Department of Computer Science

An Overview of Current Evaluation Methods Used in Medical Image Segmentation

Varduhi Yeghiazaryan and Irina Voiculescu

CS-RR-15-08

Department of Computer Science, University of Oxford
Wolfson Building, Parks Road, Oxford, OX1 3QD

An Overview of Current Evaluation Methods Used in Medical Image Segmentation

Varduhi Yeghiazaryan and Irina Voiculescu

November 3, 2015

Abstract

An important aspect of the development of image segmentation algorithms is the availability of mechanisms to evaluate them. This is necessary in order to estimate how fit a segmentation approach is for the specific task, validate its performance on data and compare it against other approaches.

Medical image segmentation is very interesting in this perspective since there is no established evaluation framework. A commonly accepted approach is to compare the segmentation output to some reference results (usually produced with manual segmentation) with a similarity/difference measure. While most authors rely on such methods there is still no agreement concerning the measures used. This is partly because there are no measures that reflect all the important features of a desirable segmentation and the existing measures do not discriminate different segmentation results in an acceptable way.

This paper provides a survey of current methods being used for medical image segmentation evaluation.

1 Introduction

The evaluation of medical image segmentation algorithms is impacted by a number of factors and conventions specific to this task. The variability of medical image data quality is one of them and manifests itself in noise level, artefacts, partial volume effects, limited resolution. Poor quality images are not only difficult to segment with automated algorithms but also reference labelling images are not always possible to construct from such data in an unambiguous way.

1.1 Contrast-enhanced images

Often medical contrast substances are administered prior to image acquisition to enhance the image quality. Depending on the imaging modality and timing of the acquisition this may cause major organs to have a distinct

appearance in the images and the segmentation algorithms to perform differently on contrast-enhanced images. For instance, minor blood vessels in a liver which form a part of the organ can be made to appear highlighted in contrast-enhanced images. Whether and how the inclusion or exclusion of these from the identified liver region should be penalised is an issue for the segmentation evaluation to address. However, current evaluation methods being used do not tackle it in any way.

1.2 Pathological formations

Another issue for the tasks of medical image segmentation and its evaluation is the presence of pathological tissue in the image. These include cysts, tumours and other lesions and change the anatomical structures in the image by impacting the intensities, size, spatial extent, shape and other features of organs. They impose the greatest challenge on medical image segmentation algorithms with most approaches showing worst results on images with pathological organs (see Campadelli et al. [4], Section 8: proposed method obtains a lower score on image set with cancerous livers).

An important problem over which there still seems to be no consensus is whether, in which cases, and to what extent pathological formations should be labelled as part of the considered regions representing anatomical structures. For instance, the MICCAI 2007 liver segmentation grand challenge [21] (see Section 3 for more details) assumed tumours, cysts should be segmented as part of liver; as a result the segmentation approaches excluding the lesion region performed poorly and received lower scores. Other literature suggesting a similar approach includes [11, 14], while some papers take the opposite perspective [22].

1.3 Intra- and inter-observer variability

A detailed discussion of requirements and format of image segmentation evaluation frameworks can be found in the paper by Udupa et al. [34]. In practice, most image segmentation algorithms applied to medical image data are evaluated by computing region similarity/difference measures for machine labelled regions and results of manual segmentation of the same images by trained operators. Since absolutely accurate reference labellings are unavailable for this type of data, expert radiologists produce labellings of the data which are used as ground truth. Such data provides high level of reliability but is still affected by intra- and inter-observer variability (see [2] for reported results).

A variety of similarity/difference measures have been suggested and used in the literature for the task of medical image segmentation evaluation for the past couple of decades. Today, authors continue to assess the performance of their segmentation approaches with different measures on data collections

of variable quality and size. In this paper, we consider the most popular measures in use for assessing final segmentation results in recent literature on the topic. Measures used specifically for intra- and inter-observer variability estimation or evaluation of initial results (recognition, bounding boxes, etc.) are not observed here. We describe the measures considered, present a discussion of their features and conduct some experiments to compare and analyse their performance in recognising reliable segmentation results and discriminating omnifarious segmentation errors.

2 Measure Type Classification

During the past couple of decades several authors conducted surveys on existing image segmentation evaluation methods and introduced classifications and comparisons of those. In a series of papers [42, 43, 44] Zhang discusses an extensive set of evaluation methods, categorises them based on the approaches used to assess image segmentation, compares these categories and reports an empirical comparison of specific evaluation methods. Evaluation methods are grouped into *analytical methods* which assess segmentation algorithms by considering their principles, *empirical goodness methods* analysing properties of segmented images, and *empirical discrepancy methods* which compare segmented images to *reference images* (gold standard, ground truth).

The empirical methods are further categorised by the quality measures they use to assess the segmentation algorithm performance for output images. The measure groups for empirical goodness methods include intra-region uniformity, inter-region contrast, region shape, moderate number of regions. The empirical discrepancy methods employ measures of number of mis-segmented voxels, position of mis-segmented voxels, number of objects in the image, feature values of segmented objects, miscellaneous quantities (region consistency, grey level difference, symmetric divergence) [44]. Also, a group of evaluation methods with measures, like amount of editing operations, visual inspection or correlation between original image and bi-level image are considered special cases, not fitting in any of the groups.

The reviewed methods are compared using criteria of generality, subjective/objective, qualitative/quantitative, complexity, consideration of segmentation applications and requirement for reference images. Additionally, the performance of several empirical (goodness and discrepancy) methods is compared on a series of simple synthetic images segmented with thresholding [42] with discrepancy methods showing better sensitivity to threshold changes.

Zhang, Fritts and Goldman [41] group image segmentation evaluation methods into five categories: *subjective evaluation*, when output image is assessed by human judges, *system level evaluation*, which analyses the seg-

mentation algorithm by considering results of systems using the segmentation, *analytical methods* (similar to [42]), *supervised methods* (corresponding to empirical discrepancy) and *unsupervised methods* (empirical goodness). They present a detailed discussion of a range of *unsupervised* evaluation methods, analysing the various measures used and their combination methods. A number of experiments are performed on synthetic and real images for representative evaluation methods to compare two segmentation results. The accuracy of these methods is reported as the percentage of cases when the segmentation result preferred by the evaluation method and human judges coincide. It is concluded from the findings that although the unsupervised evaluation methods perform well at comparing different parameterisations of a single segmentation algorithm, they are not successful in cases of machine segmentation versus different machine segmentation, or machine segmentation versus human segmentation comparisons.

The authors suggest using machine learning to combine simple unsupervised evaluation methods [40]. This approach outperforms all the previous methods on the suggested experiments showing promising results.

3 Evaluation Frameworks

A special algorithm for validating image segmentation, *simultaneous truth and performance level estimation (STAPLE)*, is introduced by Warfield, Zou and Wells [36]. This is an expectation-maximization approach for comparison of several (human or machine) segmentations. It gives a probabilistic estimate of the true segmentation and performance scores for the considered segmentation approaches. However, this evaluation method is designed to analyse different segmentation algorithms simultaneously and is not readily applicable to evaluating a single segmentation approach (unless several reference segmentations are available).

Crum et al. [13] redefine established overlap measures (Jaccard and DSC, see Section 6 for definitions) for fuzzy segmentation results and generalise these measures to consider multiple segmentation labels and images. Thus, a single total fuzzy overlap result can be generated for a series of images fuzzily segmented using several labels with two different segmentation algorithms. A corresponding distance measure is also introduced. Such generalisations turn out to be useful when an overall performance measure is sought across different labels and/or images. However, they still suffer from the shortcomings of the basic measures being redefined.

As part of MICCAI 2007 conference a competition and workshop called "3D Segmentation in the Clinic: A Grand Challenge" was organised [21] which analysed and compared the performance of several state-of-the-art automatic and semi-automatic liver segmentation algorithms for computerised tomography (CT) images. For evaluation a scoring system was used

which first calculates 5 different segmentation error measures, calibrates the results to bring the measures to a unique scale, and finally takes the average of the results in this scale to get a single score for a segmentation approach (refer to [21] for details).

Cárdenes et al. [6] define new image segmentation evaluation measures by introducing position and intensity values of misclassified voxels into the established overlap-based measures. Two other measures proposed rely on the connectivity of segmented regions and limiting overlap-based measures (Jaccard coefficient (see Section 6.4) in this case) to only region boundaries respectively. Finally, the authors combine these measures into an aggregated multimodal similarity measure. In a series of experiments on real and synthetic data they reveal that the new measures show more variability and sensitivity when comparing segmentation algorithms than the classic measures like Jaccard coefficient.

4 Evaluation Methods

One of the key aspects of a systematic classification of these measures is experimental data and examples. Crucially, none of the aforementioned surveys illustrate their findings on concrete examples.

We present a systematic summary of several empirical discrepancy evaluation methods frequently used in medical image segmentation applications. For an object label we consider the machine segmented set of voxels MS, the ground truth GT, which is commonly acquired with manual segmentation by human experts, and the image being segmented I. The operator $|\cdot|$ returns the number of pixels (or voxels in 3D) contained in a region which is proportional to the physical volume of the considered region.

A voxel in the considered region is said to be on its boundary if at least one of the neighbouring voxels does not belong to the region. There is no single consensus for the choice of the neighbourhood; for instance, some authors consider the 18- [21], others the 26-neighbourhood, etc. We denote the boundaries of MS and GT as B_{MS} and B_{GT} respectively.

While most medical image segmentation papers present some quantitative evaluation for suggested algorithms, there are also works which rely solely on qualitative analysis of the results [18, 10, 35, 24]. This is mainly due to the difficulty of acquiring high quality ground truth segmentations for medical images. Manual segmentation in a slice-by-slice manner by an expert radiologist is considered the common source of gold standard segmentations. But this process is extremely time consuming; an operator may spend several hours segmenting a 3D image volume of a few dozen slices. In addition, manual segmentation is exposed to intra- and inter-observer variability.

It is crucial not to confuse one measure for another despite naming similarities. We try to maintain the original names in order not to introduce multiple names to same measures (there is quite a lot of naming diversity in the existing literature).

5 Size Based Methods

5.1 Relative volume difference (RVD)

A very simple measure of dissimilarity of two segmented volumes is their difference in size as a fraction of the size of the reference:

$$RVD = \pm \frac{|MS| - |GT|}{|GT|} \tag{1}$$

Lim, Jeong and Ho [25], Linguraru et al. [26] use formula (1) (referred as volume error) to compare automatic segmentation results to manually traced ground truth. In [22] this measure (called volume difference) is used for inter-observer variability in hand-seeded kidney segmentation along with statistical measures. In [32, 8, 11] it is used as one of several measures combined into the MICCAI 2007 scoring system [21].

6 Overlap Based Methods

6.1 Dice's similarity coefficient (Dice), symmetric volume difference (SVD)

Dice's similarity coefficient, originally introduced by Dice [15] for ecological studies, is one of the most frequently used evaluation measures in medical image segmentation. The measure quantifies the match of two sets A and B by normalising the size of their intersection over the average of their sizes:

$$Dice = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)} = \frac{2\,|A \cap B|}{|A| + |B|} \tag{2}$$

When used for evaluating the results of machine segmentation against gold standard the coefficient looks like:

$$Dice = \frac{2\,|MS \cap GT|}{|MS| + |GT|} \tag{3}$$

In an impressive number of medical image segmentation papers [26, 37, 12, 20, 19] the authors present evaluation results using several different measures but prefer Dice's similarity coefficient as the main indicator of segmentation algorithm accuracy. In Linguraru et al. [27] *Dice* results are shown not only for computer segmentation and manual labelling overlap but also

for comparing inter-observer measurements. Nguyen and Wu [29] use the Dice coefficient for simulated (phantom) and real brain images segmentation evaluation.

In [1, 16, 31, 30] Dice's similarity coefficient is one of several metrics used for segmentation algorithm evaluation.

A number of medical image segmentation papers including [14] rely solely on equation (3) to validate their final segmentation results.

The *symmetric volume difference (SVD)*, provides a symmetric measure of the difference in volume of the segmentation result and the reference shape.

$$SVD = 1 - Dice = 1 - \frac{|MS \cap GT|}{\frac{1}{2}(|MS| + |GT|)} \tag{4}$$

The segmentation errors are estimated with SVD by others [33, 23].

The term *symmetric volume overlap (SVO)* is introduced by Campadelli et al. [2] to refer to a measure defined as $SVO = 1 - SVD$, thus arriving at *Dice's similarity coefficient*. So, Dice is also used by Campadelli, Casiraghi and their colleagues [3, 2, 5, 4] as a main similarity measure, but referred to as SVO, along with other measures.

6.2 True positive (TPVF), true negative (TNVF), false positive (FPVF) and false negative (FNVF) volume fractions

Udupa et al. [34] consider four evaluation measures — *true positive (TPVF), true negative (TNVF), false positive (FPVF) and false negative (FNVF) volume fractions* — borrowed from statistical decision theory measures (*sensitivity* and *specificity*):

$$TPVF = Sens = \frac{|TP|}{|TP + FN|} = \frac{|MS \cap GT|}{|GT|} \tag{5}$$

$$TNVF = Spec = \frac{|TN|}{|TN + FP|} = \frac{|I| - |MS \cup GT|}{|I| - |GT|} \tag{6}$$

$$FPVF = 1 - Spec = \frac{|FP|}{|TN + FP|} = \frac{|MS \setminus GT|}{|I| - |GT|} \tag{7}$$

$$FNVF = 1 - Sens = \frac{|FN|}{|TP + FN|} = \frac{|GT \setminus MS|}{|GT|} \tag{8}$$

Here TP denotes the set of object voxels labelled as object, FN is the set of object voxels labelled as non-object, the voxels comprising the non-object area but labelled as object make up FP and, finally, TN are the non-object voxels successfully identified as such by the machine segmentation approach (refer to Figure 1 for an illustration). Only two of the suggested measures

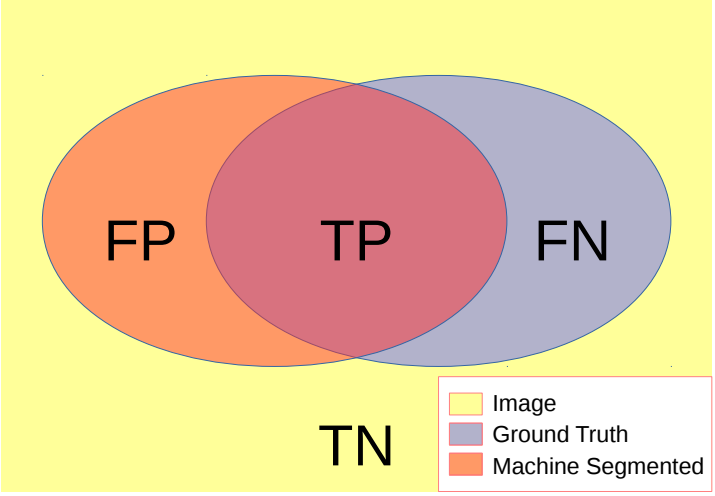

Figure 1: A graphical illustration of true positive, true negative, false positive and false negative regions when comparing machine segmentation results with ground truth.

should be used together (e.gTPVF and FPVF, but not TPVF and FNVF) due to the dependence relationships present in (5)-(8).

Some authors [32, 28] give an alternative definition for FPVF by normalising FP over the ground truth GT rather than the rest of the image $I \setminus GT$:

$$FPVF_{Alter} = \frac{|FP|}{|TP + FN|} = \frac{|MS \setminus GT|}{|GT|} \tag{9}$$

The TPVF-TNVF pair is used in [20, 16] (referenced as sensitivity and specificity) and [1] (as TPVF and 1-FPVF), while [9, 7, 8] compute the TPVF-FPVF pair. Liu et al. [28] (referred as TP and FP ratios), Ruskó et al. [32] use TPVF, $\text{FPVF}_{\text{Alter}}$ along with other evaluation measures. Liu et al. [28] also report FNVF (FN ratio) results although these are easily inferred from TPVF values.

6.3 Precision and recall

The *precision* and *recall* measures reflect the similarities in the volumes of the automatically detected regions and the ground truth. These are empirical discrepancy measures based on the number of mis-classified and correctly classified voxels.

The precision normalises the volume of the correctly segmented region, $MS \cap GT$, over the volume of the result of the segmentation, MS:

$$Precision = \frac{|MS \cap GT|}{|MS|} \tag{10}$$

The recall normalises $MS \cap GT$ over the volume of the gold standard, GT:

$$Recall = \frac{|MS \cap GT|}{|GT|} \tag{11}$$

We acknowledge that recall and true positive volume fraction (or sensitivity) are the same measure but redefine it here to emphasise its paired use with precision.

The precision does not account for under-segmentation errors, while the over-segmented volumes are not reflected in the recall. The pair of the measures is used by Campadelli, Casiraghi and their colleagues [3, 2, 5, 4] (referenced as sensitivity ratio and overlap ratio), Wolz et al. [37].

6.4 Jaccard similarity coefficient (Jaccard), volumetric overlap error (VOE)

Another measure used in Liu et al. [28] is *Jaccard similarity coefficient* (presented as *similarity ratio*) defined as the number of common voxels of the machine segmented and ground truth regions over their union:

$$Jaccard = \frac{|MS \cap GT|}{|MS \cup GT|} \tag{12}$$

Jaccard is also used in [37, 30, 19].

Volumetric overlap error (VOE) is the corresponding error measure [32]:

$$VOE = 1 - \frac{|MS \cap GT|}{|MS \cup GT|} \tag{13}$$

It is one of the measures in the MICCAI 2007 scoring system and has been used in [8, 11] as part of it.

7 Surface Distance Based Measures

Let a distance measure for a voxel x from a set of voxels A be defined as:

$$d(x, A) = \min_{y \in A} d(x, y) \tag{14}$$

where $d(x, y)$ is the Euclidean distance of the voxels incorporating the real spatial resolution of the image. A number of segmentation evaluation measures are based on this distance definition and quantify the dissimilarity of the machine segmentation from the ground truth. The most popular of these

are presented below. They are all quantified in millimetres (mm), value of 0 corresponding to perfect match between MS and GT, and greater values indicating higher errors.

7.1 Average symmetric surface distance (ASD)

The *average symmetric surface distance (ASD)* is the average of all the distances from points on the boundary of MS to the boundary of GT and from points on B_{GT} to B_{MS}, respectively:

$$ASD = \frac{1}{|B_{MS}| + |B_{GT}|} \times \left(\sum_{x \in B_{MS}} d(x, B_{GT}) + \sum_{y \in B_{GT}} d(y, B_{MS}) \right) \quad (15)$$

It is the third measure in the MICCAI 2007 scoring system, see [21, 32, 8, 11] for numeric results with this measure. Yokota et al. [39] rely only on ASD while others [23] (as Mean Distance), [26, 31] present results with several measures including ASD.

7.2 Root mean square symmetric surface distance (RMSD)

The penultimate measure in the MICCAI 2007 challenge evaluation scoring system is *root mean square symmetric surface distance (RMSD)* [21] and is defined as:

$$RMSD = \sqrt{\frac{1}{|B_{MS}| + |B_{GT}|}} \times \sqrt{\sum_{x \in B_{MS}} d^2(x, B_{GT}) + \sum_{y \in B_{GT}} d^2(y, B_{MS})} \quad (16)$$

It is used as part of the scoring system in [32, 8, 11] and as a separate measure along with others in [26, 23].

7.3 Maximum symmetric surface distance (MSD)

The *Hausdorff distance*[1] of two sets A and B defines the maximal distance from a point in the first to a nearest point in the other one [17]:

$$d_H(A, B) = \max_{x \in A} \min_{y \in B} d(x, y) = \max_{x \in A} d(x, B) \quad (17)$$

The symmetric variant of the Hausdorff metric for the boundaries of the segmented regions is referred to as *maximum symmetric surface distance (MSD)* in image segmentation evaluation:

$$MSD = \max \{ d_H(B_{MS}, B_{GT}), d_H(B_{GT}, B_{MS}) \} \quad (18)$$

[1] We define the Hausdorff metric as the directed asymmetric distance of two sets. Some literature refers to the symmetric version with the same name.

For each voxel on the boundary of MS there is guaranteed to be a voxel of GT in a distance of at most MSD, and vice versa.

This is the last of the five measures used in the MICCAI 2007 challenge [21] and within that evaluation scheme is used in [32, 8, 11]. Liu et al. [28] use MSD but refer to it as Hausdorff distance.

8 Discussion and Results

8.1 Summary

The similarity/difference measures which are popular in the literature fail to capture all the aspects of segmented regions. Size based measures rely only on the difference in size of the segmented region and the gold standard. As a result, a segmented region which does not even intersect the gold standard may receive a highest score as long as it is of the same size.

Overlap based methods account only the number of correctly or misclassified voxels without reflecting their spatial distribution. So, a segmentation result which has leaked slightly to neighbouring tissue will be regarded as equivalently good/bad as a result without leakage but with a separate disconnected region (of the same size as the leakage area). Figure 2 illustrates this.

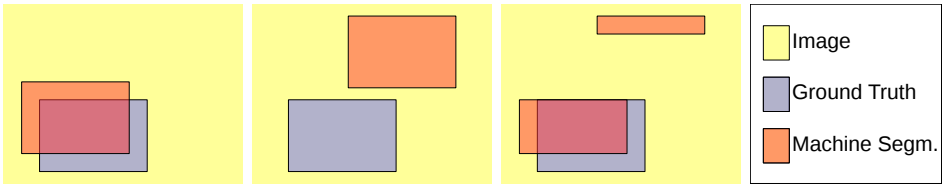

Figure 2: (a) and (b) get the same size based score; (a) and (c) get the same overlap based score

The distance based methods discussed take into account only the minimal distance from the boundary of the other region at each boundary point. Thus, they ignore the actual volume differences of the machine segmented and gold standard regions as depicted in Figure 3. Another issue of distance based measures is the value range in units of length which makes it difficult to compare results for images of different spatial resolution and quality.

We conclude that none of the measures from the literature that we presented is suitable to act as a reliable measure to reflect all the aspects of segmentation accuracies and errors. We intend to introduce new simple measures that overcome these problems, are easy to implement and have a fixed value range.

Most of the presented measures are very different in nature and authors are usually encouraged to report results with several measures to allow for

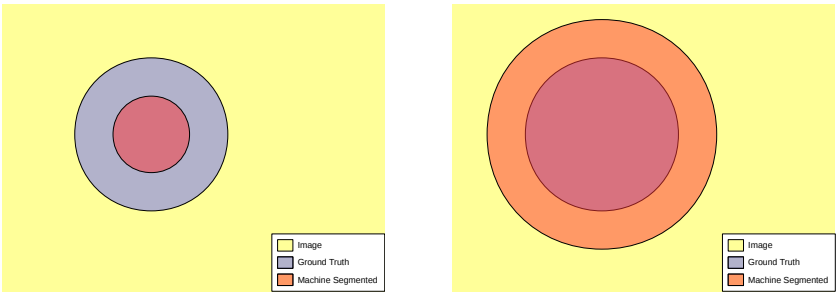

Figure 3: Segmentation results in the two images get the same score with the discussed distance based evaluation methods.

easy comparison of results with other literature. Nevertheless, the presented information can be redundant if there are dependencies between the measures. One such example is the pair Jaccard-Dice:

$$Jaccard = \frac{Dice}{2 - Dice} \tag{19}$$

Both measures have a value range $[0, 1]$, reach the interval endpoints at the same time, Dice is always bigger between the endpoints. The quartet of true positive, true negative, false positive and false negative volume fractions is another example where the true positive–false negative and true negative–false positive pairs should not be used together.

One advantage of overlap based measures over the size and surface distance based measures is their fixed value range $[0, 1]$, sometimes reported as percentage $[0, 100]$. This makes the results reported in different experiments more easily comparable (ignoring the issue of gold standard consistency), unlike the surface distance based measures with their mm range $[0, +\infty)$. An additional complication with the use of surface distance methods is the variety of the ways region borders can be defined, depending on the chosen neighbourhood size.

8.2 Experiments

Below we present results of experiments that we carried out to compare the performance of the size based and overlap based measures discussed. We ran the segmentation algorithm described in [38] on 5 different CT image volumes to label liver and left kidney for which we also produced gold standard masks. Our dataset contained abdominal CT images of low resolution (5mm slice thickness, 0.68–0.78mm pixel resolution) between 1 and 21 slices each, acquired with or without contrast agent administration and covered both healthy and diseased organs. Visual inspection of the results revealed that the liver was significantly undersegmented in the first volume (big part of it missed by the machine segmentation algorithm) and over-segmentation

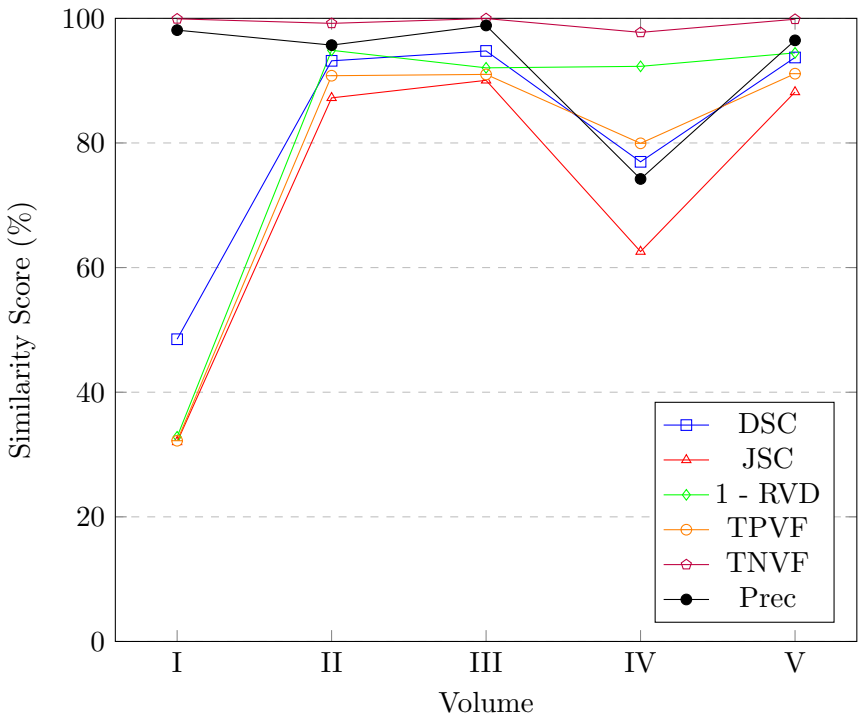

Figure 4: Liver results for 5 image volumes with different similarity measures

of kidney took place in the fourth volume (machine segmentation included a large necrotic tumour into the kidney region).

We considered size and overlap based similarity measures, replacing dissimilarity measures with the corresponding similarity measures so that we could compare like with like. Current set of measures includes:

- Dice's similarity coefficient (DSC)
- Jaccard similarity coefficient (JSC)
- 1 - relative volume difference (RVD)
- true positive volume fraction (TPVF)
- true negative volume fraction (TNVF)
- precision (Prec)

Further experiments will be carried out to include analysis of the behaviour of the surface distance based measures.

In Figure 4 we report evaluation of the liver segmentation results of five data volumes (I – V) against the manual labellings measured with the 6

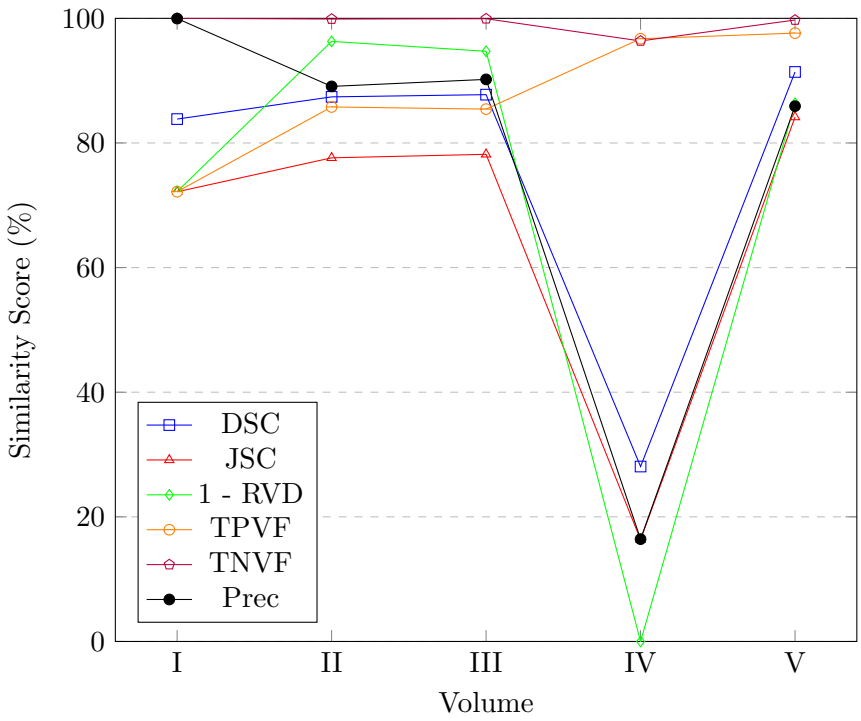

Figure 5: Left kidney results for 5 image volumes with different similarity measures

similarity measures. TNVF and Prec give a very high score to the under-segmented liver results on Volume I. In addition, TNVF shows very little variability on the 5 volumes. The outcome of left kidney segmentation evaluation is presented in Figure 5. The over-segmentation of the left kidney is missed by TPVF and TNVF. Although 1-RVD does not show extreme deviations from the average results in both tables, still its curve reveals inconsistent scoring (see liver score in Volume IV, for instance). This is because it incorporates only size information of the labelled regions ignoring the amount of common voxels.

Another experiment performed included a single volume being manually segmented by two different operators. We compared the inter-observer (IO) variability results with the 6 measures against machine segmentation (MS) evaluation outcome. In other words, with each measure we once compared manual segmentation against another manual segmentation and once against machine segmentation. The outcome is presented in Table 1. The measures which show bigger difference between the IO and MS results are more sensitive to machine segmentation errors and should be preferred. This is because a better agreement is expected between the manual segmentations than between a manual and a machine segmentation. From that perspective,

Liver segmentation similarity (%)

	DSC	JSC	1-RVD	TPVF	TNVF	Prec
IO	97.9	95.9	99.5	97.7	99.9	98.2
MS	94.8	90.1	92.1	91.0	99.9	98.9

Right kidney segmentation similarity (%)

	DSC	JSC	1-RVD	TPVF	TNVF	Prec
IO	93.9	88.4	98.4	93.1	99.9	94.6
MS	90.8	83.2	89.4	95.7	99.9	86.5

Table 1: Inter-observer (IO) similarity and machine segmentation (MS) results with different measures for a single image volume

TPVF, TNVF and Prec fail in this experiment.

Finally, we report the IO results for the single volume along with the mean and standard deviation of the segmentation scores over 5 volumes in Figure 6. The preferred measures should show an average below the IO score and a big standard deviation; the results agree with the previous findings.

9 Conclusions and Future Work

To conclude, the results of our current experiments give preference to Dice's and Jaccard similarity coefficients over 1-RVD, TPVF, TNVF and Prec. Taking into account the link between DSC and JSC, only one of those should be used for segmentation evaluation. Also, we appreciate that the rest of the measures can still be useful as segmentation evaluation measures, especially if reported together at the same time. Our main goal in these experiments was to reveal the goodness of the discussed measures as separate evaluation methods.

In future we hope to produce a set of simple synthetic images representing simulated pairs of reference and machine segmented regions and conduct experiments to assess and compare the performance of the considered evaluation measures. We intend to include a variety of reference shapes to reflect the diversity of patterns appearing in medical images. The segmentation result simulations should include errors of varied nature to represent mistakes of segmentation algorithms. We want to analyse the performance of evaluation measures on pairs of these regions and reveal their response to errors of different nature and scale. Hopefully, this will provide enough data to introduce some scoring for the discussed measures for the task of medical image segmentation evaluation.

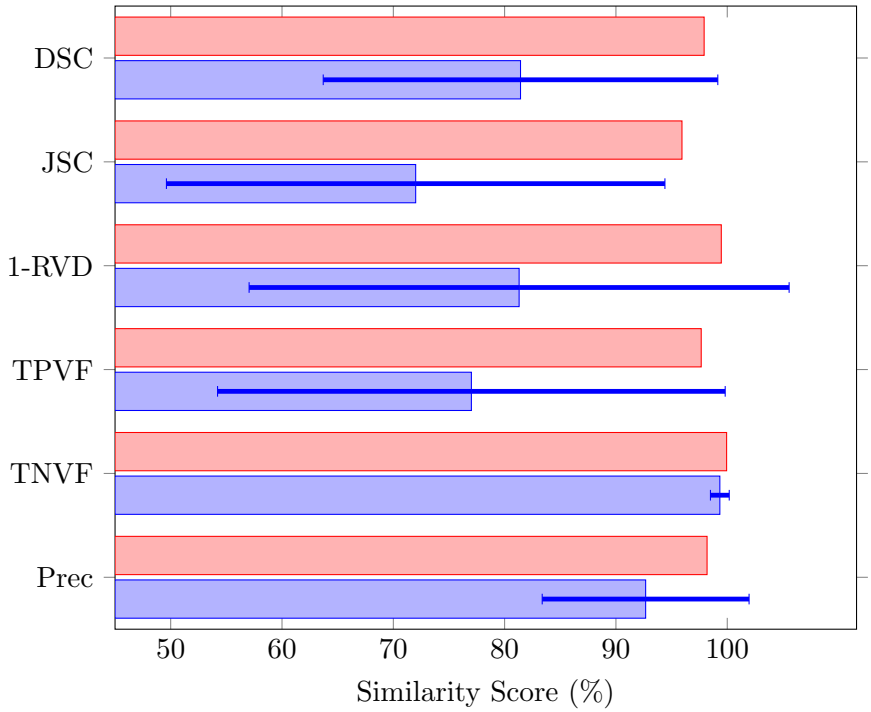

Figure 6: A comparison of the inter-observer similarity for liver in a single volume with segmentation average scores over 5 volumes (and standard deviation) for different measures.

References

[1] Ulas Bağci, Xinjian Chen, and Jayaram K. Udupa. Hierarchical scale-based multiobject recognition of 3-D anatomical structures. *IEEE Transactions on Medical Imaging*, 31(3):777–789, 2012.

[2] Paola Campadelli, Elena Casiraghi, and Andrea Esposito. Liver segmentation from computed tomography scans: A survey and a new algorithm. *Artificial Intelligence in Medicine*, 45(2-3):185–196, 2009. Computational Intelligence and Machine Learning in Bioinformatics.

[3] Paola Campadelli, Elena Casiraghi, and Stella Pratissoli. Fully automatic segmentation of abdominal organs from CT images using fast marching methods. In *21st IEEE International Symposium on Computer-Based Medical Systems, 2008. CBMS'08*, pages 554–559. IEEE, June 2008.

[4] Paola Campadelli, Elena Casiraghi, and Stella Pratissoli. A segmentation framework for abdominal organs from CT scans. *Artificial Intelligence in Medicine*, 50(1):3–11, 2010.

[5] Paola Campadelli, Elena Casiraghi, Stella Pratissoli, and Gabriele Lombardi. Automatic abdominal organ segmentation from CT images. *Electronic Letters on Computer Vision and Image Analysis*, 8(1):1–14, July 2009.

[6] Rubén Cárdenes, Rodrigo de Luis-García, and Meritxell Bach-Cuadra. A multidimensional segmentation evaluation for medical image data. *Computer Methods and Programs in Biomedicine*, 96(2):108–124, 2009.

[7] Xinjian Chen and Ulas Bagci. 3D automatic anatomy segmentation based on iterative graph-cut-ASM. *Medical Physics*, 38(8):4610–4622, August 2011.

[8] Xinjian Chen, Jayaram K. Udupa, Ulas Bagci, Ying Zhuge, and Jianhua Yao. Medical image segmentation by combining graph cuts and oriented active appearance models. *IEEE Transactions on Image Processing*, 21(4):2035–2046, 2012.

[9] Xinjian Chen, Jianhua Yao, Ying Zhuge, and Ulas Bagci. 3D automatic anatomy segmentation based on graph cut-oriented active appearance models. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 3653–3656, September 2010.

[10] Yu Chen, Zhuangzhi Yan, and Yungao Chu. Cellular automata based level set method for image segmentation. In *IEEE/ICME International Conference on Complex Medical Engineering*, pages 171–174. IEEE, 2007.

[11] Yufei Chen, Zhicheng Wang, Jinyong Hu, Weidong Zhao, and Qidi Wu. The domain knowledge based graph-cut model for liver CT segmentation. *Biomedical Signal Processing and Control*, 7(6):591–598, 2012.

[12] Najeeb Chowdhury, Robert Toth, Jonathan Chappelow, Sung Kim, Sabin Motwani, Salman Punekar, Haibo Lin, Stefan Both, Neha Vapiwala, Stephen Hahn, and Anant Madabhushi. Concurrent segmentation of the prostate on MRI and CT via linked statistical shape models for radiotherapy planning. *Medical Physics*, 39(4):2214–2228, 2012.

[13] William R. Crum, Oscar Camara, and Derek L. G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, November 2006.

[14] Rémi Cuingnet, Raphael Prevost, David Lesage, Laurent D. Cohen, Benoît Mory, and Roberto Ardon. Automatic detection and segmentation of kidneys in 3D CT images using random forests. In Nicholas Ayache, Hervé Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, volume 7512 of *Lecture Notes in Computer Science*, pages 66–74. Springer Berlin Heidelberg, 2012.

[15] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[16] Hong-Seng Gan, Tian-Swee Tan, Ahmad Helmy Abdul Karim, Khairil Amir Sayuti, and Mohammed Rafiq Abdul Kadir. Interactive medical image segmentation with seed precomputation system: Data from the osteoarthritis initiative. In *IEEE Conference on Biomedical Engineering and Sciences (IECBES)*, pages 315–318. IEEE, December 2014.

[17] Guido Gerig, Matthieu Jomier, and Miranda Chakos. Valmet: A new validation tool for assessing and improving 3D object segmentation. In Wiro J. Niessen and Max A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001*, volume 2208 of *Lecture Notes in Computer Science*, pages 516–523. Springer Berlin Heidelberg, 2001.

[18] Payel Ghosh, Sameer K. Antani, L. Rodney Long, and George R. Thoma. Unsupervised grow-cut: Cellular automata-based medical image segmentation. In *First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 40–47. IEEE, 2011.

[19] Stuart Golodetz. *Zipping and Unzipping: The Use of Image Partition Forests in the Analysis of Abdominal CT Scans*. DPhil thesis, University of Oxford, 2011.

[20] Vicente Grau, A. U. J. Mewes, M. Alcañiz, Ron Kikinis, and Simon K. Warfield. Improved watershed transform for medical image segmentation using prior information. *IEEE Transactions on Medical Imaging*, 23(4):447–458, April 2004.

[21] Tobias Heimann, Bram van Ginneken, Martin A. Styner, Yulia Arzhaeva, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, August 2009.

[22] Claude Kauffmann and Nicolas Piché. Seeded ND medical image segmentation by cellular automaton on GPU. *International Journal of Computer Assisted Radiology and Surgery*, 5(3):251–262, 2010.

[23] Hans Lamecker, Thomas Lange, and Martin Seebass. Segmentation of the liver using a 3D statistical shape model. Technical Report 4-9, ZIB, Takustr.7, 14195 Berlin, 2004.

[24] Bing Nan Li, Chee Kong Chui, Stephen Chang, and S. H. Ong. Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. *Computers in Biology and Medicine*, 41(1):1–10, 2011.

[25] Seong-Jae Lim, Yong-Yeon Jeong, and Yo-Sung Ho. Automatic liver segmentation for volume measurement in CT images. *Journal of Visual Communication and Image Representation*, 17(4):860–875, 2006.

[26] Marius George Linguraru, John A. Pura, Vivek Pamulapati, and Ronald M. Summers. Statistical 4D graphs for multi-organ abdominal segmentation from multiphase CT. *Medical Image Analysis*, 16(4):904–914, 2012.

[27] Marius George Linguraru, Jianhua Yao, Rabindra Gautam, James Peterson, Zhixi Li, W. Marston Linehan, and Ronald M. Summers. Renal tumor quantification and classification in contrast-enhanced abdominal CT. *Pattern Recognition*, 42(6):1149–1161, 2009.

[28] Yan Liu, H. D. Cheng, Jianhua Huang, Yingtao Zhang, and Xianglong Tang. An effective approach of lesion segmentation within the breast ultrasound image based on the cellular automata principle. *Journal of Digital Imaging*, 25(5):580–590, 2012.

[29] Thanh Minh Nguyen and Q. M. Jonathan Wu. Robust Student's-t mixture model with spatial constraints and its application in medical image segmentation. *IEEE Transactions on Medical Imaging*, 31(1):103–116, 2012.

[30] Toshiyuki Okada, Marius George Linguraru, Masatoshi Hori, Ronald M. Summers, Noriyuki Tomiyama, and Yoshinobu Sato. Abdominal multi-organ CT segmentation using organ correlation graph and prediction-based shape and location priors. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, volume 8151 of *Lecture Notes in Computer Science*, pages 275–282. Springer Berlin Heidelberg, 2013.

[31] Renzo Phellan, Alexandre X. Falcão, and Jayaram Udupa. Improving atlas-based medical image segmentation with a relaxed object search. In Yongjie Jessica Zhang and João Manuel R. S. Tavares, editors, *Computational Modeling of Objects Presented in Images. Fundamentals, Methods, and Applications*, volume 8641 of *Lecture Notes in Computer Science*, pages 152–163. Springer International Publishing, 2014.

[32] László Ruskó, György Bekes, and Márta Fidrich. Automatic segmentation of the liver from multi- and single-phase contrast-enhanced CT images. *Medical Image Analysis*, 13(6):871–882, 2009.

[33] Andrea Schenk, Guido Prause, and Heinz-Otto Peitgen. Efficient semiautomatic segmentation of 3Dobjectsinmedicalimages. In Scott L. Delp, Anthony M. DiGoia, and Branislav Jaramaz, editors, *Medical Image Computing and Computer-Assisted Intervention*, volume 1935 of *Lecture Notes in Computer Science*, pages 186–195. Springer Berlin Heidelberg, 2000.

[34] Jayaram K. Udupa, Vicki R. LeBlanc, Ying Zhuge, Celina Imielinska, Hilary Schmidt, Leanne M. Currie, Bruce E. Hirsch, and James Woodburn. A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics*, 30(2):75–87, 2006.

[35] Vladimir Vezhnevets and Vadim Konouchine. "GrowCut" - interactive multi-label N-D image segmentation by cellular automata. In *Proceedings of Graphicon*, pages 150–156, 2005.

[36] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[37] Robin Wolz, Chengwen Chu, Kazunari Misawa, Kensaku Mori, and Daniel Rueckert. Multi-organ abdominal CT segmentation using hierarchically weighted subject-specific atlases. In Nicholas Ayache, Hervé Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, volume 7510 of *Lecture Notes in Computer Science*, pages 10–17. Springer Berlin Heidelberg, 2012.

[38] Varduhi Yeghiazaryan and Irina Voiculescu. The use of fast marching methods in medical image segmentation. Technical Report CS-RR-15-07, Department of Computer Science, University of Oxford, Oxford, UK, 2015.

[39] Futoshi Yokota, Toshiyuki Okada, Masaki Takao, Nobuhiko Sugano, Yukio Tada, Noriyuki Tomiyama, and Yoshinobu Sato. Automated CT segmentation of diseased hip using hierarchical and conditional statistical shape models. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pages 190–197. Springer Berlin Heidelberg, 2013.

[40] Hui Zhang, Sharath Cholleti, Sally A. Goldman, and Jason E. Fritts. Meta-evaluation of image segmentation using machine learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1138–1145. IEEE, 2006.

[41] Hui Zhang, Jason E. Fritts, and Sally A. Goldman. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2):260–280, 2008.

[42] Yu Jin Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.

[43] Yu Jin Zhang. A review of recent evaluation methods for image segmentation. In *Sixth International Symposium on Signal Processing and its Applications*, volume 1, pages 148–151. IEEE, 2001.

[44] Yu Jin Zhang. Image segmentation evaluation in this century. *Encyclopedia of Information Science and Technology. Beijing, Tsinghua University, China*, pages 1812–1817, 2009.