Chapter 6
Evaluation Metrics for Medical Organ Segmentation and Lesion Detection

Abdel Aziz Taha and Allan Hanbury

Abstract This chapter provides an overview of the metrics used in the VISCERAL segmentation benchmarks, namely Anatomy 1, 2 and 3. In particular, it provides an overview of 20 evaluation metrics for segmentation, from which four metrics were selected to be used in VISCERAL benchmarks. It also provides an analysis of these metrics in three ways: first by analysing fuzzy implementations of these metrics using fuzzy segmentations produced either synthetically or by fusing participant segmentations and second by comparing segmentation rankings produced by these metrics with rankings performed manually by radiologists. Finally, a metric selection is performed using an automatic selection framework, and the selection result is validated using the manual rankings. Furthermore, this chapter provides an overview of metrics used for the Lesion Detection Benchmark.

Source code is available at:
https://github.com/visceral-project/EvaluateSegmentation

6.1 Introduction

The importance of using suitable metrics in evaluation stems from the fact that there are different metrics, and each of them has particular sensitivities and thus measures particular aspects of similarity/discrepancy between the objects being evaluated and the corresponding ground truth. Poorly defined metrics may lead to inaccurate conclusions about the state-of-the-art algorithms, which negatively impacts system

A.A. Taha · A. Hanbury (✉)
Institute of Software Technology and Interactive Systems, TU Wien,
Favoritenstraße 9-11/188, 1040 Vienna, Austria
e-mail: hanbury@ifs.tuwien.ac.at

A.A. Taha
e-mail: taha@ifs.tuwien.ac.at

A. Hanbury et al. (eds.), *Cloud-Based Benchmarking of Medical Image Analysis*, DOI 10.1007/978-3-319-49644-3_6

development. This chapter provides an overview of metrics used for the Anatomy and Detection Benchmarks of the VISCERAL project [1].

Segmentation methods with high accuracy and high reproducibility are a main goal in medical image processing. Therefore, assessing the accuracy and the quality of segmentation algorithms is of great importance, which is a matter of the evaluation methodology. Segmentation evaluation is the task of comparing two segmentations by measuring the distance or similarity between them, where one is the segmentation to be evaluated and the other is the corresponding ground truth segmentation. In this chapter, we provide an overview of a metric pool consisting of twenty metrics for evaluating medical image segmentations and a subset of four metrics that were considered in the VISCERAL segmentation benchmarks.

The knowledge about the metrics in terms of their strength, weakness, sensitivities, bias, as well as their ability to deal with fuzzy segmentation, is essential for taking the decision about which metrics are to be used in the evaluation. In this chapter, we provide an analysis of metrics with respect to their fuzzy definitions and discussion about selecting suitable metrics for evaluating segmentation from a metric pool.

Apart from segmentation, the VISCERAL project had also the Lesion Detection Benchmark, where lesions are to be localized by detection algorithms. In this chapter, we provide an overview of the metrics and evaluation methodologies that were used for the Detection Benchmark.

The remainder of this chapter is organized as follows: in Sect. 6.2, we provide an overview of the metrics that were used in the VISCERAL Anatomy and Detection Benchmarks. In Sect. 6.3, we validate a subset of the segmentations of the Anatomy 2 Benchmark against synthetic fuzzy variants of the ground truth and discuss the results. In Sect. 6.4, we present an analysis based on the comparison between rankings produced by the segmentation metrics and manual rankings made by radiologists. Finally, this chapter is concluded in Sect. 6.5.

6.2 Metrics for VISCERAL Benchmarks

In this section, we provide an overview of the metrics that were used in the VISCERAL benchmarks. In particular, we provide a pool of metrics for evaluating medical image segmentation, from which four metrics were selected for the VISCERAL Anatomy Benchmarks. Furthermore, we provide an overview of metrics that were used for the Detection Benchmark.

6.2.1 Metrics for Segmentation

Medical image segmentation assigns each voxel of a medical image to a class, e.g. an anatomical structure. While this assignment is crisp in binary segmentation, it takes

Table 6.1 Overview of evaluation metrics for 3D image segmentation. The symbols in the second column are used to denote the metrics throughout the chapter. The column "category" assigns each metric to one of the categories above. The column "Fuzzy" indicates whether a fuzzy implementation of the metric is available

Metric	Symbol	Category	Fuzzy
Dice coefficient	$DICE$	Spatial overlap based	yes
Jaccard index	JAC	Spatial overlap based	yes
True-positive rate (sensitivity, recall)	TPR	Spatial overlap based	yes
True-negative rate (specificity)	TNR	Spatial overlap based	yes
False-positive rate (= 1-specificity, fallout)	FPR	Spatial overlap based	yes
False-negative rate (= 1-sensitivity)	FNR	Spatial overlap based	yes
F-measure (F1-measure = Dice)	FMS	Spatial overlap based	yes
Global consistency error	GCE	Spatial overlap based	no
Volumetric similarity	VS	Volume based	yes
Rand index	RI	Pair counting based	yes
Adjusted Rand index	ARI	Pair counting based	yes
Mutual information	MI	Information theoretic based	yes
Variation of information	VOI	Information theoretic based	yes
Interclass correlation	ICC	Probabilistic based	no
Probabilistic distance	PBD	Probabilistic based	yes
Cohen's kappa	KAP	Probabilistic based	yes
Area under ROC curve	AUC	Probabilistic based	yes
Hausdorff distance	HD	Spatial distance based	no
Average distance	AVD	Spatial distance based	no
Mahalanobis distance	MHD	Spatial distance based	no

other forms in fuzzy segmentation, e.g. the degree of membership or the probability that a particular voxel belongs to a particular class. An automatic segmentation is validated by comparing it with the corresponding ground truth segmentation using an evaluation metric.

We describe the metrics for validating medical segmentation in Table 6.1, which were selected based on a literature review of papers in which medical volume segmentations are evaluated. Only metrics with at least two references (papers) of use are considered. These metrics were implemented in the EvaluateSegmentation[1] tool for evaluating medical image segmentation. Taha and Hanbury [4] provide definitions and a comprehensive analysis of these metrics as well as guidelines for metric selection based on the properties of the segmentations being evaluated and the segmentation goal.

[1] EvaluateSegmentation is open source software for evaluating medical image segmentation available at https://github.com/visceral-project/EvaluateSegmentation.

Based on the relations between the metrics, their nature and their definition, we group them into six categories, namely:

- **Spatial overlap based (Category 1)**: These are metrics defined based on the spatial overlap between the two segmentations being compared, namely the four basic overlap cardinalities—true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).
- **Volume based (Category 2)**: Metrics from this category are based on comparing the volume of the segmented region, i.e. they aim to measure the number of voxels segmented compared with the number of voxels in the true segmentation (ground truth).
- **Pair counting based (Category 3)**: Metrics from this category are based on $\binom{n}{2}$ tuples that represent all possible voxel pairs in the image. These tuples can be grouped into four categories depending on where the voxels of each pair are placed according to each of the segmentations being compared. These four groups are Group I: if both voxels are placed in the same segment in both segmentations; Group II: if both voxels are placed in the same segment in the first segmentation but in different segments in the second; Group III: if both voxels are placed in the same segment in the second segmentation but in different segments in the first; and Group IV: if both voxels are placed in different segments in both segmentations.
- **Information theoretic based (Category 4)**: Metrics of this category are based on basic values of information theory such as entropy and mutual information.
- **Probabilistic based (Category 5)**: These metrics consider the segmentations being compared as two distributions. Under this consideration, the metrics are defined based on the classic comparison methods of statistics of these distributions.
- **Spatial distance based (Category 6)**: These metrics aim to summarize distances between all pairs of voxels in the two segmentations being compared, i.e. they provide a one-value measure that represents all pairwise distances.

The aim of this grouping is to enable a reasonable selection when a subset of metrics is to be used, i.e. selecting metrics from different groups to avoid biased results.

For the evaluation of medical image segmentation in the VISCERAL Anatomy Benchmarks, four metrics were selected from the 20 metrics presented in Table 6.1. The following criteria were considered:

- The metrics were selected so that they cover as many different categories as possible from those categories described above.
- From those metrics that meet the criteria above, metrics were selected that have the highest correlation with the rest of the metrics in each category.

Based on these criteria, the following metrics were considered for validating segmentations in all the segmentation benchmarks of the VISCERAL project: the Dice coefficient (DICE), the average distance (AVD), the interclass correlation (ICC) and the adjusted Rand index (ARI).

6.2.2 Metrics for Lesion Detection

The Detection Benchmark considered pathology instead of anatomy. The goal of the benchmark is to automatically detect lesions in images acquired in clinical routine.

In the Detection Benchmark, an annotated lesion, L_i, is represented by three points, namely the centre of the lesion, C_i, and two other points, $D1_i$ and $D2_i$, indicating the diameter of the lesion. Participating algorithms are expected to provide per lesion exactly one point, P_i, as near as possible to the centre of the lesion, C_i.

As mentioned above, it is expected that exactly one point per lesion is retrieved by each participating algorithm. To penalize algorithms that may try to improve the evaluation results by providing many points per lesion, all other points retrieved are considered as false positives. However, annotators have looked at specific regions of the volume, which means that one cannot be sure that other regions are free of lesions. In other words, participating algorithms could detect lesions that were not annotated. To avoid penalizing such lesions, binary masks are used for each volume, which mask only those regions that were manually annotated. Retrieved points that lie outside the mask are not considered in the confusion matrix.

The evaluation of the Detection Benchmark takes place at three different levels:

1. Lesion level: For each annotated lesion, two values are measured, namely
   - Minimum Euclidean distance, $min(d_i)$: For each annotated lesion, the distance to the nearest point retrieved by the participating algorithm is measured as shown in Fig. 6.1. This distance is provided for each annotated lesion, regardless of whether the lesion is considered as detected or not.
   - Detection: A lesion is considered as detected if the point P_i, provided by the algorithm, is within the sphere centred on C_i and has the diameter given by the points $D1_i$ and $D2_i$. In particular, a radius of the sphere, r, is considered, which is equal to the distance between the centre C_i and the farthest of the points $D1_i$ and $D2_i$. That is, a lesion is detected iff $min(d) < r$. In Fig. 6.1, the lesion is detected by the point $P1_i$, but not by $P2_i$.

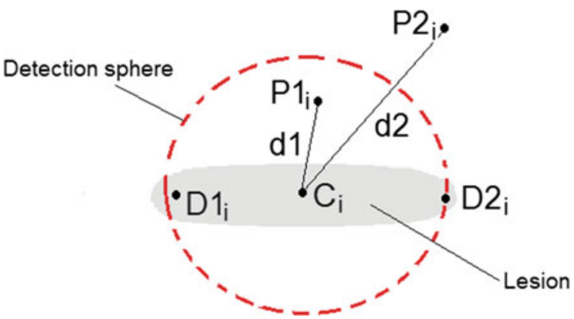


Fig. 6.1 Schematic representation of a lesion annotated by the centre C_i and the two diameter points $D1_i$ and $D2_i$. The points $P1_i$ and $P2_i$ are retrieved by participating algorithms. $P1_i$ lies within the detection sphere and is thus considered as detected in contrast to the point $P2_i$

2. Volume level: The confusion matrix (true positives, false positives, true negatives and false negatives) is calculated per volume, based on the detection values calculated at lesion level. From this confusion matrix, the precision (percentage of correctly detected lesions) and the recall (percentage of total lesions detected) are calculated for each volume and participating algorithm. As it is expected that algorithms provide exactly one point per lesion, all further points provided by an algorithm for the same lesion are considered as false positives.
3. Anatomical structure (organ) average level: To test whether the scores of lesion detection are generally dependent on the anatomical structure in which the lesions are, we calculate the score averages (the averages of the Euclidean distances between lesion centres and detection points) over each organ.

6.3 Analysis of Fuzzy Segmentation Metrics

Sometimes, medical volume segmentations are fuzzy. Such segmentations can be the result of averaging annotations done by different annotators. Fuzzy segmentation can also be the result of fusing automatic segmentations, which results in a silver corpus [2]. Depending on the approach used, automatic segmentations generated by segmentation algorithms can also be fuzzy. In contrast to binary segmentation, fuzzy segmentations are represented as memberships of voxels in classes (anatomical structures). Instead of a binary association, a voxel is rather associated with a class with a probability specifying the degree of membership to this class. Note that binary segmentation is just a special case of fuzzy segmentation, where the degree of memberships to a particular class can be either zero or one.

In this section, we analyse the impact of using fuzzy metrics in evaluating medical image segmentation. This is done by analysing the rankings produced by binary and fuzzy metrics of segmentations as well as segmentation algorithms. Segmentation ranking here means ordering segmentations according to their similarities to their corresponding ground truth segmentations. We analyse this from several sides trying to answer the following questions: (1) considering the case when the segmentations being evaluated/ranked are of mixed types (fuzzy and binary), which of the following two evaluation methods is to be used: (a) evaluating both types using fuzzy metrics based on the fact that binary segmentation is a special case of fuzzy segmentation, or (b) cutting fuzzy segmentations at a particular threshold and then using binary evaluation metrics? (2) The same question holds for the case when the ground truth segmentations and the segmentations being evaluated are of different types?

In the following, we define some notations and settings to be used in this section. Since binary segmentation is a special case of fuzzy segmentation, in which probabilities are either 0 or 1, this implies that fuzzy metrics can be used to compare the following combinations of segmentations, which we will denote as evaluation cases throughout this section:

Case i: binary segmentation evaluated against binary ground truth
Case ii: binary segmentation evaluated against fuzzy ground truth

Case iii: fuzzy segmentation evaluated against binary ground truth
Case iv: fuzzy segmentation evaluated against fuzzy ground truth

We define two types of evaluation that can be used for each of the evaluation cases above. The first type is **threshold evaluation**. Here, the ground truth segmentation, as well as the segmentation being evaluated, is cut at a threshold of 0.5 as a first step and then compared using an evaluation metric. The second type is **fuzzy evaluation** in which the segmentations are compared directly using fuzzy metrics.

The aim of this analysis is to infer how sensitive metrics are against image fuzzification. This analysis is motivated by the following: on the one hand, if there is fuzzy ground truth available and the segmentations being evaluated are fuzzy as well (Case iv), then metrics with high fuzzification sensitivity are required to distinguish the accuracy of the systems. On the other hand, when binary segmentations are to be compared with fuzzy ones (Case ii and Case iii), the question to be answered is, which type of evaluation (threshold evaluation or fuzzy evaluation) should be used?

In the Anatomy 1 and 2 Benchmarks, only binary ground truth segmentation has been used. Most of the participating algorithms provided binary segmentation, i.e. from Case i. However, only one of the participating algorithms produced fuzzy segmentations, i.e. Case iii. This algorithm is denoted as *Algorithm A* throughout this section. To complete the analysis, the other cases (Case ii and Case iv) and different types of segmentations are involved, which are described in the following:

- Binary ground truth (BGT): This is the official binary ground truth, used for validating the challenge.
- Synthetic fuzzy ground truth (FGT): Since there are only binary ground truth segmentations available, the fuzzy ground truth was generated synthetically: from each of the ground truth segmentations, a fuzzy variant was produced by smoothing the corresponding ground truth using a mean filter.
- Fuzzy silver ground truth (FSGT): In another variant, a fuzzy silver corpus is generated by fusing all the automatic segmentations.
- Binary silver ground truth (BSGT) [2]: The silver corpus was generated by fusing all the automatic segmentations and then cutting them at threshold 0.5, i.e. BSGT is FSGT cut at 0.5.
- Fuzzy automatic segmentation (FAS): These are the fuzzy segmentations produced by one of the participating algorithms, namely Algorithm A.
- Binary automatic segmentations (BAS): These are the automatic segmentations produced by all of the participating algorithms except Algorithm A.

Metrics considered in this analysis are those metrics in Table 6.1 that have fuzzy implementation (column "Fuzzy"). More about the fuzzy implementation of the metrics is available in [4].

In the remainder of this section, two experiments regarding fuzzy metrics are presented. In Sect. 6.3.1, the sensitivity of metrics to fuzzification is investigated by considering for each metric the discrepancy of similarities measured in two cases: the first is when binary segmentations are compared, and the second is when fuzzy representations of the same segmentation are compared. In Sect. 6.3.2, the impact of comparing segmentations of different types (fuzzy and binary) on the evaluation results is investigated, e.g. it is tested whether using binary ground truth to validate the fuzzy segmentation using fuzzy metrics has a negative impact on the evaluation result compared with using a binary representation of the segmentations by cutting them at a threshold of 0.5 as a prior step.

6.3.1 Metric Sensitivity Against Fuzzification

The aim of this experiment is to infer how invariant metrics are against fuzzification of images. To this end, we compare each binary volume in the silver corpus (BSGT) with its corresponding volume from the fuzzy silver corpus (FSGT) using each of the 16 metrics for which fuzzy implementations exist. This results in 16 metric values (similarities and distances) per comparison (segmentation pair), which are then averaged over all pairs to get 16 average metric values, presented in Fig. 6.2. The assumption is that metrics that measure less average discrepancy between the binary volumes and their fuzzy variants are more invariant against fuzzification.

Results in Fig. 6.2 show that metrics are differently invariant against fuzzification, that is, they have different capabilities in discovering changes due to fuzzification. Metrics that include the true negatives (TN) in their definitions (e.g. ARI, ACU and TNR) are in general less sensitive to fuzzification, in contrast to other metrics not considering the TN, such as DICE, KAP and JAC. Also, one can observe that the discrepancy metrics FPR, PBD and VOI are also invariant against fuzzification because they provide very small distances ($<<$ 0.02 voxel) between binary images and their corresponding smoothed images.

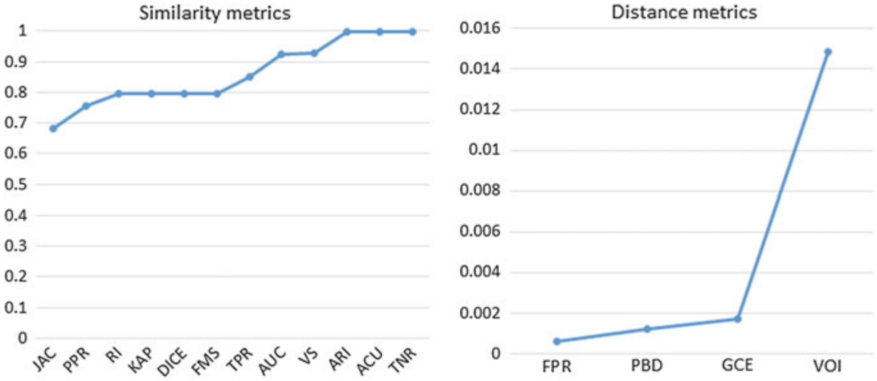


Fig. 6.2 The average similarity between binary volumes and their corresponding fuzzy variant

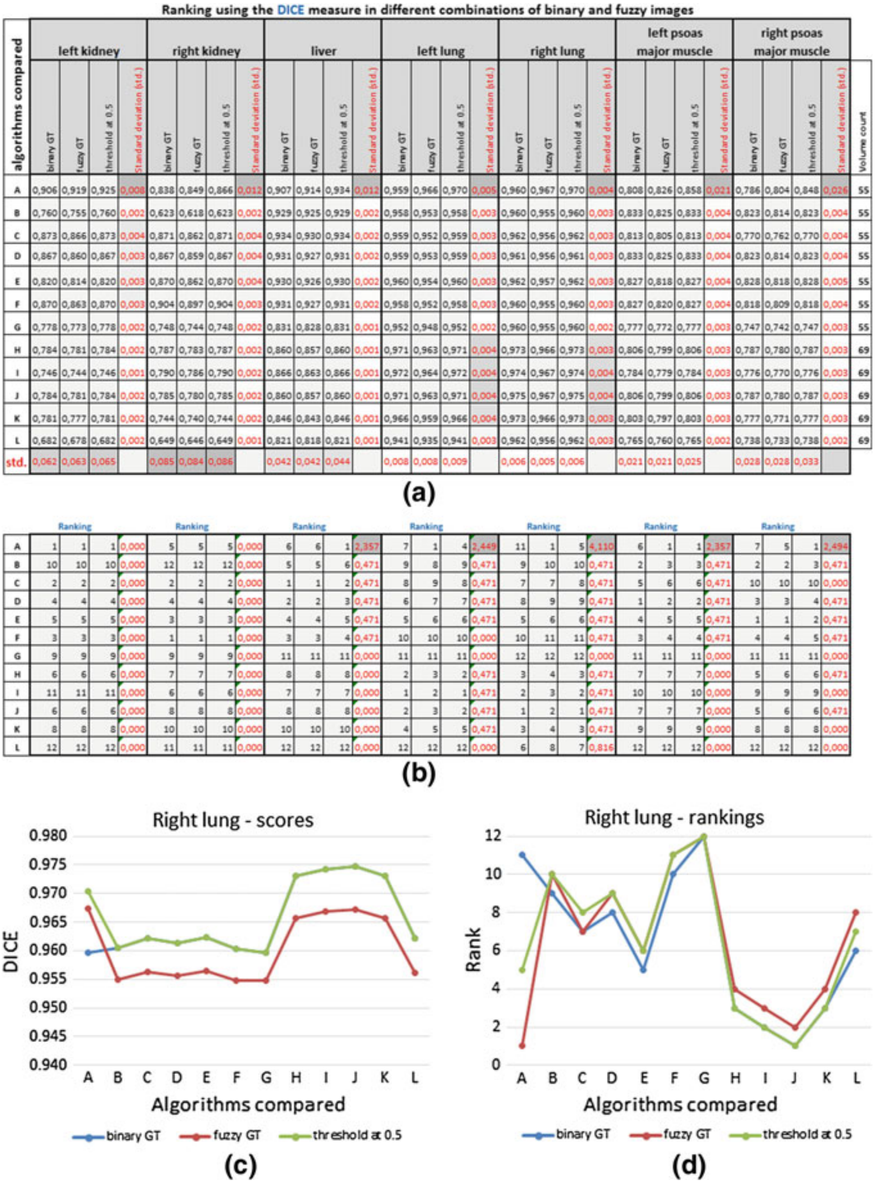
6.3.2 *Ranking Systems Using Binary/Fuzzy Ground Truth*

The aim of this experiment is to infer how system rankings, using metrics, change when using fuzzy instead of binary ground truth in two cases: when the segmentations being evaluated are binary (Case i and Case ii) and when they are fuzzy (Case iii and Case iv). The segmentations that were used in this experiment are BGT, FGT, BAS and FAS. Figures 6.3, 6.4 and 6.5 show the results of the experiment performed for three selected metrics, namely Dice coefficient (DICE), interclass correlation (ICC) and adjusted Rand index (ARI), respectively. The three metrics are selected to represent three different metric categories in Table 6.1 for which fuzzy implementations exist. There are seven systems (Systems A to L) to be ranked according to their performance, which is measured by average quality of the segmentations produced by these systems, i.e. the metric values resulting from comparing these segmentations with the corresponding ground truth. The averages are built separately for each of the seven organs (left kidney, right kidney, liver, left lung, right lung, left psoas major muscle and right psoas major muscle), which means the systems are ranked for each organ separately. The participating algorithms B to L produce only binary volumes, whereas Algorithm A produces only fuzzy segmentations.

The ranking is performed in three different configurations: in the first, which we denote by “binary GT”, the ground truth is binary (BGT) and the segmentations are unchanged (fuzzy for Algorithm A and binary otherwise). This covers Case i and Case iii. In the second configuration, which we denote by “fuzzy GT”, the ground truth is fuzzy (FGT) and the segmentations are unchanged. This covers Case ii and Case iv. In the third configuration, denoted by “threshold at 0.5”, the fuzzy segmentations of Algorithm A are cut at a 0.5 threshold to get binary representations. The other binary segmentations and the ground truth are unchanged; thus, all images involved in this case are binary. In the first and second configurations, fuzzy evaluation metrics are used, whereas in the third configuration, binary evaluation metrics (threshold evaluation) are used.

In the figures, we included standard deviation columns and a standard deviation row to indicate the discrepancy (deviation) between the algorithms as well as between the three cases.

The first observation is regarding Algorithm A, which produces fuzzy segmentations. Here, Algorithm A has the best ranking when the corresponding segmentations are evaluated using a 0.5 threshold or against a fuzzy ground truth, but it has a considerable disadvantage when using the binary ground truth. Thus, it is strongly recommended to use a threshold option when the segmentations/ground truth is mixed in terms of binary and fuzzy modes. The second observation is that the sensitivity in the resulting rankings is dependent on the deviations between the average scores of the systems; the lower the deviation, the more the rankings change between the three cases. That is, if the algorithms are similar in their performance, then using a binary instead of a fuzzy ground truth, or the opposite, has a considerable impact on the system ranking. For example, the average scores of the systems have the highest deviation with kidney and liver, so the rankings of the systems are exactly the same



Ranking using the DICE measure in different combinations of binary and fuzzy images

algorithms compared	left kidney				right kidney				liver				left lung				right lung				left psoas major muscle				right psoas major muscle				
	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	Volume count
A	0,906	0,919	0,925	0,008	0,838	0,849	0,866	0,012	0,907	0,914	0,934	0,012	0,959	0,966	0,970	0,005	0,960	0,967	0,970	0,004	0,808	0,826	0,858	0,021	0,786	0,804	0,848	0,026	55
B	0,760	0,755	0,760	0,002	0,623	0,618	0,623	0,002	0,929	0,925	0,929	0,002	0,958	0,953	0,958	0,003	0,960	0,955	0,960	0,003	0,833	0,825	0,833	0,004	0,823	0,814	0,823	0,004	55
C	0,873	0,866	0,873	0,004	0,871	0,862	0,871	0,004	0,934	0,930	0,934	0,002	0,959	0,952	0,959	0,003	0,962	0,956	0,962	0,003	0,813	0,805	0,813	0,004	0,770	0,762	0,770	0,004	55
D	0,867	0,860	0,867	0,003	0,867	0,859	0,867	0,004	0,931	0,927	0,931	0,002	0,959	0,953	0,959	0,003	0,961	0,956	0,961	0,003	0,833	0,825	0,833	0,004	0,823	0,814	0,823	0,004	55
E	0,820	0,814	0,820	0,003	0,870	0,862	0,870	0,004	0,930	0,926	0,930	0,002	0,960	0,954	0,960	0,003	0,962	0,957	0,962	0,003	0,827	0,818	0,827	0,004	0,828	0,818	0,828	0,005	55
F	0,870	0,863	0,870	0,003	0,904	0,897	0,904	0,003	0,931	0,927	0,931	0,002	0,958	0,952	0,958	0,003	0,960	0,955	0,960	0,003	0,827	0,820	0,827	0,004	0,818	0,809	0,818	0,004	55
G	0,778	0,773	0,778	0,002	0,748	0,744	0,748	0,002	0,831	0,828	0,831	0,001	0,952	0,948	0,952	0,002	0,960	0,955	0,960	0,002	0,777	0,772	0,777	0,003	0,747	0,742	0,747	0,003	55
H	0,784	0,781	0,784	0,002	0,787	0,783	0,787	0,002	0,860	0,857	0,860	0,001	0,971	0,963	0,971	0,004	0,973	0,966	0,973	0,003	0,806	0,799	0,806	0,003	0,787	0,780	0,787	0,003	69
I	0,746	0,744	0,746	0,001	0,790	0,786	0,790	0,002	0,866	0,863	0,866	0,001	0,972	0,964	0,972	0,004	0,974	0,967	0,974	0,004	0,784	0,779	0,784	0,003	0,776	0,770	0,776	0,003	69
J	0,784	0,781	0,784	0,002	0,785	0,780	0,785	0,002	0,860	0,857	0,860	0,001	0,971	0,963	0,971	0,004	0,975	0,967	0,975	0,004	0,806	0,799	0,806	0,003	0,787	0,780	0,787	0,003	69
K	0,781	0,777	0,781	0,002	0,744	0,740	0,744	0,002	0,846	0,843	0,846	0,001	0,966	0,959	0,966	0,004	0,973	0,966	0,973	0,003	0,803	0,797	0,803	0,003	0,777	0,771	0,777	0,003	69
L	0,682	0,678	0,682	0,002	0,649	0,646	0,649	0,001	0,821	0,818	0,821	0,001	0,941	0,935	0,941	0,003	0,962	0,956	0,962	0,003	0,765	0,760	0,765	0,002	0,738	0,733	0,738	0,002	69
std.	0,062	0,063	0,065		0,085	0,084	0,086		0,042	0,042	0,044		0,008	0,008	0,009		0,006	0,005	0,006		0,021	0,021	0,025		0,028	0,028	0,033		

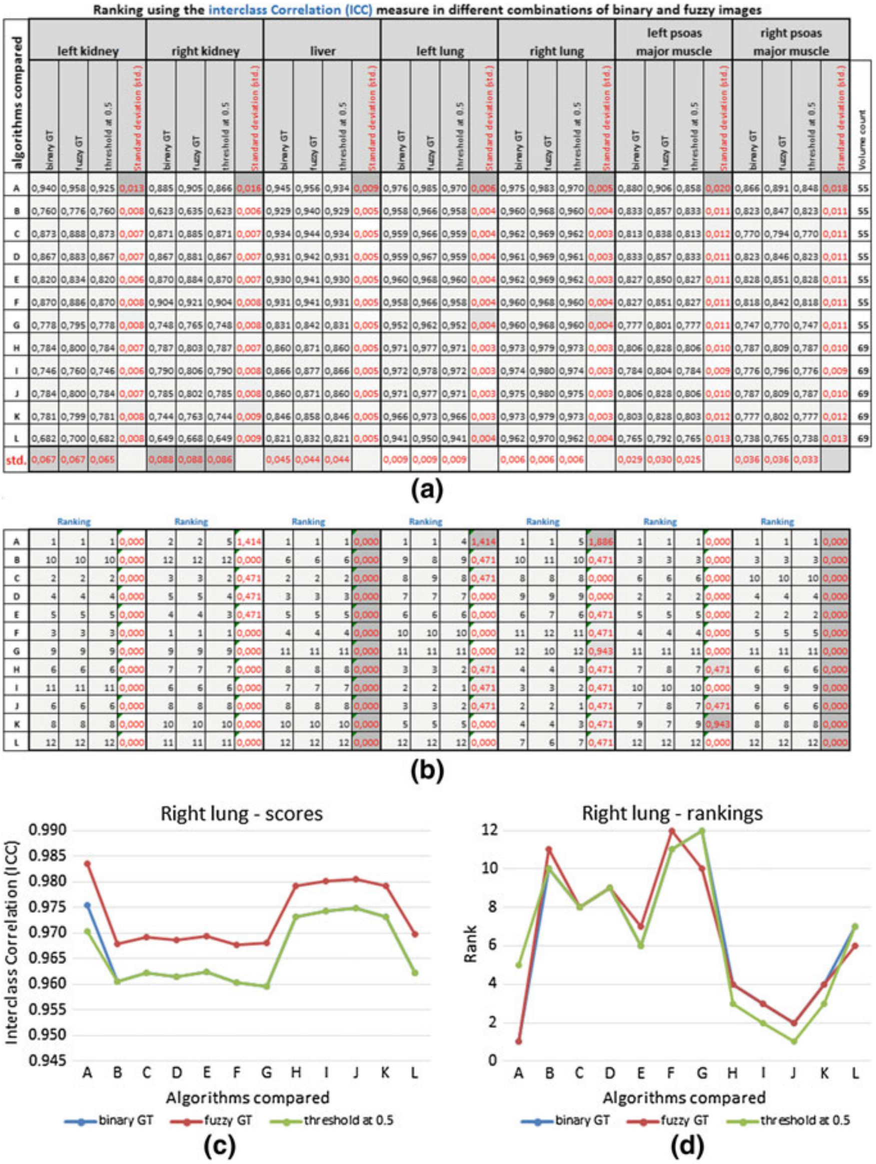
(a)

	Ranking				Ranking				Ranking				Ranking				Ranking				Ranking				Ranking			
A	1	1	1	0,000	5	5	5	0,000	6	6	1	2,357	7	1	4	2,449	11	1	5	4,110	6	1	1	2,357	7	5	1	2,494
B	10	10	10	0,000	12	12	12	0,000	5	5	6	0,471	9	8	9	0,471	9	10	10	0,471	2	3	3	0,471	2	2	3	0,471
C	2	2	2	0,000	2	2	2	0,000	1	1	2	0,471	8	9	8	0,471	7	7	8	0,471	5	6	6	0,471	10	10	10	0,000
D	4	4	4	0,000	4	4	4	0,000	2	2	3	0,471	6	7	7	0,471	8	9	9	0,471	1	2	2	0,471	3	3	4	0,471
E	5	5	5	0,000	3	3	3	0,000	4	4	5	0,471	5	6	6	0,471	5	6	6	0,471	4	5	5	0,471	1	1	2	0,471
F	3	3	3	0,000	1	1	1	0,000	3	3	4	0,471	10	10	10	0,000	10	11	11	0,471	3	4	4	0,471	4	4	5	0,471
G	9	9	9	0,000	9	9	9	0,000	11	11	11	0,000	11	11	11	0,000	12	12	12	0,000	11	11	11	0,000	11	11	11	0,000
H	6	6	6	0,000	7	7	7	0,000	8	8	8	0,000	2	3	2	0,471	3	4	3	0,471	7	7	7	0,000	5	6	6	0,471
I	11	11	11	0,000	6	6	6	0,000	7	7	7	0,000	1	2	1	0,471	2	3	2	0,471	10	10	10	0,000	9	9	9	0,000
J	6	6	6	0,000	8	8	8	0,000	8	8	8	0,000	2	3	2	0,471	1	2	1	0,471	7	7	7	0,000	5	6	6	0,471
K	8	8	8	0,000	10	10	10	0,000	10	10	10	0,000	4	5	5	0,471	3	4	3	0,471	9	9	9	0,000	8	8	8	0,000
L	12	12	12	0,000	11	11	11	0,000	12	12	12	0,000	12	12	12	0,000	6	8	7	0,816	12	12	12	0,000	12	12	12	0,000

(b)

(c) (d)

Fig. 6.3 **a** Validating segmentations using the DICE in three different combinations of binary/fuzzy segmentations. The standard deviations of the scores are to show the quality variance between the algorithms and the score variance between the combinations. **b** The resulting system ranking. **c** Score details of the right lung as a selected case. **d** The resulting system ranking for the right lung



Ranking using the interclass Correlation (ICC) measure in different combinations of binary and fuzzy images

algorithms compared	left kidney				right kidney				liver				left lung				right lung				left psoas major muscle				right psoas major muscle				
	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	Volume count
A	0,940	0,958	0,925	0,013	0,885	0,905	0,866	0,016	0,945	0,956	0,934	0,009	0,976	0,985	0,970	0,006	0,975	0,983	0,970	0,005	0,880	0,906	0,858	0,020	0,866	0,891	0,848	0,018	55
B	0,760	0,776	0,760	0,008	0,623	0,635	0,623	0,006	0,929	0,940	0,929	0,005	0,958	0,966	0,958	0,004	0,960	0,968	0,960	0,004	0,833	0,857	0,833	0,011	0,823	0,847	0,823	0,011	55
C	0,873	0,888	0,873	0,007	0,871	0,885	0,871	0,007	0,934	0,944	0,934	0,005	0,959	0,966	0,959	0,004	0,962	0,969	0,962	0,003	0,813	0,838	0,813	0,012	0,770	0,794	0,770	0,011	55
D	0,867	0,883	0,867	0,007	0,867	0,881	0,867	0,007	0,931	0,942	0,931	0,005	0,959	0,967	0,959	0,004	0,961	0,969	0,961	0,003	0,833	0,857	0,833	0,011	0,823	0,846	0,823	0,011	55
E	0,820	0,834	0,820	0,006	0,870	0,884	0,870	0,007	0,930	0,941	0,930	0,005	0,960	0,968	0,960	0,004	0,962	0,969	0,962	0,003	0,827	0,850	0,827	0,011	0,828	0,851	0,828	0,011	55
F	0,870	0,886	0,870	0,008	0,904	0,921	0,904	0,008	0,931	0,941	0,931	0,005	0,958	0,966	0,958	0,004	0,960	0,968	0,960	0,004	0,827	0,851	0,827	0,011	0,818	0,842	0,818	0,011	55
G	0,778	0,795	0,778	0,008	0,748	0,765	0,748	0,008	0,831	0,842	0,831	0,005	0,952	0,962	0,952	0,004	0,960	0,968	0,960	0,004	0,777	0,801	0,777	0,011	0,747	0,770	0,747	0,011	55
H	0,784	0,800	0,784	0,007	0,787	0,803	0,787	0,007	0,860	0,871	0,860	0,005	0,971	0,977	0,971	0,003	0,973	0,979	0,973	0,003	0,806	0,828	0,806	0,010	0,787	0,809	0,787	0,010	69
I	0,746	0,760	0,746	0,006	0,790	0,806	0,790	0,008	0,866	0,877	0,866	0,005	0,972	0,978	0,972	0,003	0,974	0,980	0,974	0,003	0,784	0,804	0,784	0,009	0,776	0,796	0,776	0,009	69
J	0,784	0,800	0,784	0,007	0,785	0,802	0,785	0,008	0,860	0,871	0,860	0,005	0,971	0,977	0,971	0,003	0,975	0,980	0,975	0,003	0,806	0,828	0,806	0,010	0,787	0,809	0,787	0,010	69
K	0,781	0,799	0,781	0,008	0,744	0,763	0,744	0,009	0,846	0,858	0,846	0,005	0,966	0,973	0,966	0,003	0,973	0,979	0,973	0,003	0,803	0,828	0,803	0,012	0,777	0,802	0,777	0,012	69
L	0,682	0,700	0,682	0,008	0,649	0,668	0,649	0,009	0,821	0,832	0,821	0,005	0,941	0,950	0,941	0,004	0,962	0,970	0,962	0,004	0,765	0,792	0,765	0,013	0,738	0,765	0,738	0,013	69
std.	0,067	0,067	0,065		0,088	0,088	0,086		0,045	0,044	0,044		0,009	0,009	0,009		0,006	0,006	0,006		0,029	0,030	0,025		0,036	0,036	0,033		

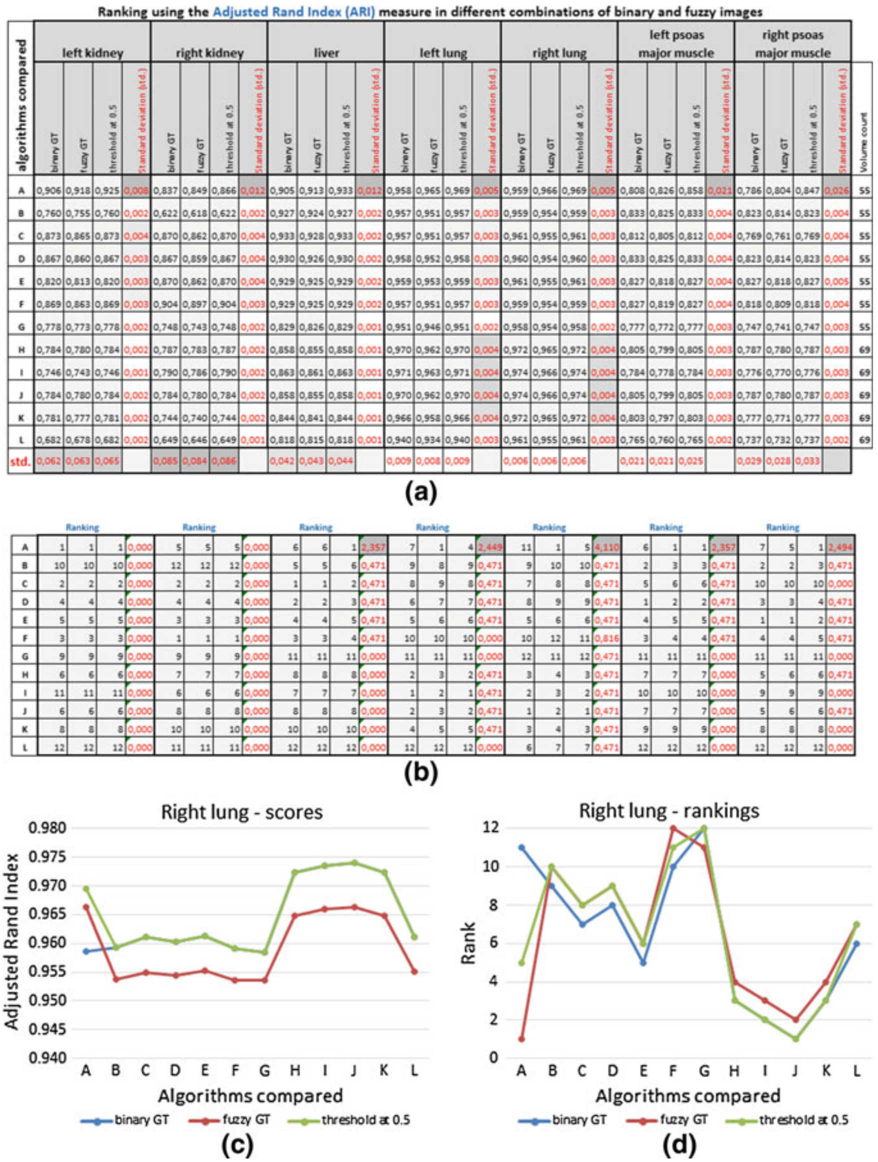
(a)

	Ranking				Ranking				Ranking				Ranking				Ranking				Ranking				Ranking			
A	1	1	1	0,000	2	2	5	1,414	1	1	1	0,000	1	1	4	1,414	1	1	5	1,886	1	1	1	0,000	1	1	1	0,000
B	10	10	10	0,000	12	12	12	0,000	6	6	6	0,000	9	8	9	0,471	10	11	10	0,471	3	3	3	0,000	3	3	3	0,000
C	2	2	2	0,000	3	3	2	0,471	2	2	2	0,000	8	9	8	0,471	8	8	8	0,000	6	6	6	0,000	10	10	10	0,000
D	4	4	4	0,000	5	5	4	0,471	3	3	3	0,000	7	7	7	0,000	9	9	9	0,000	2	2	2	0,000	4	4	4	0,000
E	5	5	5	0,000	4	4	3	0,471	5	5	5	0,000	6	6	6	0,000	6	7	6	0,471	5	5	5	0,000	2	2	2	0,000
F	3	3	3	0,000	1	1	1	0,000	4	4	4	0,000	10	10	10	0,000	11	12	11	0,471	4	4	4	0,000	5	5	5	0,000
G	9	9	9	0,000	9	9	9	0,000	11	11	11	0,000	11	11	11	0,000	12	10	12	0,943	11	11	11	0,000	11	11	11	0,000
H	6	6	6	0,000	7	7	7	0,000	8	8	8	0,000	3	3	2	0,471	4	4	3	0,471	7	8	7	0,471	6	6	6	0,000
I	11	11	11	0,000	6	6	6	0,000	7	7	7	0,000	2	2	1	0,471	3	3	2	0,471	10	10	10	0,000	9	9	9	0,000
J	6	6	6	0,000	8	8	8	0,000	8	8	8	0,000	3	3	2	0,471	2	2	1	0,471	7	8	7	0,471	6	6	6	0,000
K	8	8	8	0,000	10	10	10	0,000	10	10	10	0,000	5	5	5	0,000	4	4	3	0,471	9	7	9	0,943	8	8	8	0,000
L	12	12	12	0,000	11	11	11	0,000	12	12	12	0,000	12	12	12	0,000	7	6	7	0,471	12	12	12	0,000	12	12	12	0,000

(b)

Fig. 6.4 The results of the same experiment as in Fig. 6.3, but performed using the interclass correlation (ICC) as an evaluation metric

in the three cases. On the contrary, system average scores have low deviations with lungs and psoas major muscles; therefore, the rankings of the systems considerably change between the three cases. We recommend therefore to take the score deviations into account when there are mixed fuzzy and binary segmentations/ground truth.



Ranking using the Adjusted Rand Index (ARI) measure in different combinations of binary and fuzzy images

algorithms compared	left kidney				right kidney				liver				left lung				right lung				left psoas major muscle				right psoas major muscle				
	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	binary GT	fuzzy GT	threshold at 0.5	Standard deviation (std.)	Volume count
A	0,906	0,918	0,925	0,008	0,837	0,849	0,866	0,012	0,905	0,913	0,933	0,012	0,958	0,965	0,969	0,005	0,959	0,966	0,969	0,005	0,808	0,826	0,858	0,021	0,786	0,804	0,847	0,026	55
B	0,760	0,755	0,760	0,002	0,622	0,618	0,622	0,002	0,927	0,924	0,927	0,002	0,957	0,951	0,957	0,003	0,959	0,954	0,959	0,003	0,833	0,825	0,833	0,004	0,823	0,814	0,823	0,004	55
C	0,873	0,865	0,873	0,004	0,870	0,862	0,870	0,004	0,933	0,928	0,933	0,002	0,957	0,951	0,957	0,003	0,961	0,955	0,961	0,003	0,812	0,805	0,812	0,004	0,769	0,761	0,769	0,004	55
D	0,867	0,860	0,867	0,003	0,867	0,859	0,867	0,004	0,930	0,926	0,930	0,002	0,958	0,952	0,958	0,003	0,960	0,954	0,960	0,003	0,833	0,825	0,833	0,004	0,823	0,814	0,823	0,004	55
E	0,820	0,813	0,820	0,003	0,870	0,862	0,870	0,004	0,929	0,925	0,929	0,002	0,959	0,953	0,959	0,003	0,961	0,955	0,961	0,003	0,827	0,818	0,827	0,004	0,827	0,818	0,827	0,005	55
F	0,869	0,863	0,869	0,003	0,904	0,897	0,904	0,003	0,929	0,925	0,929	0,002	0,957	0,951	0,957	0,003	0,959	0,954	0,959	0,003	0,827	0,819	0,827	0,004	0,818	0,809	0,818	0,004	55
G	0,778	0,773	0,778	0,002	0,748	0,743	0,748	0,002	0,829	0,826	0,829	0,001	0,951	0,946	0,951	0,002	0,958	0,954	0,958	0,002	0,777	0,772	0,777	0,003	0,747	0,741	0,747	0,003	55
H	0,784	0,780	0,784	0,002	0,787	0,783	0,787	0,002	0,858	0,855	0,858	0,001	0,970	0,962	0,970	0,004	0,972	0,965	0,972	0,004	0,805	0,799	0,805	0,003	0,787	0,780	0,787	0,003	69
I	0,746	0,743	0,746	0,001	0,790	0,786	0,790	0,002	0,863	0,861	0,863	0,001	0,971	0,963	0,971	0,004	0,974	0,966	0,974	0,004	0,784	0,778	0,784	0,003	0,776	0,770	0,776	0,003	69
J	0,784	0,780	0,784	0,002	0,784	0,780	0,784	0,002	0,858	0,855	0,858	0,001	0,970	0,962	0,970	0,004	0,974	0,966	0,974	0,004	0,805	0,799	0,805	0,003	0,787	0,780	0,787	0,003	69
K	0,781	0,777	0,781	0,002	0,744	0,740	0,744	0,002	0,844	0,841	0,844	0,001	0,966	0,958	0,966	0,004	0,972	0,965	0,972	0,004	0,803	0,797	0,803	0,003	0,777	0,771	0,777	0,003	69
L	0,682	0,678	0,682	0,002	0,649	0,646	0,649	0,001	0,818	0,815	0,818	0,001	0,940	0,934	0,940	0,003	0,961	0,955	0,961	0,003	0,765	0,760	0,765	0,002	0,737	0,732	0,737	0,002	69
std.	0,062	0,063	0,065		0,085	0,084	0,086		0,042	0,043	0,044		0,009	0,008	0,009		0,006	0,006	0,006		0,021	0,021	0,025		0,029	0,028	0,033		

(a)

	Ranking				Ranking				Ranking				Ranking				Ranking				Ranking				Ranking			
A	1	1	1	0,000	5	5	5	0,000	6	6	1	2,357	7	1	4	2,449	11	1	5	4,110	6	1	1	2,357	7	5	1	2,494
B	10	10	10	0,000	12	12	12	0,000	5	5	6	0,471	9	8	9	0,471	9	10	10	0,471	2	3	3	0,471	2	2	3	0,471
C	2	2	2	0,000	2	2	2	0,000	1	1	2	0,471	8	9	8	0,471	7	8	8	0,471	5	6	6	0,471	10	10	10	0,000
D	4	4	4	0,000	4	4	4	0,000	2	2	3	0,471	6	7	7	0,471	8	9	9	0,471	1	2	2	0,471	3	3	4	0,471
E	5	5	5	0,000	3	3	3	0,000	4	4	5	0,471	5	6	6	0,471	5	6	6	0,471	4	5	5	0,471	1	1	2	0,471
F	3	3	3	0,000	1	1	1	0,000	3	3	4	0,471	10	10	10	0,000	10	12	11	0,816	3	4	4	0,471	4	4	5	0,471
G	9	9	9	0,000	9	9	9	0,000	11	11	11	0,000	11	11	11	0,000	12	11	12	0,471	11	11	11	0,000	11	11	11	0,000
H	6	6	6	0,000	7	7	7	0,000	8	8	8	0,000	2	3	2	0,471	3	4	3	0,471	7	7	7	0,000	5	6	6	0,471
I	11	11	11	0,000	6	6	6	0,000	7	7	7	0,000	1	2	1	0,471	2	3	2	0,471	10	10	10	0,000	9	9	9	0,000
J	6	6	6	0,000	8	8	8	0,000	8	8	8	0,000	2	3	2	0,471	1	2	1	0,471	7	7	7	0,000	5	6	6	0,471
K	8	8	8	0,000	10	10	10	0,000	10	10	10	0,000	4	5	5	0,471	3	4	3	0,471	9	9	9	0,000	8	8	8	0,000
L	12	12	12	0,000	11	11	11	0,000	12	12	12	0,000	12	12	12	0,000	6	7	7	0,471	12	12	12	0,000	12	12	12	0,000

(b)

Fig. 6.5 The results of the same experiment as in Fig. 6.3, but performed using the adjusted Rand index (ARI) as an evaluation metric

6.4 Analysis of Metrics Using Manual Rankings

In this section, we provide an analysis of the metrics based on the two manual rankings of segmentations, done by two medical experts. Manual rankings provide a reference for judging metrics and evaluation methods. That is, when evaluating segmentations by comparing them with the corresponding ground truth using distance or similarity metrics, one gets scores denoting how similar or different the segmentations are from the ground truth. However, since different metrics provide different scores, which produce different rankings, the aim of this analysis is to find the metric(s) with the highest correlation with the manual rankings. Another aim of this analysis is to validate the selection of the subset of four metrics from Table 6.1 used for the evaluation of medical image segmentation in the VISCERAL Anatomy 1, 2 and 3 Benchmarks.

In Sect. 6.4.1, we describe the dataset that has been manually ranked and the ranking methodology used. We then analyse the correlation between the manual ranking and the rankings produced by metrics: in Sect. 6.4.2, the ranking is done at segmentation level, while in Sect. 6.4.3, the ranking is done at system level. Finally, we discuss the results of the manual ranking analysis in Sect. 6.4.4.

6.4.1 Dataset

To provide a manual ranking, 483 segmentations were selected by medical experts from the output of the Anatomy 2 Benchmark participant algorithms. This segmentation set has the following properties:

- The segmentations correspond to six organs/structures, namely liver, pancreas, urinary bladder, aorta, left lung and right kidney. These structures were selected by medical experts so that they cover different sizes, shapes and boundary complexities.
- The segmentations correspond to 110 different volumes each representing a medical case, where a medical case is defined as an anatomical structure in a particular ground truth volume (e.g. the liver in each ground truth is considered a different medical case).
- The segmentations were produced by seven participating algorithms. However, different volumes (medical cases) were segmented by different numbers of algorithms. This means that for some volumes, seven segmentations are available, but for other volumes, there are fewer than seven. For the ranking analysis, only those volumes were considered for which at least three segmentations are available. These are only 92 volumes.

The segmentations described above have been ranked by two different radiologists separately, resulting in two different rankings, which we call Manual Ranking 1

Table 6.2 Criteria for the subjective scoring system used for manual ranking

Score	Ranking criteria
1	Severe deviation to other organs, no connection with expected organ segmentation
2	Evident crossing of organ border, organ parts missing from segmentation
3	Irregular segmentation with respect to manual segmentation guidelines
4	Minor deviations from segmentation guidelines
5	Optimal segmentation, organ borders and adherence to segmentation guidelines

(MRK1) and Manual Ranking 2 (MRK2). The ranking was performed in a double-blind way. The ranking criteria in Table 6.2 have been considered.

In each of the manual rankings, all segmentations corresponding to the same medical case were considered as one group, within which these segmentations are ranked using the criteria in Table 6.2.

Note that according to this ranking system, different segmentations may have the same rank. For example, it is common with manual ranking that five segmentations are ranked with 1, 2, 2, 2, 3, which is not common in case of ranking based on metric values except if the metric values are discretized.

In order to test how the two manual rankers agree, the Pearson correlation coefficient between the two manual rankings was measured. The correlation between the manual rankings, RNK1 and RNK1, is 0.62. This is a moderate correlation, which means that there is a non-negligible discrepancy between the manual rankings.

6.4.2 Manual Versus Metric Rankings at Segmentation Level

We analyse the correlation between rankings of groups of segmentations produced by each of the metrics in Table 6.1 and rankings of the same segmentations based on the manual rankings (MRK1 and MRK2). This analysis is to infer which metrics have the most correlation with the manual ranking.

The rankings in this experiment are at segmentation level, which means that individual segmentations corresponding to the same medical case are ranked. To this end, the segmentations were grouped so that each group consists of a medical case and the corresponding segmentations. The segmentations in each group are then ranked using each of the metrics by comparing each of the segmentations with its corresponding ground truth. The segmentation with the lowest match is given the lowest rank, and the best match is given the highest rank. This is in order to get a ranking that is comparable with the manual ranking.

Table 6.3 shows the correlations between each of the metrics presented in Table 6.1 and each of the manual rankings, RNK1 and RNK2. The metrics are sorted according to the correlation with RNK1. Note that the highest correlation value (0.64) is a moderate correlation, and many of the metrics have weak correlation. This is

Table 6.3 Pearson correlation coefficient (CORR.) between each of the metrics presented in Table 6.1 and the manual rankings MRK1 and MRK2 at segmentation level. The metrics are sorted according to the decreasing correlation

Manual Ranking 1 (MRNK 1)			Manual Ranking 2 (MRNK 2)		
Metric		CORR.	Metric		CORR.
Average distance	AVD	0.57	Rand index	RI	0.56
Adjusted Rand index	ARI	0.54	Variation of information	VOI	0.56
Dice	DICE	0.54	Average distance	AVD	0.56
F-measure	FMS	0.54	Accuracy	ACU	0.56
Interclass correlation	ICC	0.54	Global consistency error	GCE	0.55
Cohen's kappa	KAP	0.54	Adjusted Rand index	ARI	0.52
Probabilistic distance	PBD	0.54	Dice	DICE	0.52
Rand index	RI	0.54	F-measure	FMS	0.52
Jaccard index	JAC	0.54	Interclass correlation	ICC	0.52
Accuracy	ACU	0.53	Cohen's kappa	KAP	0.52
Variation of information	VOI	0.53	Jaccard index	JAC	0.52
Global consistency error	GCE	0.53	Probabilistic distance	PBD	0.51
Mutual information	MI	0.47	Mutual information	MI	0.46
Mahalanobis distance	MHD	0.44	Mahalanobis distance	MHD	0.41
Hausdorff distance	HD	0.43	Hausdorff distance	HD	0.40
Area under ROC curve	AUC	0.39	Positive predictive value	PPR	0.38
True-positive rate (sensitivity)	TPR	0.39	Area under ROC curve	AUC	0.36
Volumetric similarity	VS	0.27	True-positive rate (sensitivity)	TPR	0.36
Positive predictive value	PPR	0.27	Volumetric similarity	VS	0.30
Fallout	FPR	0.17	Fallout	FPR	0.26
True-negative rate (specificity)	TNR	0.17	True-negative rate (specificity)	TNR	0.26

expected, since ranking at segmentation level using the metrics considers very small changes, which do not necessarily reflect an improvement, e.g. differences caused by chance. For this reason, we provide another correlation analysis at system level, in Sect. 6.4.3, that uses significance testing to decide whether one system has better performance than another.

6.4.3 Manual Versus Metric Rankings at System Level

In this experiment, the evaluation metrics in Table 6.1 are validated by considering the system (algorithm) rankings produced by these metrics. In contrast to the ranking at segmentation level in Sect. 6.4.2, here the systems are ranked based on averages of the metrics of segmentations produced by these systems. In particular, for each metric, (i) we build a system ranking by comparing metric values of the segmentations produced by the systems using significance testing, and (ii) we calculate the correlation between this ranking and a system ranking based on the manual ranks. The resulting correlation for each metric is used as a quality measure of the metric, i.e. the best metrics are those having the highest correlation with the manual ranking. In the remainder of this section, the experiment is described and discussed in detail.

Validating a particular metric using a manual ranking goes in the following steps: Separately for each organ, the average of the metric values for each system is calculated, i.e. the metric values of all segmentations corresponding to a particular organ and produced by a particular system are averaged. We denote the resulting average by the system score for the organ considered. This system score is used to build a system ranking as discussed below. Note that although each organ is considered separately, it is different from the experiment in Sect. 6.4.2 (at segmentation level) because here we are averaging the metric values of more than one medical case, all of them corresponding to the same organ, but in different volumes.

Based on these system scores, the systems are ranked using a significance test (the sign test) to ensure that the difference between the system scores is significant. To this end, the systems are sorted according to their average scores ascending. Then, the ranks are given as follows: starting with the first system S_1 having the lowest system score, it is given the rank 1. Then, for each next system S_i, if there is a significant difference to the previous system S_{i-1}, according to a sign test, then S_i is assigned the next rank; otherwise, it is assigned the same rank as S_{i-1}.

Now, we want to judge the resulting ranking using each of the manual rankings as ground truth. However, the manual rankings available are at segmentation level. Therefore, the manual ranks are averaged analogously over all segmentations produced by a particular system corresponding to the organ considered. The resulting averages of the manual ranks are used to build a ground truth system ranking using the same method as with the metric ranking (i.e. significance sign test). Now, the correlation between the two rankings (system ranking based on the metrics and system ranking based on the manual ranks) is calculated. Since each organ is considered separately, we get a correlation value per organ for each metric, which are averaged to get the overall correlation of the metric.

Table 6.4 shows, for each metric, the overall correlation (correlation averaged over all organs). The same experiment is performed separately for each of the manual rankings (MNRK 1 and MNRK 2).

6.4.4 Discussion of the Manual Ranking Analysis

The following conclusions can be inferred from the results of the analysis using the manual rankings (results presented in Tables 6.3 and 6.4).

Table 6.4 Pearson correlation coefficient between each of the metrics presented in Table 6.1 and the manual rankings MRK1 and MRK2 at system level. The metrics are sorted according to the decreasing correlation

Manual Ranking 1 (MNRK 1)			Manual Ranking 2 (MNRK 2)		
Metric		CORR.	Metric		CORR.
Volumetric similarity	VS	0.81	Mahalanobis distance	MHD	0.75
Jaccard index	JAC	0.81	Hausdorff distance	HD	0.66
Dice	DICE	0.81	Adjusted Rand index	ARI	0.65
F-measure	FMS	0.81	Dice	DICE	0.64
Interclass correlation	ICC	0.81	F-measure	FMS	0.64
Cohen's kappa	KAP	0.81	Interclass correlation	ICC	0.64
Adjusted Rand index	ARI	0.80	Cohen's kappa	KAP	0.64
Area under ROC curve	AUC	0.72	Jaccard index	JAC	0.62
True-negative rate (specificity)	TNR	0.72	Accuracy	ACU	0.56
Accuracy	ACU	0.71	Global consistency error	GCE	0.56
Global consistency error	GCE	0.71	Rand index	RI	0.56
Rand index	RI	0.71	Variation of information	VOI	0.56
Variation of information	VOI	0.71	Average distance	AVD	0.54
Positive predictive value	PPR	0.64	Positive predictive value	PPR	0.53
Mahalanobis distance	MHD	0.47	Fallout	FPR	0.48
Probabilistic distance	PBD	0.41	True-positive rate (sensitivity)	TPR	0.48
Average distance	AVD	0.39	Volumetric similarity	VS	0.47
Hausdorff distance	HD	0.38	Probabilistic distance	PBD	0.36
Fallout	FPR	0.23	Area under ROC curve	AUC	0.34
True-positive rate (sensitivity)	TPR	0.23	True-negative rate (specificity)	TNR	0.34
Mutual information	MI	0.19	Mutual information	MI	0.14

Table 6.4 shows the correlations at system level that are significantly stronger than the correlations of rankings at segmentation level (Table 6.3). Actually, this is intuitive because the errors (differences from the manual ranking) in the ranking at segmentation level are higher than in rankings at system level. This stems from the fact that ranking single segmentations using metrics is sensitive to small differences in the metrics, i.e. a segmentation with a higher similarity is ranked as better, regardless of how small the similarity difference is. This is in contrast to manual rankings, where small differences in the quality are ignored. Using significance testing in ranking at system level solves the problem, since the ranking becomes similar to the manual ranking: only systems that have significant performance difference are assigned different rankings, otherwise the same rank. The results of this experiment show the necessity of using significance tests for ranking.

The four metrics selected for evaluating segmentation in the VISCERAL project, namely the Dice coefficient (DICE), the interclass correlation (ICC), the average Hausdorff distance (AVD) and the adjusted Rand index (ARI), are in general (except for the AVG in Ranking 1) ranked at the top, which means they have strong correlation with expert ranking. These four metrics have been selected from the 20 metrics based on a correlation analysis on brain tumour segmentations from the BRATS challenge [3], using the automatic metric selection method proposed in [5].

One observation is interesting for a further analysis, namely the differences in how the metrics are placed in Table 6.4 for MNRK 1 and MNRK 2. For example, the volumetric similarity (VS) is placed at the top for MNRK 1, but at the bottom in MNRK 2. This is also the case for many other metrics. This can be explained by the weak correlation between the two rankers, namely 0.62 (Sect. 6.4.1). However, these differences should be related to the criteria considered in the manual ranking by each of the rankers, i.e. the subjective rating of the different qualities of the segmentations.

6.5 Conclusion

We provide an overview of 20 evaluation metrics for medical volume segmentation that have been implemented in the evaluation tool EvaluateSegmentation. From these metrics, we select four metrics to be used for evaluating the segmentation tasks of the VISCERAL benchmarks. We show in an analysis on synthetic fuzzy segmentations, generated using smoothing functions, that using binary ground truth to evaluate fuzzy segmentations or the opposite (fuzzy ground truth to evaluate binary segmentations) has a considerable impact on the system ranking, if the systems are similar in their performance. Therefore, it is strongly recommended to always evaluate using a threshold of 0.5 if the segmentations/ground truth is mixed in terms of fuzzy and binary modes. Furthermore, we show that different metrics are differently invariant against fuzzification, i.e. differently sensitive to the combinations of fuzzy/binary volumes. In an analysis using manual rankings provided by two radiologists, compared to the rankings produced by the 20 evaluation metrics, we show that the correlation between metric rankings and manual rankings is significantly stronger when using significance tests, since small performance differences are mostly ignored by manual rankers. We also provide an evaluation methodology and metrics for evaluating the VISCERAL Detection Benchmark.

Acknowledgements The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement 318068 (VISCERAL).

References

1. Hanbury A, Müller H, Langs G, Weber MA, Menze BH, Fernandez TS (2012) Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In: Catarci T, Forner P, Hiemstra D, Peñas A, Santucci G (eds) CLEF 2012. LNCS, vol 7488. Springer, Heidelberg, pp 24–29. doi:10.1007/978-3-642-33247-0_3
2. Krenn M, Dorfer M, Jiménez del Toro OA, Müller H, Menze B, Weber MA, Hanbury A, Langs G (2016) Creating a large-scale silver corpus from multiple algorithmic segmentations. In: Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, Cai W, Metaxas D (eds) MCV 2015. LNCS, vol 9601. Springer, Cham, pp 103–115. doi:10.1007/978-3-319-42016-5_10
3. Menze B, Jakab A, Bauer S, Reyes M, Prastawa M, Leemput KV (eds) (2012) MICCAI 2012 challenge on multimodal brain tumor segmentation BRATS2012, MICCAI, Nice, France
4. Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 15:29
5. Taha AA, Hanbury A, Jiménez del Toro OA (2014) A formal method for selecting evaluation metrics for image segmentation. In: 2014 IEEE international conference on image processing (ICIP), Paris, France, pp 932–936