Validation of Liver Tumor Segmentation in CT Scans by Relating Manual and Algorithmic Performance – A Preliminary Study

Jan Hendrik Moltz, Jan Rühaak, Christiane Engel, Ulrike Kayser, and Heinz-Otto Peitgen

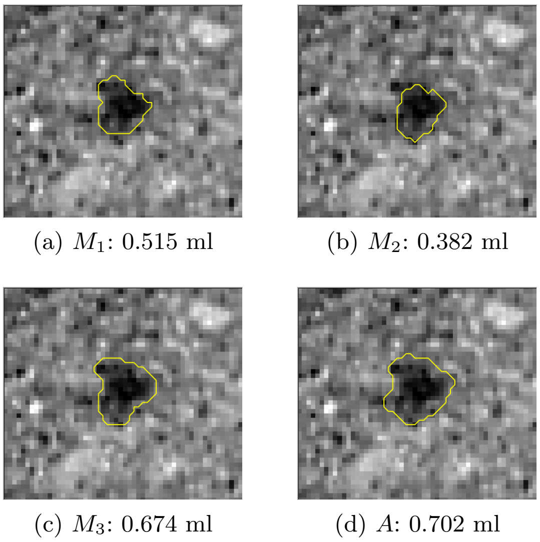
Fraunhofer MEVIS – Institute for Medical Image Computing, Bremen, Germany
jan.moltz@mevis.fraunhofer.de

Abstract. The development of segmentation algorithms for liver tumors in CT scans has found growing attention in recent years. The validation of these methods, however, is often treated as a subordinate task. In this article, we review existing approaches and present first steps towards a new methodology that evaluates the quality of an algorithm in relation to the variability of manual delineations. We obtained three manual segmentations for 50 liver lesions and computed the results of a segmentation algorithm. We compared all four masks with each other and with different ground truth estimates and calculated scores according to the validation framework from the MICCAI challenge 2008. Our results show some cases where this more elaborate evaluation reflects the segmentation quality in a more adequate way than traditional approaches. The concepts can also be extended to other similar segmentation problems.

1 Introduction

In oncological therapy monitoring, the estimation of tumor growth from consecutive CT scans is an important aspect in deciding whether the given treatment is adequate for the patient. Traditionally, this is done by measuring and comparing the largest axial diameters of each lesion manually, but recent advances in image processing also allow a semi-automatic volumetry, which has the potential to enhance the accuracy and reproducibility of growth estimation. Lesion segmentation is an essential prerequisite for volumetry and efficient algorithms are needed for different kinds of tumors.

This article focuses on the segmentation of liver metastases in CT scans. They are among the clinically most important entities in this context, but also pose a particular challenge for segmentation algorithms due to the great diversity in their appearance. In recent years, a considerable amount of effort has been spent on the development of semi-automatic and fully automatic algorithms for liver tumor segmentation, especially triggered by a competition at MICCAI 2008 [1]. The validation of these methods, however, seems to play a subordinate role in most publications. Although it is well known that manual segmentations even by experts will always show some degree of variability (Fig. 1), algorithms are



(a) M_1: 0.515 ml

(b) M_2: 0.382 ml

(c) M_3: 0.674 ml

(d) A: 0.702 ml

Fig. 1: Three manual and an algorithmic segmentation of the same liver tumor. The volumes differ almost by a factor of 2 but all masks are visually plausible.

often evaluated by comparing their results to a single reference segmentation which is considered to be the “ground truth”. Various similarity measures can be computed from these two masks, but it is not clear how the results reflect the actual accuracy of an algorithm, especially from a clinical user’s point of view.

From our experience, a fair and objective validation of a segmentation algorithm should take the uncertainty in manual segmentations into account. This means that more than one reference mask should be obtained and that a validation metric should relate the performance of the algorithm to the variability between these reference masks. If we find, for example, that the volume of a lesion computed by an algorithm is within the range of variation of several manually determined volumes, this may give a more conclusive statement about the quality of the method than just giving a difference to one of these reference volumes.

In this article, we take a closer look at evaluation concepts in the recent literature on liver tumor segmentation and show first steps towards a new validation methodology that follows the ideas mentioned above. They will be exemplified by a study on our own algorithm using 50 liver lesions and three reference segmentations for each of them.

2 Validation with Multiple Reference Segmentations: A Critical Review

We reviewed publications from the last ten years that focus partially or completely on liver tumor segmentation in CT with regard to the validation techniques that were applied to assess the quality of the algorithms. In this section, we show different strategies that have been used to deal with multiple reference segmentations.

Yim and Foran [8] compared the reproducibility of manual and semi-automatic area measurements from repeated reading.

Popa *et al.* [4] obtained four reference segmentations by different readers and estimated a ground truth using the STAPLE algorithm [7]. Comparison metrics were computed for one of the reference masks and the algorithmic result versus the estimated ground truth. Unfortunately, the results are biased since the algorithmic segmentation was not used for STAPLE while the reference mask was.

Zhao *et al.* [9] computed concordance correlation coefficients (CCC) for volume measurements of three readers and their algorithm as well as an overall CCC to assess the accuracy of their results.

Ray *et al.* [5] used reference masks of four readers from three reading sessions each. They analyzed the development of the volume measurements between the sessions and compared the measured volumes with those from the algorithm.

A methodology that can be considered the current state of the art was used at several MICCAI segmentation challenges and first introduced by Heimann *et al.* [2]. Deng and Du [1] adapted it for the liver tumor segmentation challenge in 2008. A scoring system was developed that evaluates segmentation results in consideration of the variability in manual delineations. For this purpose, two reference masks were acquired, one of which is used as the “ground truth” to compare the algorithmic segmentations with. This reference mask was confirmed by both readers as correct. The other reference mask is used to estimate the typical deviation of an independent observer. Five comparison metrics are transformed into a score between 0 and 100 such that 100 corresponds to complete agreement with the “ground truth” and 90 to the average deviation of the second reference mask. The total score is the mean of the five scores obtained from the metrics.

This is an interesting concept that tries to increase the objectivity of segmentation validation. It does, however, not capture the uncertainty of human measurements completely. First, it can only be used for lesions where the readers are able to agree on a “perfect” segmentation. The other cases were discarded for the MICCAI challenge, so the problem of validating segmentations where manual segmentations differ and no actual ground truth is available was left aside. In practice, however, such cases are not rare, so the validation result may sometimes depend strongly on which reference mask is chosen.

Figure 2 shows a graphical example where one reference mask is just a dilation of the other, i. e., one reader produced a mask that is systematically larger. Although this scenario is simplified for illustration, results like this can easily occur when the two readers use different window settings or when the contrast between the lesion and the parenchyma is low. The areas with high scores are significantly different depending on the choice of the “ground truth” as is shown in Fig. 2a and 2b. Given the information from both reference masks, a symmetric distribution as in Fig. 2c would be more desirable.

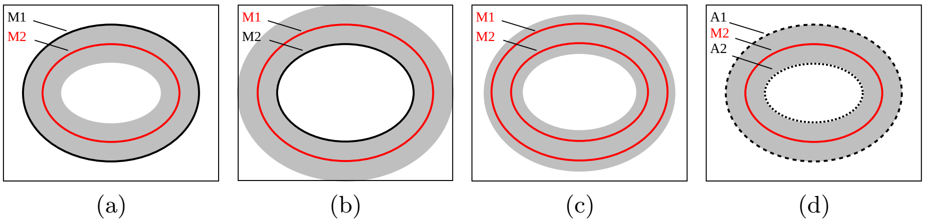


Fig. 2: Graphical illustration of some problems with the MICCAI scoring system. Solid lines denote manual segmentations, red ones are the "ground truth", dotted lines are algorithmic segmentations. The gray area covers the segmentations that would get a score above 90. (a) M_2 is used as the ground truth. (b) M_1 is used as the ground truth. (c) M_1 and M_2 are treated equally. (d) M_2 is used as the ground truth. A_1 and A_2 would get the same score although A_1 is obviously better if both reference masks are taken into account.

But even if the readers are able to agree on a "ground truth", there is still a degree of uncertainty about the true segmentation which is reflected in the independent segmentations of the other readers. This information, however, is effectively discarded by the MICCAI scoring system since the direction or "sign" of the deviation is not taken into account. This effect is illustrated in Fig. 2d. If we choose the red segmentation as the ground truth, A_1 and A_2, which are just an erosion and a dilation of the ground truth, would get the same score. But from the other reference mask we know that A_1 is probably better than A_2.

In order to see the impact on real data, we get back to Fig. 1 where the manual segmentations are very different. Depending on the choice of the "ground truth", the scores for the relative volume difference are 71.9 (M_1), 35.3 (M_2) and 96.9 (M_3), respectively, which covers the range from an excellent to a rather poor agreement. This example shows that using three reference masks instead of one or two allows us to get a better idea of the most probable ground truth and of the inherent uncertainty in a particular segmentation problem. A reliable score should therefore incorporate all available information, for example by averaging the individual scores for all reference masks or by estimating an aggregate ground truth with an algorithm such as STAPLE.

3 Data and Validation Methodology of our Study

In order to exemplify the ideas pointed out above and to examine our concepts more closely, we initiated a validation study for our own liver tumor segmentation algorithm [3]. This algorithm combines adaptive thresholding with model-based morphological postprocessing and adds special treatment for peripheral and rim-enhancing lesions. The input to the method is a stroke across the tumor that should roughly indicate its largest diameter.

The data set that was used for our study consists of 38 CT scans from several hospitals and CT scanners that were acquired for liver surgery planning. For a total number of 50 lesions, manual segmentations were obtained from three experienced radiology technicians. These lesions were selected randomly from a larger amount of data and can be considered as a representative collection of segmentable liver tumors. Very small, very inhomogeneous, and not clearly delimitable lesions had been excluded initially. We applied our algorithm to the lesions, using the largest axial diameters of the union of all manual segmentations as strokes.

We call the three manual segmentations M_1, M_2 and M_3 and the automatic one A. With these four masks, we performed two experiments. First, we computed the comparison measures and scores as in the MICCAI challenge 2008 for all six pairs of masks. The results are divided into three pairs which contain A and three which do not. A comparison of the values among both groups gives us an idea of how well the algorithm performs in consideration of the uncertainty of the human observers.

In the second experiment, we estimated a ground truth from the four masks using four different techniques – voxelwise voting (2-out-of-4 and 3-out-of-4), STAPLE [7], and shape-based averaging (SBA) [6] – and compared each of the four masks with the result. We also included A into the computation because comparing the scores would not be fair if some of the masks had influenced the ground truth estimate and some had not. Since the three reference masks still dominate the result, incorporating A does not cause a significant bias.

4 Results

Table 1a shows the medians of the comparison measures for all pairs of masks. It can be seen that the agreement among the reference masks is better for all metrics but the difference is not very large. The average score is 83.5 for manual pairs and 78.0 for manual-algorithmic pairs.

In Table 1b we report the median MICCAI scores for all masks compared with the ground truth estimates. The results are consistent among the different ground truth estimation methods as well as with the results of the pairwise comparison.

In addition to the overall results, we give a closer look at some cases where our validation strategy provides more information or is more consistent with the visual impression than the MICCAI scores for the individual reference masks.

The first example, shown in Fig. 3, is a case where the average of the pairwise scores as well as the score against the STAPLE ground truth is higher for A than for any of the reference masks. This means that A is a good compromise between the divergent reference masks which is confirmed by visual inspection and by the high similarity to the STAPLE ground truth. This effect could not have been identified with the MICCAI score which would suggest slightly below average segmentation quality. Interestingly, the results for the SBA ground truth are completely different and yield the best agreement with M_1. This suggests that the behavior of the ground truth estimators can show subtle discrepancies in individual cases that should be examined in more detail.

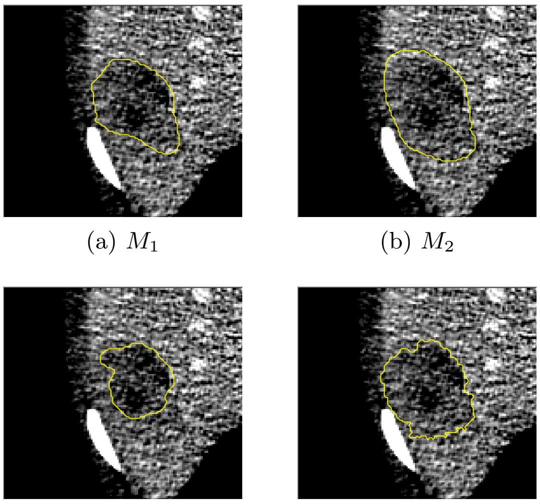
	M_1M_2	M_1M_3	M_2M_3	AM_1	AM_2	AM_3
Volume overlap	68.9	66.5	72.5	60.0	64.8	66.5
Relative volume difference	19.2	25.3	13.7	28.0	25.1	13.6
Avg. surface distance	0.67	0.69	0.53	0.97	0.89	0.87
RMS surface distance	0.88	0.93	0.77	1.25	1.20	1.20
Max. surface distance	3.08	3.58	2.66	4.04	4.01	4.01
MICCAI score	83.2	81.0	86.2	76.9	77.4	79.8

(a)

	M_1	M_2	M_3	A
2-out-of-4 voting	88.2	90.2	88.2	83.8
3-out-of-4 voting	84.0	91.5	91.2	79.8
STAPLE	87.9	90.4	88.2	83.6
Shape-based averaging	87.1	90.2	90.7	84.1

(b)

Table 1: (a) Results of the pairwise comparison between manual and algorithmic segmentations (medians over all lesions). (b) Results of the comparison of individual segmentations with the estimated ground truth according to the different methods (medians of MICCAI score over all lesions).



(a) M_1

(b) M_2

(c) M_3

(d) A

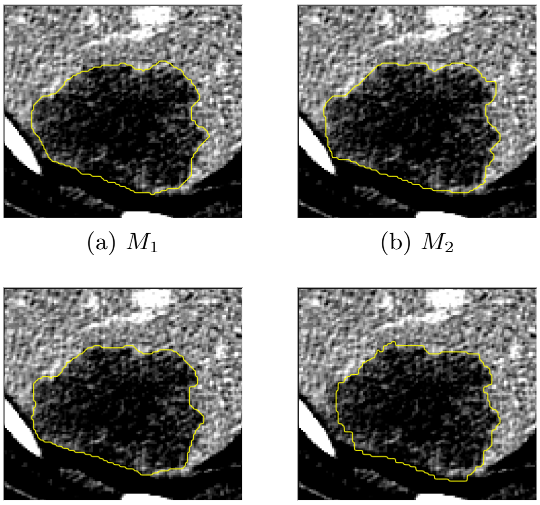
M_1M_2	M_1M_3	M_2M_3	AM_1	AM_2	AM_3
64.1	65.1	37.5	67.2	78.2	41.2

(e) Pairwise scores

	M_1	M_2	M_3	A
STAPLE	74.3	83.3	47.0	88.3
SBA	87.1	64.5	69.7	69.5

(f) Ground truth scores

Fig. 3: Example from our study where the algorithmic segmentation has better scores than the reference masks and provides a visually satisfying compromise.



(a) M_1

(b) M_2

(c) M_3

(d) A

M_1M_2	M_1M_3	M_2M_3	AM_1	AM_2	AM_3
89.3	90.5	90.8	84.5	81.8	84.4

(e) Pairwise scores

	M_1	M_2	M_3	A
STAPLE	93.1	93.5	94.1	84.7
SBA	93.6	91.9	94.1	87.7

(f) Ground truth scores

Fig. 4: Example from our study where the reference masks have a very good agreement and the algorithmic segmentation shows a deviation in spite of the good scores.

The second example in Fig. 4 shows the opposite effect. From the pairwise scores, the accuracy of A seems to be above average, but the scores between the reference masks and especially the scores of the reference masks against the ground truth estimates are even higher. In this case, the reference masks show a very good agreement and the validation should be less tolerant to the inaccuracy that can be seen in Fig. 4d.

5 Discussion

In our preliminary study, we examined how the results of a segmentation algorithm can be evaluated by relating them to the performance of manual segmentations. At a first glance, it seems that pairwise comparison of all available masks leads to similar conclusions as comparison with an estimated ground truth in most cases, but this should be investigated further. In particular, cases such as the one shown in Fig. 3 where different methods yield different results should be analyzed in more detail in future studies.

So far, we have presented the general ideas behind our validation methodology and shown the possible benefit by some examples. The goal of our ongoing work is to derive new quantitative metrics that allow an objective accuracy assessment of a segmentation algorithm and a fair comparison between different algorithms on different data. These metrics should adapt their tolerance depending on the uncertainty of the experts which is reflected in the variability of the reference segmentations. This could be done by merging a set of individual comparison measures into a new score or by estimating a fuzzy ground truth. Figures 3 and 4 show examples where this could give a more adequate validation than the MICCAI score.

This concept raises another important question: How many reference segmentations are needed for this kind of validation? Even though we want to establish a fuzzy ground truth, we still require a statistical stability to achieve reproducible results. This will also be a subject of further research.

We would like to point out that our concept is not restricted to liver tumor segmentation. It can be used for other tumor entities and potentially be extended to general segmentation problems where no ground truth is available.

Acknowledgment. This work was funded in part by the European Regional Development Fund.

References

1. Deng, X., Du, G.: Editorial: 3d segmentation in the clinic: A grand challenge II - liver tumor segmentation. In: Proceedings MICCAI Workshop on 3D Segmentation in the Clinic (2008), http://grand-challenge2008.bigr.nl/proceedings/pdfs/lts08/00_Editorial.pdf
2. Heimann, T., van Ginneken, B., Styner, M.A., et al.: Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Transactions on Medical Imaging 28(8), 1251–1265 (2009)
3. Moltz, J.H., Bornemann, L., Kuhnigk, J.M., et al.: Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans. IEEE Journal of Selected Topics in Signal Processing 3(1), 122–134 (2009)
4. Popa, T., Ibanez, L., Levy, E., et al.: Tumor volume measurement and volume measurement comparison plug-ins for VolView using ITK. In: Proceedings SPIE Medical Imaging. vol. 6914, pp. 69141B1–69141B8 (2006)
5. Ray, S., Hagge, R., Gillen, M., Cerejo, M., Shakeri, S.: Comparison of two-dimensional and three-dimensional iterative watershed segmentation methods in hepatic tumor volumetrics. Medical Physics 35(12), 5869–5881 (2008)
6. Rohlfing, T., Maurer Jr., C.R.: Shape-based averaging. IEEE Transactions on Image Processing 16(1), 153–161 (2007)
7. Warfield, S.K., Zou, K.H., Wells, W.M.: 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)
8. Yim, P.J., Foran, D.J.: Volumetry of hepatic metastases in computed tomography using the watershed and active contour algorithms. In: Proceedings IEEE Symposium on Computer-Based Medical Systems. pp. 329–335 (2003)
9. Zhao, B., Schwartz, L.H., Jiang, L., et al.: Shape-constraint region growing for delineation of hepatic metastases on contrast-enhanced computed tomograph scans. Investigative Radiology 41(10), 753–762 (2006)