set 2. b), f) True Segmentations. c), g) Results of segmenter 1. d), h) Results of segmenter 2.

for segmenter 2. Figs 4-5 show the normalized values of the seven metrics for region 6 along the 43 segmented images by segmenters 1 and 2 on test set 2. As can be seen on Figs 2-5, the metrics of absolute difference of mean, σ, area and perimeter between the experimental and true segmentations take values close to cero, meaning there is a good preservation of statistics and geometry on the segmented results of both segmenters. The values of the Tanimoto index and Overlap Metric also take values close to 1, indicating good agreement between experimental and true segmentations. Finally the MCR is close to 0 as expected. But on closer examination of the composite index applied to test set 1, one might conclude that segmenter 1 better preserves the area and the perimeter than segmenter 2, even though the values of the Tanimoto index are very similar. For test set 2 both segmenters have similar behaviour with respect to area and perimeter preservation. Over test set 2, segmenter 1 outperforms segmenter 1 on MCR values.

IV. Discussion

The proposed prototype index allows us to realistically compare the performance of segmenter 1 and 2. According to the composite index, segmenter 1 better preserves the area and perimeter of segmented test set 1 than segmenter 2, even though both segmenters have similar Tanimoto indices and Overlap Metrics. We only present the results for one segmented region, because applying the composite index region-wise over the images gives a more precise measurement of segmentation results. The distributed application of the index allows the user to identify problem zones in the images and to detect the algorithm's weak points. The prototype index must be included in a complete validation framework that specifies, among other things, the right methodology for segmentation validation, including a description of how the composite index is measured and reported. Our next step will be to incorporate measures of the segmented boundary curvature, and preservation of entropy and topology.

V. Conclusion

In this paper we present our initial results towards a complete framework for the quantitative validation of segmentation algorithms. We introduce a prototype composite index consisting of seven validation metrics. Our proposed index can be read as a global metric or as a series of algorithmic ratings that will allow the user to compare how an algorithm performs under many categories. The composite index allows users to make statements of the form: While segmentation algorithm 1 performs better in area and perimeter preservation, algorithm 2 better preserves boundary curvature and tissue statistics. To complete the validation framework we will program some of the more popular image segmentation algorithms, trying to represent all algorithmic categories. We plan on comparing active contour models, neural networks, statistical algorithms, fuzzy algorithms, segmentation by texture and morphological operators. We will segment the same test images using all the algorithms and will then assign a composite index to each segmentation result. We will also incorporate additional metrics to the composite index. We believe a validation framework is relevant and necessary for uniform, and realistic comparisons of current and future segmentation algorithms. Such comparisons are not possible with the single metric indices that are currently used in literature for quantitative segmentation evaluation.

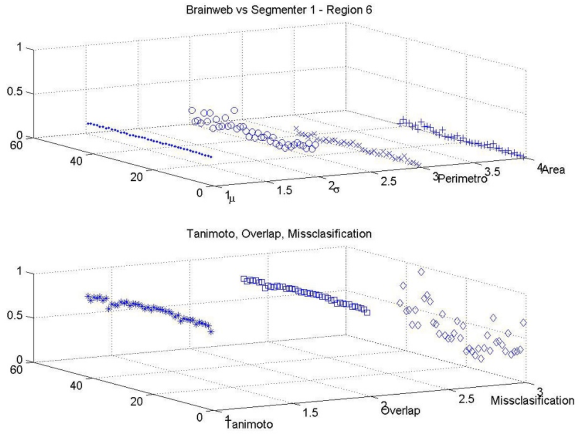

Fig 2. Quantitative results of composite index for segmenter 1 compared to the true segmentation of test set 1 (region 6). above) Absolute difference in mean, σ, perimeter and area, below) Tanimoto, Overlap Metric and MCR.

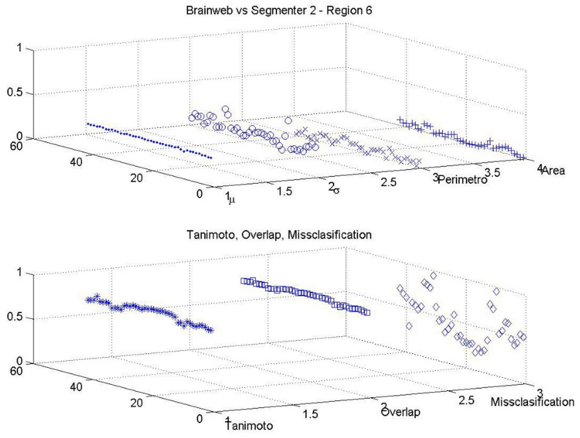

Fig 3. Quantitative results of composite index for segmenter 2 compared to the true segmentation of test set 1 (region). above) Absolute difference in mean, σ, perimeter and area, below) Tanimoto, Overlap Metric and MCR.

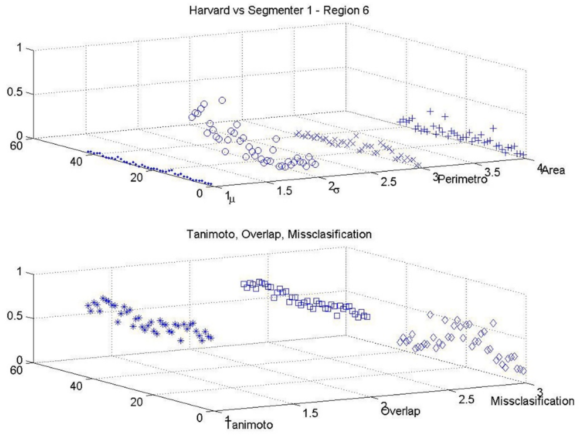

Fig 4. Quantitative results of composite index for segmenter 1 compared to the true segmentation of test set 2 (region 6). above) Absolute difference in mean, σ, perimeter and area, below) Tanimoto, Overlap Metric and MCR.

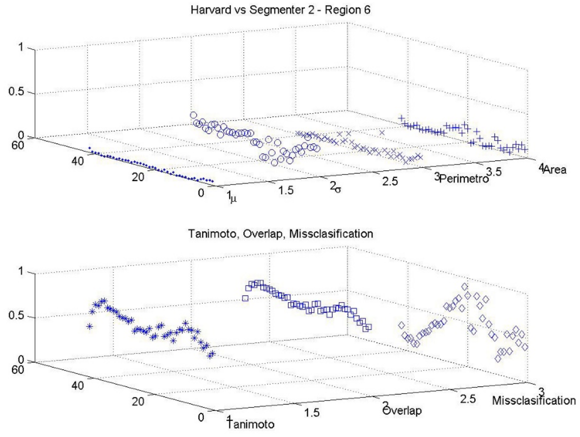

Fig 5. Quantitative results of composite index for segmenter 2 compared to the true segmentation of test set 2 (region 6). above) Absolute difference in mean, σ, perimeter and area, below) Tanimoto, Overlap Metric and MCR.

REFERENCES

[1] J. C. Rajapakse, J.N. Giedd, and J.L.Rapoport, "Statistical Approach to Segmentation of Single-Channel Cerebral MR Images", *IEEE Trans. Med. Imag*., vol. 16, pp.176-186, Apr. 1997.

[2] D. L Pham, and J.L Prince, "Adaptive Fuzzy Segmentation of Magnetic Resonance Images", *IEEE Trans. Med. Imag*. vol. 18, no.9, pp.737-752, Sept. 1999.

[3] M. Ozkan, B.Dawant, and R. Maciunas, "Neural-network-based segmentation of multimodal medical images: A comparative and prospective study," *IEEE. Trans. Med. Imag*., vol. 12, pp.534-544, Sept. 1993.

[4] C. Han, W. S. Kerwin, T. S. Hatsukami, J-N. Hwang, and C. Yuan, "Detecting Objects in Image Sequences Using Rule-Based Control in an Active Contour Model*", IEEE Trans. Biomedical Engineering*, vol. 50, no. 6, pp705-710, Jun. 2003.

[5] J. Serra,"Image Analysis and Mathematical Morphology: vol. 1", Academic Press, 1984

[6] G. B. Aboutanos, J. Nikanne, N. Watkins, and B. M. Dawant, "Model Creation and Deformation for the Automatic Segmentation of the Brain in MR Images", *IEEE Trans. Biomedical Engineering*, vol. 46, no. 11, Nov. 1999.

[7] J. Krumm, S.A. Shafer, "Texture segmentation and shape in the same image", *IEEE 5th Int.Conf. on Comp.Vision*, pp. 121, June 20-23 1995

[8] M. E. Algorri, F. Flores-Mangas, "Classification of Anatomical Structures in MR Brain Images using Fuzzy Logic*", IEEE Trans. Biomedical Engineering*, to appear

[9] R. K.-S. Kwan, A.C.Evans, and G.B. Pike, " MRI Simulation-Based Evaluation of Image-Processing and Classification Methods", *IEEE Trans. Med. Imag*., vol. 18, no. 11, pp. 1085-1097, Nov. 1999.

[10] D.L. Collins, A.P. Zijdenbos, V. Kollokian, J.G. Sled, N.J. Kabani, C.J. Holmes, A.C. Evans , "Design and Construction of a Realistic Digital Brain Phantom" , *IEEE Trans. Med. Imag*., vol.17, No.3, p.463--468, June 1998.

[11] McGill Brainweb Simulator. http://www.bic.mni.mcgill.ca/brainweb/

[12] http://www.cma.mgh.harvard.edu/isbr

[13] T. Tanimoto, An elementary mathematical theory of classification and prediction. Technical report, IBM Corp., 1958.

[14] 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 Trans. Med. Imag.,* vol. 13, no. 4, pp. 716-724, Dec. 1994

[15] W. C. Liew, H. Yan, "An Adaptive Fuzzy Clustering Algorithm for 3-D MR Image Segmentation", *IEEE Trans. Med. Imag*., vol. 22, no. 9, pp. 1063-1075, Sept. 2003.