Boundary Overlap for Medical Image Segmentation Evaluation

Varduhi Yeghiazaryan and Irina Voiculescu
Department of Computer Science, University of Oxford

Abstract

All medical image segmentation algorithms need to be validated and compared, and yet no evaluation framework is widely accepted within the imaging community. Collections of segmentation results often need to be compared and ranked by their effectiveness. Evaluation measures which are popular in the literature are based on region overlap or boundary distance. None of these are consistent in the way they rank segmentation results: they tend to be sensitive to one or another type of segmentation error (size, location, shape) but no single measure covers all error types. We introduce a new family of measures, with hybrid characteristics. These measures quantify similarity/difference of segmented regions by considering their overlap around the region boundaries. This family is more sensitive than other measures in the literature to combinations of segmentation error types. We compare measure performance on collections of segmentation results sourced from carefully compiled 2D synthetic data, and also on 3D medical image volumes. We show that our new measure (1) penalises errors successfully, especially those around region boundaries; (2) gives a low similarity score when existing measures disagree, thus avoiding overly inflated scores; and (3) scores segmentation results over a wider range of values. We consider a representative measure from this family and the effect of its only free parameter on error sensitivity, typical value range, and running time.

Introduction

The most popular empirical discrepancy methods (as per Zhang's classification [7]) compare segmented images to (manually produced) ground truth by exploring similarity/difference of the labelled regions. We consider three common classes of such measures — overlap, size, and boundary distance based — and examine popular members of these classes. These measures are only sensitive to limited, mostly separate, ranges of segmentation errors; they produce contradicting scores and rankings. [2] suggests a framework for combining results by several measures to produce a single score. Nevertheless, this remains an open question: more recent reports comparing segmentation algorithms avoid this by using one measure as main indicator of segmentation quality [3] or by averaging rankings of algorithms obtained with different measures [1].

We construct a dataset of 43 synthetic images to simulate segmentation results with varied deviations from a gold standard. We quantify these images with a range of evaluation measures and reveal how well these react to specific segmentation mistakes. Additionally, we compare the performance of the same measures on real 3D data using automatically segmented medical images [6]. Based on the results illustrated, in order to increase insight into segmentation algorithm performance, we propose that our new measure from the *boundary overlap family* be considered alongside (or instead of) some existing measures.

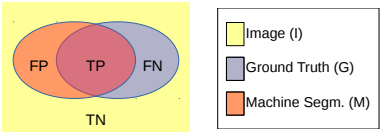
Novel Methods based on Boundary Overlap

The lack of unique established evaluation framework in medical image segmentation results in various evaluation methods being used in literature. Although new evaluation methods have been suggested, measures with simple definitions are still the most popular methods in use in the segmentation literature. We consider several popular measures based on our own literature survey [5], definitions are presented in Table 1.

We introduce a new class of evaluation measures to quantify similarity/difference of segmented regions by considering their overlap around the region boundaries. For this we borrow features from both overlap based measures and boundary distance based measures. In order to compare regions A and B, we consider all the points on the boundaries $\partial A, \partial B$. However, instead of considering distances of points on ∂A to ∂B and vice versa, as in boundary distance based measures, we look at the local neighbourhoods of those boundary points and determine the overlap of the two regions in those neighbourhoods.

Here we formally define one boundary overlap method and consider it as a representative of the class. Let N_x be the local neighbourhood of some radius r of a point x: $N_x = \{y, d(x,y) \leq r\}$, where $d(x,y)$ is the distance of the points. Let $A(N_x)$ and $B(N_x)$ be the portions of regions A and B in that neighbourhood N_x. Let us denote the Dice similarity coefficient between these portions by $DSC(N_x) = \frac{2|A(N_x) \cap B(N_x)|}{|A(N_x)|+|B(N_x)|}$. Now we can define *directed boundary Dice (DBD)* as the average of these overlap scores among all points on the first boundary:

$$DBD(A,B) = \frac{\sum\limits_{x \in \partial A} \frac{2|A(N_x) \cap B(N_x)|}{|A(N_x)|+|B(N_x)|}}{|\partial A|}$$



Similarity Measure	**Difference Measure**
DSC = $\frac{2 \times \text{TP}}{\text{FP} + \text{TP} + \text{TP} + \text{FN}} = \frac{2\,\lvert M \cap G \rvert}{\lvert M \rvert + \lvert G \rvert}$	SVD = 1 - DSC
JSC = $\frac{\text{DSC}}{2 - \text{DSC}} = \frac{\text{TP}}{\text{FP} + \text{TP} + \text{FN}} = \frac{\lvert M \cap G \rvert}{\lvert M \cup G \rvert}$	VOE = 1 - JSC
TPVF (Rec) = $\frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\lvert M \cap G \rvert}{\lvert G \rvert}$	FNVF = 1 - TPVF = $\frac{\text{FN}}{\text{TP} + \text{FN}} = \frac{\lvert G \setminus M \rvert}{\lvert G \rvert}$
TNVF = $\frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\lvert I \rvert - \lvert M \cup G \rvert}{\lvert I \rvert - \lvert G \rvert}$	FPVF = 1 - TNVF = $\frac{\text{FP}}{\text{TN} + \text{FP}} = \frac{\lvert M \setminus G \rvert}{\lvert I \rvert - \lvert G \rvert}$
Prec = $\frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\lvert M \cap G \rvert}{\lvert M \rvert}$	
	RVD = $\left\lvert \frac{\text{FP} - \text{FN}}{\text{TP} + \text{FN}} \right\rvert = \left\lvert \frac{\lvert M \rvert - \lvert G \rvert}{\lvert G \rvert} \right\rvert$
	HD = $\max \left\{ \max_{x \in \partial G} d(x, \partial M), \max_{y \in \partial M} d(y, \partial G) \right\}$
	ASSD = $\frac{\sum_{x \in \partial G} d(x, \partial M) + \sum_{y \in \partial M} d(y, \partial G)}{\lvert \partial G \rvert + \lvert \partial M \rvert}$
Proposed boundary Dice measures	
DBD1 = DBD(G, M) = $\frac{\sum_{x \in \partial G} DSC(N_x)}{\lvert \partial G \rvert}$	
DBD2 = DBD(M, G) = $\frac{\sum_{y \in \partial M} DSC(N_y)}{\lvert \partial M \rvert}$	
SBD = $\frac{\sum_{x \in \partial G} DSC(N_x) + \sum_{y \in \partial M} DSC(N_y)}{\lvert \partial G \rvert + \lvert \partial M \rvert}$	

Table 1: (1) Overlap based: Dice similarity coefficient (DSC) and symmetric volume difference (SVD); Jaccard similarity coefficient (JSC) and volumetric overlap error (VOE); true positive (TPVF), true negative (TNVF), false positive (FPVF) and false negative (FNVF) volume fractions; precision and recall. (2) Size based: relative volume difference (RVD). (3) Boundary distance based: Hausdorff distance (HD) and average symmetric surface distance (ASSD). (4) Our new measure is SBD.

Similarly, the symmetric version of this measure can be defined — *symmetric boundary Dice (SBD)*:

$$SBD(A, B) = \frac{\sum_{x \in \partial A} DSC(N_x) + \sum_{y \in \partial B} DSC(N_y)}{\lvert \partial A \rvert + \lvert \partial B \rvert}$$

According to these definitions, both our measures have a fixed range of values between 0 and 1 with higher values indicating better match between the regions. We can change the separate components in the definitions above, thus getting other measures in the same family. These new measures can be defined, for instance by replacing the boundary averaging by other boundary distance features, such as taking the maximum. Similarly, the overlap measure in the neighbourhood can be replaced by others. Corresponding difference measures can also be introduced, such as $1 - \text{SBD}$.

The popular similarity/difference measures fail to capture all the aspects of the segmented regions. Boundary overlap measures introduce a family of new methods with hybrid features. We show how these display increased sensitivity to a wider range of errors than the classes of measures considered.

Parameterisation. Each boundary overlap measure can be parameterised by the neighbourhood radius to increase/decrease visibility of local overlap. In our current experiments we consider cube–shaped Moore neighbourhoods of radius 1, i.e. 9– and 27–neighbourhoods in 2D and 3D respectively, to radius 5 and compare the measure performance with these experimentally. This systematic study will be illustrated in the full paper.

Complexity. In terms of complexity boundary overlap measures allow for straightforward implementation, which is linear in region size. From this perspective they are more similar to overlap measures, and outperform boundary distance measures which have to consider all pairs of points from the two boundaries (although nearly linear implementations exist for these [4]).

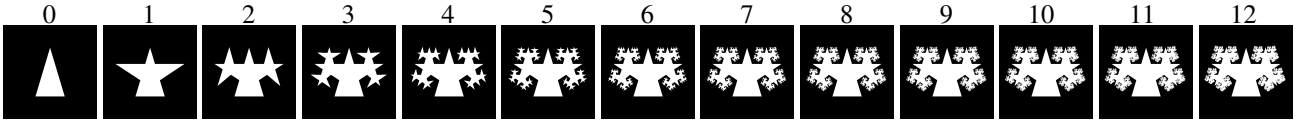


Figure 1: Sequence of synthetic fractal images with new isosceles triangles added at each recursive step (0–12)

Experimental Results

Synthetic data. We compare SBD against existing measures on a carefully constructed set of 2D synthetic images. It contains instances of simple geometric shapes (rectangle, ellipse, star) and fractals. For each shape there is an image to be used as ground truth and a sequence of images to simulate segmentation results. These features depart from the ground truth in size, location, orientation, 'salt–and–pepper' errors, or combinations thereof.

Our experiments for simple shapes will be illustrated in the full paper. They rely on the comparison of the 'segmented' images against the corresponding 'ground truth' using the measures listed above. For a given sequence of images, SBD assigns different scores to all the cases whereas most existing measures are unable to differentiate clearly between segmentations. Our results show that SBD generally penalises segmentation errors more than the other measures, resulting in a wider (and hence easier to use) range of scores. For images with 'salt–and–pepper' errors, where other measures mostly disagree, SBD averages the overoptimistic scores of overlap based measures and the overpessimistic scores of boundary distance based measures. SBD penalises the random false negatives inside and false positives outside the expected shapes more than the overlap measures and less than the boundary distance measures.

The evaluation scores of existing measures heavily disagree with each other on images simulating potential errors: some measures are sensitive to specific types of errors while other measures are not, and vice versa for other images. SBD assigns a low score to all of such images thus showing better sensitivity to a larger span of errors. The directional variants of boundary Dice show interestingly contradicting results for images with evident disagreement from other measures; these contradictions are neutralised in its symmetric variant (SBD).

Consistently changing the Moore neighbourhood radius from 1 to 5 for the synthetic data we reveal that SBD similarity scores gradually increase for larger radii (for most images). The growth is around 9–10% between radius 1 to 5 in average, although for specific segmentation results it reaches 46–47% (more details in full paper).

Additional to the simple shapes, the synthetic data contains a sequence of fractals (Figure 1). These help reveal the measures' sensitivity to boundary errors, where the boundary consists of increasingly intricate detail. The initial region is an isosceles triangle. At each step of the recursion new triangles with the same ratio of side lengths $(4 : 7)$ are added on the equal sides of each triangle. The image from step 6 of the recursion is assigned as ground truth, producing the results reported in Figure 2.

The pair of graphs reveal interesting patterns. Images 0–5 are undersegmented, and images 7–12 are oversegmented. Note, on the one hand, that the conventional measures DSC (and JSC), TPVF, HD, ASSD show more variability for the undersegmented images; they fail to differentiate properly the oversegmented images. On the other hand, TNVF, Prec are insensitive to undersegmentation errors. The directional variants of the new measure show similar behaviour to the first and the second group of measures respectively. RVD performs relatively better but still fails to capture differences at the level of local neighbourhoods on the region boundary.

By contrast, SBD easily identifies the under– and oversegmentation errors and reflects those in its scoring. It shows high sensitivity to even small differences between the boundaries of the ground truth and 'segmented' images and uses a wide scoring range to illustrate those differences. These examples showcase how well SBD outperforms all the other measures, especially for segmentation errors near the boundary of the intended region.

Real data. We ran our own segmentation algorithm described in [6] on 9 different CT scan volumes to label 13 kidneys (left or right), for which we also produced gold standard masks manually. The results of validation of the segmented 3D volumes against the ground truths are presented in Figure 2.

Since the machine segmentation results are produced by a single automated algorithm, errors of a similar nature are expected in all cases. The small ranges of score values produced by the various measures illustrate this. As was the case with 2D synthetic images, SBD tends to penalise errors more heavily, producing smaller similarity scores — and hence higher difference scores — than most existing measures. Furthermore, in case of strong disagreement between classes of measures or several measures inside a class, like in volumes 10–11, it acknowledges the errors without being overly sensitive to the error.

In order to preserve the level of detail in the graphs, Figure 2 does not include data points for two kidney volumes with pathological cases where the gold standard disagreed strongly with the machine segmentation. For these images the machine segmented 3D region was more than 3.5 times larger than the ground truth (the gold standard did not include necrotic tissue). SBD scored these volumes around 0.33 and 0.38, sharing the sensitivity of other measures: DSC of 0.28 and 0.34, and HD of 97 and 85mm. RVD flags this stark difference.

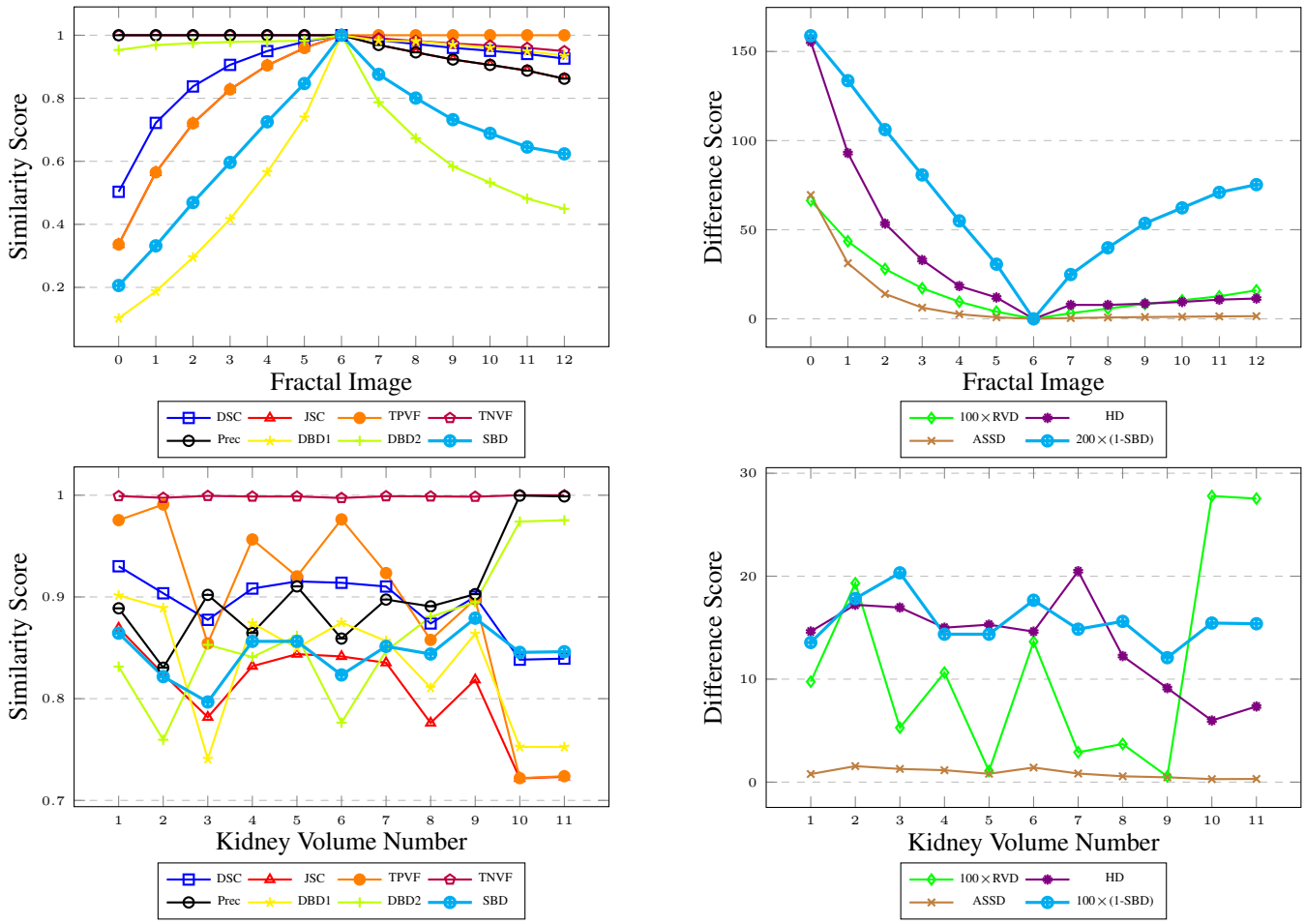


Figure 2: Results for synthetic fractal images and medical kidney volumes. Fractal at step 6 of the recursion considered as ground truth. Results for 11 kidney volumes (left or right) from 9 3D medical CT scans.

Conclusions

Both the results illustrated here and our other experiments (which will be shown in the full paper) demonstrate how, by contrast with the existing measures considered separately, SBD reacts to a wider range of error types. It penalises errors more, producing a wider spread of similarity scores (lower for more errors). Cases of strong disagreement between existing measures received acknowledgement of errors from SBD but avoiding overly inflated scores. Being easy to implement and quick to run, SBD can be used to evaluate segmentation algorithms alongside existing evaluation measures to increase insight into the algorithm performance. Future work will include other measures from this class (such as symmetric boundary TPVF). In the pipeline are experiments with a wider range of segmentation results from medical scan data, and further analysis of the strengths and weaknesses of this new family in comparison to measures recently suggested in the literature.

References

[1] ISLES challenge 2015. `www.isles-challenge.org/ISLES2015/`.

[2] T Heimann et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans Med Imag*, 28(8):1251–1265, 2009.

[3] Bjoern H Menze, Andras Jakab, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Medical Imaging*, 34(10):1993–2024, 2015.

[4] Abdel A Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):1–28, 2015.

[5] V Yeghiazaryan and I Voiculescu. An overview of current evaluation methods used in medical image segm. Technical Report CS-RR-15-08, Dept Comp Sci, U Oxford, 2015.

[6] Varduhi Yeghiazaryan and Irina D Voiculescu. Automated 3D renal segmentation based on image partitioning. In *Proc. SPIE 9784, MI 2016: IP*, page 97842E. SPIE, 2016.

[7] Y J Zhang. A surv on eval meth for im segm. *Patt Rec*, 29(8):1335–1346, 1996.