New Disagreement Metrics Incorporating Spatial Detail – Applications to Lung Imaging

Alberto M Biancardi, Jim M Wild

Polaris (IICD Dept.) and INSIGNEO
The University of Sheffield, Sheffield, U.K
a.biancardi+miua@sheffield.ac.uk
j.m.wild+miua@sheffield.ac.uk

Abstract. Evaluation of medical image segmentation is increasingly important. While set-based agreement metrics are widespread, they assess the absolute overlap, but fail to account for any spatial information related to the differences or to the shapes being analyzed. In this paper, we propose a family of new metrics that can be tailored to deal with a broad class of assessment needs.

Keywords: Disagreement, Evaluation metrics, Segmentation, Medical images, Lung MRI

1 Introduction

As computer-supported segmentation of medical images becomes increasingly commonplace, evaluating the outcomes plays an even more important role – for instance, for validation purposes (especially on large datasets) or for performance comparison. The typical goal of these evaluations is the assessment of the agreement, or disagreement, expressed as a measure of their spatial overlap. The most common approaches, across several applicative domains [1-4], are based on the quantification of the spatial agreement by means of set operations (Dice similarity, Jaccard index, etc.). However, these approaches assume that the assessed elements are independent among themselves (as entailed by the definition of these similarity assessments itself, being based on set operations), while, in the imaging domain, the segmentation-region elements (pixels, voxels, etc.) are characterized by their spatial location and this location introduces a correlation among the set elements. In this paper, a new family of metrics that quantify various aspects of the spatial differences between two regions is presented. We demonstrate the use of the proposed metrics amongst set-based techniques in the analysis of lung MRI.

2 Set-Based Measurements

Following [5] and [6], let a scalar (medical) image be represented by a function defined on a regular grid $I: \mathcal{G} \rightarrow V$. Typically the elements of $\mathcal{G}$ are indexed by a subset

adfa, p. 1, 2011.

of $\mathbb{Z}^n$, where n is the image dimensionality, and V is a subset of $\mathbb{Z}$ or $\mathbb{Q}$. Additionally we assume that consistent spatial locations are assigned to all of the elements of $\mathcal{G}$ and that, therefore, a metric is defined between grid element pairs.

We define a binary image as an image with two possible values:

$$b: \mathcal{G} \to [0,1] \tag{1}$$

As we are going to analyze only single-label segmentations, a segmentation is represented by a binary image and the subset of $\mathcal{G}$ having a value of 1:

$$S = \{x: b(x) = 1, x \in \mathcal{G}\} \tag{2}$$

The function b, by definition, induces a partition of $\mathcal{G}$ by means of its two inverse images; therefore, the segmentation background, i.e. $b^{-1}(0)$, will be simply denoted by S', being the complement of S.

Given two segmentations T and R, typical set-based assessments of spatial overlap are defined by computing the cardinality of selected subsets of $\mathcal{G}$. If T is a test segmentation and R is a reference segmentation, for which its values have been considered to be in accordance with the expected outcome (or ground truth), then the customary confusion matrix can be expressed as:

	R	R'
T	$TP = T \cap R$	$FP = T \cap R'$
T'	$FN = T' \cap R$	$TN = T' \cap R'$

The performance parameters can then be expressed either in term of set operations, where there is no assumption of truth (see also section 4), or based on the confusion matrix cardinalities:

Measure	Set-based	Truth-based
$Dice$	$\frac{2\,\|T \cap R\|}{\|T\| + \|R\|}$	$\frac{2\,TP}{2TP + FN + FP}$
$Jaccard$	$\frac{\|T \cap R\|}{\|T \cup R\|}$	$\frac{TP}{TP + FN + FP}$
$Sensitivity$ or $Recall$	$\frac{\|T \cap R\|}{\|R\|}$	$\frac{TP}{TP + FN}$
$Specificity$	$\frac{\|T' \cap R'\|}{\|R'\|}$	$\frac{TN}{FP + TN}$

3 Spatial Impact of the Image Domain

Even if the set-based measurements are sometimes referred to as assessing the spatial overlap, the extent to which the actual spatial characteristics of the two segmentations under evaluation are assessed is limited: only the exact overlap of the voxels is tested, while any level of proximity is lost. Additionally every element is given the same weight, regardless of possible constrains brought forth by the specific application where the evaluation takes place.

Figure 1 shows the impact that the spatial location of the segmented voxel plays when assessing the dissimilarity between two segmentations, assuming that no a-priori knowledge is available as regards the region to be segmented. The reference region R, shown in (a), is assumed to be the ground truth while regions shown in (b) and (c) have 7 additional elements with respect to R (false positives highlighted in red). It is easy to see, that, no matter what cardinality-based measurement is chosen, the four regions will have the same outcome measure of agreement, regardless of the position of the red pixels:

	(b)	(c)
$Dice$	$\frac{2\ \lvert R\rvert}{\lvert R\rvert+7+\ \lvert R\rvert}=\frac{2\ \lvert R\rvert}{2\lvert R\rvert+7}=\frac{280}{287}=\mathbf{97.56}\%$	
$Jaccard$	$\frac{\lvert R\rvert}{\lvert R\rvert+7}=\frac{140}{147}=\mathbf{95.24}\%$	
$Sensitivity$	$\frac{\lvert R\rvert}{\lvert R\rvert}=\ \mathbf{100}\%$	
$Specificity$	$\frac{\lvert R'\rvert-7}{\lvert R'\rvert}=97.30\%$	

Analogously, regions (d) and (e) have 7 pixels missing from the reference region (a), false negatives highlighted in orange. The lack of any spatial insight in the evaluation produces 4 values, for these image examples, that are always the same:

	(d)	(e)
$Dice$	$\frac{2\ (\lvert R\rvert-7)}{\lvert R\rvert-7+\ \lvert R\rvert}=\frac{2\ (\lvert R\rvert-7)}{2\lvert R\rvert-7}=\frac{266}{273}=\mathbf{97.44}\%$	
$Jaccard$	$\frac{\lvert R\rvert-7}{\lvert R\rvert}=\frac{133}{140}=\mathbf{95.00}\%$	
$Sensitivity$	$\frac{\lvert R\rvert-7}{\lvert R\rvert}=\frac{133}{140}=\mathbf{95.00}\%$	
$Specificity$	$\frac{\lvert R'\rvert}{\lvert R'\rvert}=100\%$	

It is worth mentioning that, when dealing with medical images, the cardinalities of all the subsets that include segmentation complements (T', R') are somewhat arbitrary, being easily affected by crop operations that leave the segmented regions untouched.

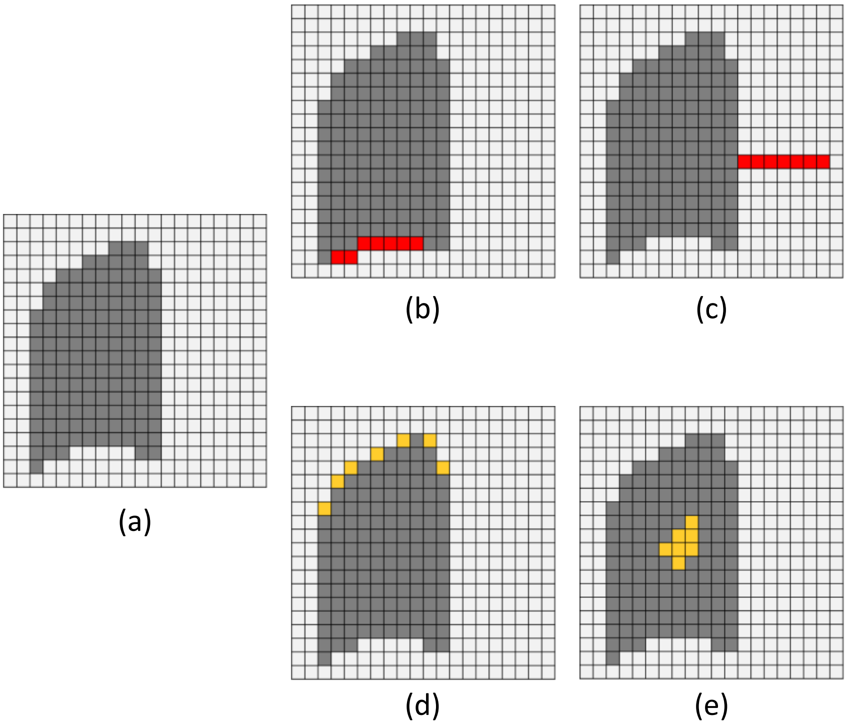

Fig. 1. Two groups of hypothetical segmentations having the same set-based agreement with the reference region (a). Regions (b) and (c) have additional pixels (i.e., false positives), marked in red/dark gray; regions (d) and (e) are missing pixels (i.e., false negatives), marked in orange/light gray

4 Roles

Before proceeding, it is important to highlight a key difference between the two possible scenarios where an agreement assessment takes place: (i) comparison with respect to a reference region (typically called ground truth), and (ii) comparison between two regions with equal standing. Sub-figures 1a and 1e demonstrate this difference. If 1a is the reference region and we want to assess the agreement of 1e with 1a, then clearly 1e is in error and, possibly, by an important one as the region has a hole in the middle of a supposedly filled region[1]. On the other hand, if 1a and 1e are equally reliable (we call it the symmetric case), then the hole in the middle might be part of the correct result, but we cannot infer it from the data that we have available. In the following sections, the discussion will assume the existence of a reference region; a preliminary analysis of the symmetric scenario is presented in section 8.

[1] Notice that a specular reasoning still holds if the roles of 1a and 1e are swapped - with 1e being the reference and 1a being the assessed region.

5 Initial Considerations for New Metrics

Based on the previous considerations, the key aspect we would like to introduce in our metrics is the acknowledgment of and a grading of the different spatial positions where the disagreements occur. For instance, we would like to switch from a cardinality-based disagreement as in Equation 3 (or, re-written to loop over all the image-domain elements, Equation 4) to a disagreement metric where the disagreement is weighted by a spatially dependent function w as in Equation 5 (and the normalization is scaled by function n)

$$dis(T,R) = \frac{|T \triangle R|}{|R|} \tag{3}$$

$$dis(T,R) = \frac{\sum |S_{ij} - R_{ij}|}{\sum R_{ij}} \tag{4}$$

$$new\_dis(T,R) = \frac{\sum w(i,j,R)\, |S_{ij} - R_{ij}|}{\sum n(i,j,R)} \tag{5}$$

As the new disagreement measures in (5) are still based on the cardinality of $\mathcal{G}$ subsets, they will clearly satisfy the conditions of a metric as long as the weighting functions are strictly positive.

6 A Family of Disagreement Metrics

A convenient family of metric-defining weighting functions can easily be built by using the signed Euclidian distance transform (SEDT) [6,7] of the reference region, where the internal elements are given a positive value, while those outside the region are given a negative value, as shown in Figure 2. It is worth highlighting how this approach, thanks to its use of the SEDT, provides several potential advantages (e.g., with respect to the perceptual-based approach of [8,9]):

- it maps the n-dimensional image domain to a single dimension;
- it overcomes the need to account for the pixel/voxel size;
- it makes it possible to structure the weighting function according to problem-specific (anatomical) sizes, expressed in real word lengths.

Therefore, the weighting function can be written as $w(i,j,R) = a\big(\, d_R(i,j,R)\big)$ where the metric-proxy $a\colon \mathbb{R} \ \ \rightarrow \mathbb{R}_0^+$ defines the amount of disagreement according to the signed distance from the reference region border(s) and provides ample freedom in expressing the wanted grading of the disagreement. The following equations exemplify this relationship between disagreement location and its measurement. The metric-proxy a_1 (Equation 6) provides an example of a grading that is proportional to the distance from the region borders; the metric-proxy a_2 (Equation 7) is designed to tolerate errors up to 10 mm from the border and then flag anything further up; the metric-proxy a_3 (Equation 8) highlights errors near the border and discounts the other discrepancies:

$$a_1(x) \;=\; |x| \tag{6}$$

$$a_2(x) \;= (x/10)^4 \tag{7}$$

$$a_3(x) \;= e^{-(x/10)^4} \tag{8}$$

Figure 2 shows the signed distance transform computed on the sample reference region in Figure 1a, and the resulting disagreement values for all the other regions in Table 1, according to different choices of function a (Equations 6,7,8) and having the corresponding normalization function n defined so that a complete disagreement with the reference region is graded at 100%[2]. All the results are computed assuming a pixel size of 3 mm by 3 mm. The set-based disagreement is always 5% for all sub-figures 1b to 1e.

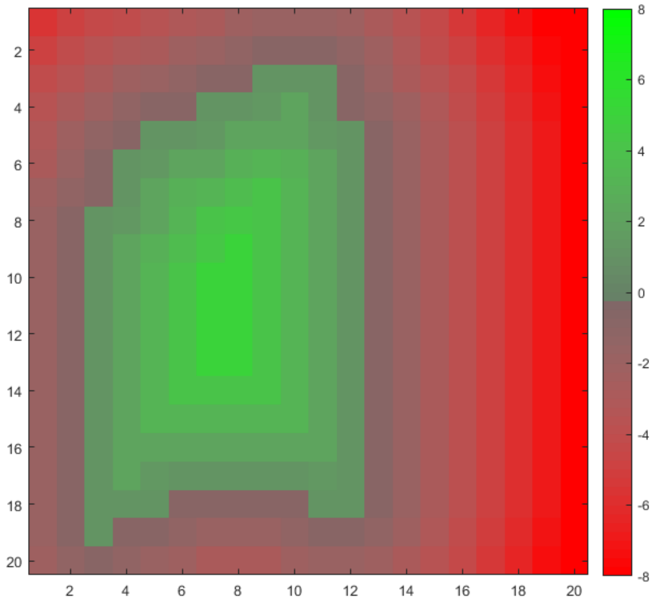

Fig. 2. The signed distance transform of the reference region R (shown in Figure 1a).

[2] The upper level of disagreement is arbitrary. The amount of 100% was chosen to be comparable with the set-based formulation.

Fig.1 Region	**Metric Proxies**		
	a_1	**a_2**	**a_3**
b	1%	0.01%	7%
c	7%	24%	3%
d	1%	0.01%	7%
e	7%	9%	3%

Table 1. Disagreement performance of the 4 hypothetical segmentations of Figure 1 with respect to the example metrics

7 Application to Lung Imaging

In order to demonstrate the effectiveness of the new metrics in supplying meaningful summaries of disagreements' spatial distributions, they were applied in the evaluation of different thresholding levels of a chest anatomical scan, acquired on a GE HDx 1.5T MR scanner[3] (3D spoiled gradient echo sequence, 1.5625 x 1.5625 x 5 mm^3 voxel size). Three threshold levels (th_3, th_5, th_7) were computed as the lowest values of a multi-threshold Otsu algorithm [10,11] with 3, 5, and 7 clusters, respectively. The image had previously been segmented manually to produce the reference segmentation. Representative coronal slices of the reference segmentation and the three thresholded regions are shown in Figure 3. As the number of clusters increases, the threshold values decrease, causing the resulting regions to exclude areas with denser tissue such as vessels and airway walls.

Table 2 reports the values for the set-based disagreement and for the metric-proxies a_2 and a_3 . While the threshold th_7 is too low according to any metric, the set-based disagreement is unable to summarize the slight differences between th_3 and th_5. By considering all the values from a_2 and a_3, it is straightforward to acknowledge that, if one is limited to simple thresholding, a tradeoff must be chosen – as th_3 has a better performance in the inner parts of the lungs, whereas th_5 is considerably better at capturing the lung borders.

Table 2. Disagreement values for the shapes in Figure 2a and 2b.

Threshold	**Set-based**	**Metric Proxies**	
		a_2	**a_3**
th_3	4.99%	0.05%	3.40%
th_5	5.11%	0.64%	3.29%
th_7	14.8%	2.49%	9.32%

[3] GE Healthcare, Milwaukee, IL, USA

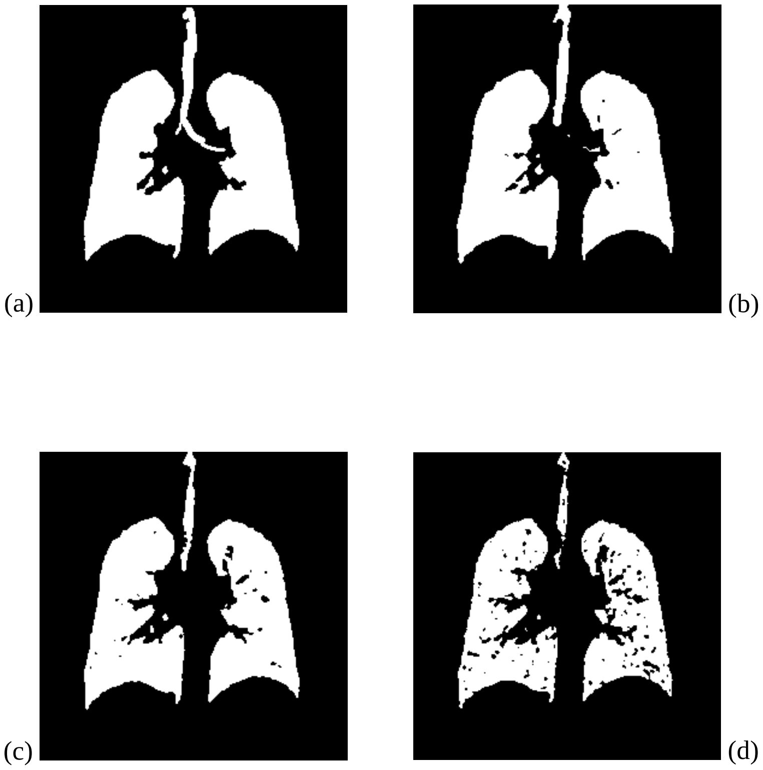

Fig. 3. Assessment of multiple thresholding levels in an MRI chest image. Representative slices of (a) the ground truth segmentation, (b,c,d) thresholded images at values th_3, th_5, th_7.

8 Region Assessment Without Reference

Let us now consider two regions, S and T, neither of which is the reference region (ground-truth). In these cases, the set-based overlap metrics of regions S and T can be interpreted as measuring the size of the region where there is agreement, the intersection, against an estimation of the reference region size – the size average for Dice, the union for Jaccard. Similar normalization approaches can be used as estimates for the set-based disagreement metric (3):

$$dis_{Dice} = \frac{2\,|S \triangle T|}{|S| + |T|} \tag{9}$$

$$dis_{Jaccard} = \frac{|S \vartriangle T|}{|S \cup T|} \tag{10}$$

If we are to take into account the spatial component, then a different estimate is required and the natural solution for the ground truth estimation appears to be a shape that is the average of S and T [12]. In our example, a shape interpolation [13] was performed using the itksnap tool [14,15]. With a proper estimate R of the reference region, a spatially aware disagreement can be expressed both as individual disagreements for each region (i.e., S with R and T with R) and as combined disagreement as the sum of the respective individual disagreements:

$$dis_S = new\_dis(S, R) \tag{11}$$

$$dis_T = new\_dis(T, R) \tag{12}$$

$$dis_{new} = dis_S + new_T \tag{13}$$

Clearly, there will be no disagreements deep inside there reference region because R is an average of the two regions. Nonetheless, the use of metric-proxies like a_3 can prove useful in summarizing the disagreement behavior around the region borders. This is well demonstrated by a simple example, where the disagreement between two shapes – a disk and a star, are investigated. Their set-based disagreements are 36.5% and 30.8% for dis_{Dice} and $dis_{Jaccard}$ respectively. Figure 4 shows the two shapes, the interpolated shape and the SEDT of the interpolated shape. Table 3 reports the values of the metric proxies. In the case when the reference ground truth is missing, metric proxies that limit the effect of large positive values, such as a_4 in Equation 14, can be used to assess the individual disagreements and, together with the other metrics, gain an additional understanding of the way spatial disagreement is distributed:

$$a_4(x) = 1 - \frac{1}{1 + e^{-x}} \tag{14}$$

Table 3. Disagreement values for the shapes in Figures 2a and 2b.

Shape	**Metric Proxies**		
	a_1	**a_2**	**a_3**
Disc	3.5%	9.4%	141.8%
Star	3.5%	9.4%	181.6%

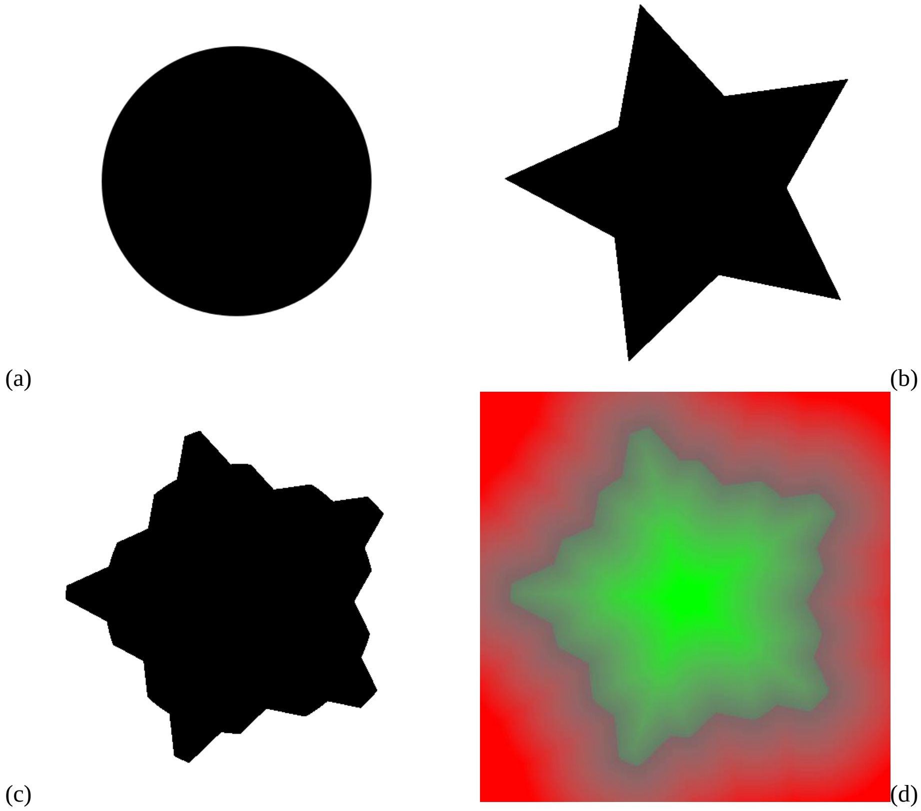

Fig. 4. Example of disagreement assessment without ground truth: (a) and (b) the shapes to be evaluated, (c) the interpolated shape as ground truth estimation, (d) the SEDT of c.

In our example, the a_4 disagreement for the star is 234.6% and for the disc is 490.8%. The higher number for the disk, in combination with a smaller a_3 value, tells us that the disk has much more disagreement on the outside of the reference region and that this disagreement is rarely far from the reference border.

9 Conclusions

Assessment of image regions plays an important role in computer-supported analysis because of the various outcomes relying on those regions for their computations. A new family of image disagreement metrics was introduced; these metrics can be easily adapted to the specific anatomical sizes under analysis and give a much richer summary of where the disagreement occurs when compared to set-based disagreement metrics. Preliminary applications show the potential usefulness of these additional spatial insights; however, further aspect can be investigated. Future work will study the relationship between these metrics and boundary/surface-based metrics such as the Hausdorff distance [16] and evaluate the extension to multi-object scenarios.

Acknowledgements

This work was supported by NIHR grant NIHR-RP-R3-12-027 and MRC grant MR/M008894/1.

The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health

References

1. Cuingnet R, Prevost R, Lesage D, Cohen LD, Mory B, Ardon R. (2012) Automatic detection and segmentation of kidneys in 3D CT images using random forests. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Heidelberg, pp. 66-74.
2. Van Leemput K, Bakkour A, Benner T, Wiggins G, Wald LL, Augustinack J, Dickerson BC, Golland P, Fischl B. (2009) Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus 19(6):549-57.
3. Sun S, Bauer C, Beichel R. (2012) Automated 3-D segmentation of lungs with lung cancer in CT data using a novel robust active shape model approach. IEEE Trans Med Imaging 31(2):449-60.
4. Korez R, Ibragimov B, Likar B, Pernuš F, Vrtovec T. (2015) A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation. IEEE Trans Med Imaging 34(8):1649-62.
5. Johnson HJ, McCormick MM, Ibáñez, and the Insight Software Consortium (2017) The ITK Software Guide http://itk.org/ItkSoftwareGuide.pdf
6. Felzenszwalb PF, Huttenlocher DP (2012) Distance Transforms of Sampled Functions. Theory of Computing 8:415-428.
7. Maurer CR, Qi R, Raghavan V. (2003) A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. IEEE Transactions on Pattern Analysis and Machine Intelligence. 25(2):265-70.
8. Villegas P, Marichal X. (2004) Perceptually-weighted evaluation criteria for segmentation masks in video sequences. IEEE Transactions on Image Processing. 13(8):1092-103.
9. Gavet Y, Fernandes M, Debayle J, Pinoli JC (2014) Dissimilarity criteria and their comparison for quantitative evaluation of image segmentation. Application to human retina vessels. Mach. Vision and Appl. 25(8):1953-1966.
10. Otsu N (1979) A Threshold Selection Method from Gray-Level Histograms. IEEE Trans on Systems, Man, and Cybernetics 9(1):62-66.
11. Liao PS, Chen TS, Chung PC (2001) A Fast Algorithm for Multilevel Thresholding. J. Inf. Sci. Eng. 17 (5): 713–727.
12. Raya SP, Udupa JK (1990) Shape-based interpolation of multidimensional objects. IEEE Trans. Med. Imaging 9: 32–42.

13. Albu AB, Beugeling T, Laurendeau D (2008) A morphology-based approach for interslice interpolation of anatomical slices from volumetric images. IEEE Transactions on Biomed Eng 55(8):2022-38.
14. Zukic D, Vicory J, McCormick M, Wisse LE, Gerig G, Yushkevich P, Aylward S. (2016) nD morphological contour interpolation. The Insight Journal http://hdl.handle.net/10380/3563
15. Yushkevich PA, Piven J, Cody Hazlett H, Gimpel Smith R, Ho S, Gee JC, Gerig G (2006) User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 31(3):1116-28.
16. Taha AA, Hanbury A. (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC medical imaging. 15(1):29.