CHAOS Metrics and Evaluation Method

Last update: September 2019

Abstract

This document explains the evaluation metrics for CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation challenge, which held at The IEEE International Symposium on Biomedical Imaging (ISBI) in 2019 [1]. In CHAOS, evaluation is handled by four different metrics. Final score of a submission is calculated by aggregation of the metrics via scoring system. The reason for using multiple metrics, definitions of the metrics and scoring system are briefly defined. For more information, you may refer to https://chaos.grand-challenge.org/Evaluation/ page.

1 Metrics

The primary aim of medical image segmentation is to develop tools for clinical needs such as diagnosing, surgery planning, and organ transplant operations. Hence, tolerance of error is minimal. According to previous studies [2, 3], there is no single metric that evaluates 3D segmented data completely and fairly in terms of clinically acceptable results. Since the results have to be analysed from many perspectives, the aggregation of multiple evaluation metrics was preferred [4]. Spatial overlap-based, volume-based, and spatial distance-based metrics were chosen here to analyse different aspects of segmentation in terms of different aspects. Also a mini experiment can be found at evaluate2D_metric_compare.m in GitHub repository to observe response of different metrics to different segmentation errors.

1.1 Dice coefficient (DICE)

DICE is one of the most frequently used metrics in medical image segmentation. It was introduced by Lee R. Dice for ecological studies [5]. The metric quantifies the match between two sets by the average size of their intersection.

$$DICE = \frac{2\,|V_{Ref} \cap V_{Seg}|}{|V_{Ref}| + |V_{Seg}|}, \tag{1}$$

where V_{Ref} is the set of voxels identified as foreground in the reference volume (ground truth) and V_{Seg} is the set of voxels identified as foreground in the segmented volume. As always, cardinality is denoted by $|.|$ and is the total number of elements in a set.

The value of DICE is between 0 and 1. A perfect segmentation yields value 1 while a segmentation with no intersection of the foregrounds of the reference and the segmented image yields value 0.

1.2 Relative absolute volume difference (RAVD)

Relative absolute volume difference is based on volumetric similarity of two 3D objects. RAVD has been used in many studies for evaluating segmentation [6, 7]. The absolute difference between the cardinalities of V_{seg} and V_{Ref} is divided by the cardinality of V_{Ref}. A perfect segmentation gets value 0. The maximum value of RAVD (assuming that V_{ref} is not empty) is $k_1 \times k_2 \times k_3 - 1$, where k_1, k_2, and k_3 are the three dimensions of the volume in image coordinates. Typically, the value is returned as a percentage:

$$RAVD = \frac{|\;|V_{Seg}| - |V_{Ref}|\;|}{|V_{Ref}|} \times 100. \tag{2}$$

1.3 Average symmetric surface distance (ASSD)

Symmetric surface distance (SSD) measures the distance between the segmented foregrounds of two 3D volumes, say, A and B. To calculate SSD, the border voxels of the two segmentations are determined. Border voxes are those marked as forground, whose 27-neighbourhood ($3 \times 3 \times 3$) contains at least one background voxel. The relative voxel positions in the image are transformed to real world coordinates (millimeters) by using parameters in DICOM tags. For each border voxel from volume A, the closest border voxel in volume B is determined, and the distance between the two is calculated. The same is applied to volume B. The distances for both A and B are stored. After that, the average of all these distances is calculated, hence the name Average Symmetric Surface Distance (ASSD).

Define the distance of voxel x to a set of voxels A as

$$d(x, A) = \min_{y \in A} d(x, y), \tag{3}$$

where $d(x, y)$ is the Euclidean distance of between voxels x and y in real-world cooridnates? The average of all stored distances gives the average symmetric surface distance (ASSD):

$$ASSD = \frac{1}{|B_{ref}| + |B_{seg}|} \times \left(\sum_{x \in B_{seg}} d(x, B_{ref}) + \sum_{y \in B_{ref}} d(y, B_{seg}) \right), \tag{4}$$

where B_{ref} and B_{seg} are the border voxel sets of the reference and the segmented volume, respectively.

For a perfect segmentation ASSD achieves value value is 0 mm. There is no reasonable upper limit for the worst segmentation beside the largest diagonal of the cuboid containing the 3D image calculated as $\Delta = \sqrt{d_1^2 + d_2^2 + d_3^2}$, where d_i is the ith dimension of the volume in mm.

1.4 Maximum symmetric surface distance (MSSD)

Maximum symmetric surface distance (MSSD) (or Hausdorff distance) is calculated as the maximum of all symmetric voxel distances $d(x, B_{ref})$ and $d(y, B_{seg})$:

$$MSSD = \max \left(\max_{x \in B_{seg}} (d(x, B_{ref})), \max_{y \in B_{ref}} (d(y, B_{seg})) \right). \tag{5}$$

The value of MSSD is 0 mm for a perfect segmentation. As with ASSD, the upper limit exists (Δ) but is impractical to use.

2 Rescaling and aggregating of metrics

The aggregated measure detailed below was used for the CHAOS challenge. Since each of the four metrics has a different scale and expected distribution of its values, we chose a heuristic threshold approach for making the values of these measures more commensurable. The thresholds were determined by examining intra- and inter-user similarity in creating the ground truth reference for our data. The same image volumes were manually segmented multiple times by the same expert and by different experts. Pairwise metrics of these different manual segmentations were calculated. The differences in the values of the four metrics were used as a guide to determine the respective thresholds. The thresholds are detailed below:

- A DICE value lower than 0.8 gets a score of 0. Values higher than 0.8 are multiplied by 100. Denote by s the raw DICE value. Then the score which we take forward is

$$S_{DICE} = \begin{cases} 0, & s < 0.8 \\ 100s & s \geq 0.8 \end{cases} \tag{6}$$

- RAVD values higher than 5% get a score of of 0. RAVD values between 5% and 0% are mapped to the range of $[0, 100]$. Since lower RAVD indicates better performance, the mapping from actual value to score has an inverse proportion. Again, let s be the raw RAVD value. Then

$$S_{RAVD} = \begin{cases} 0, & s > 5\% \\ 100 - 20s & s \leq 5\% \end{cases} \tag{7}$$

- ASSD values greater than 15mm get a score of 0. ASSD values lower than 15mm are mapped to $[0, 100]$ so that the value 0 correspond to 100% match of the two segmentations. Denoting the raw ASSD measure by s, we have

$$S_{ASSD} = \begin{cases} 0, & s > 15mm \\ 100 - \frac{20}{3}s & s \leq 15mm \end{cases} \tag{8}$$

- The same scaling as above is applied to the MSSD values with threshold 60mm. Denoting the raw ASSD measure by s, we have

$$S_{ASSD} = \begin{cases} 0, & s > 60mm \\ 100 - \frac{5}{3}s & s \leq 60mm \end{cases} \tag{9}$$

A summary of the metrics and their thresholds is presented in Table 1.

Table 1: Summary of metrics and threshold values.

Metric name	Best value	Worst value	Threshold
DICE	1	0	DICE >0.8
RAVD	0%	100%	RAVD <5%
ASSD	0 mm	Δ	ASSD <15 mm
MSSD	0 mm	Δ	MSSD <60 mm

Finally, since the measures are now all scaled within 0–100%, where higher values are desirable, we propose to average the four scores in order to give one final score for the segmentation.

References

[1] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. Gezer. (2019) CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge - Grand Challenge. Accessed: 2019-09-12. [Online]. Available: https://chaos.grand-challenge.org/

[2] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz *et al.*, "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature Communications*, vol. 9, no. 1, p. 5217, dec 2018. [Online]. Available: http://www.nature.com/articles/s41467-018-07619-7

[3] V. Yeghiazaryan, I. Voiculescu, V. Yeghiazaryan, and I. Voiculescu, "An Overview of Current Evaluation Methods Used in Medical Image Segmentation," Oxford, UK, Department of Computer Science, Tech. Rep., 2015. [Online]. Available: https://www.cs.ox.ac.uk/publications/publication10110-abstract.html

[4] A. N. Langville and C. D. C. D. Meyer, *Who's #1? : the science of rating and ranking.* Princeton University Press, 2013, p. 247. [Online]. Available: https://dl.acm.org/citation.cfm?id=2613650

[5] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, jul 1945. [Online]. Available: http://doi.wiley.com/10.2307/1932409

[6] T. Heimann, B. van Ginneken, and M. Styner, "Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1251–1265, aug 2009. [Online]. Available: http://ieeexplore.ieee.org/document/4781564/

[7] A. Salimi, M. A. Pourmina, and M. S. Moin, "Fully automatic prostate segmentation in MR images using a new hybrid active contour-based approach," *Signal, Image and Video Processing*, vol. 12, no. 8, pp. 1629–1637, nov 2018. [Online]. Available: http://link.springer.com/10.1007/s11760-018-1320-y