A tool for efficient creation of probabilistic expert segmentations

Jan Hendrik Moltz[1]
jan.moltz@mevis.fraunhofer.de
Christiane Steinberg[2]
christiane.steinberg@miw.uni-luebeck.de
Benjamin Geisler[1]
benjamin.geisler@mevis.fraunhofer.de
Horst Karl Hahn[1]
horst.hahn@mevis.fraunhofer.de

[1] Fraunhofer MEVIS
Institute for Medical Image Computing
Bremen, Germany

[2] University of Lübeck
Lübeck, Germany

Abstract

The validation of segmentation algorithms is often based on manual expert delineations, but they are subject to variability. The standard approach of using a single binary reference segmentation may therefore provide misleading results. While using multiple references increases reliability, the effort required from the experts may become infeasible. As a solution, we developed a tool that allows individual experts to create probabilistic segmentations by expressing their uncertainty about the true segmentation. An explicit distinction between statistical and semantic uncertainty is made. In a study, we compared the results of three users using our new tool for delineating liver tumors in CT with ten users drawing conventional contours. We found that with our tool more variability could be captured by a lower number of experts.

1 Introduction

The development of segmentation algorithms for different anatomical structures and imaging protocols is an important task in medical image analysis. The validation of these methods, however, is often treated as a subordinate problem. Algorithms are often evaluated by comparing their results to a single reference segmentation which is considered to be the "ground truth", although it is well known that manual delineations even by experts always show some degree of variability. This variability reflects the uncertainty of the experts about the true segmentation.

For example, in a previous publication [3] we have analyzed the variability among ten expert delineations for liver tumors in CT. Using the average segmentation as a reference, we found that any subset of the experts makes a significant error. A closer look at the individual delineations reveals that two kinds of uncertainty should be distinguished. *Statistical uncertainty* can be modeled by a mean contour and an uncertainty margin of a particular width. It can be caused by differing perception of the object size, for example due to different window settings. If the contrast is low, some readers may tend to draw the outline around all possible

object voxels, while others mark only the region that certainly belongs to the object. In this case, it can be assumed that the experts essentially agree about the segmentation. An algorithm that produces any of their segmentations and anything in between can be considered correct. *Semantic uncertainty*, on the other hand, cannot be modeled by deviation around a mean contour. Instead, larger regions, not just narrow bands of voxels, are included by some experts and excluded by others, resulting in a fuzzy segmentation with distinct areas of a particular probability. Then, a good algorithm should be close to at least one of the expert delineations, whereas a compromise between them would not be desirable.

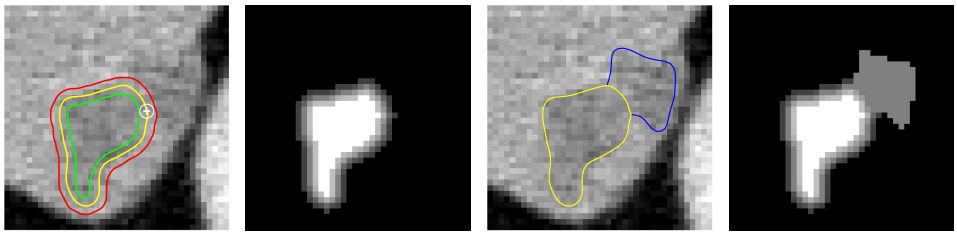
These observations suggest that the common approach of using a hard "ground truth" is not adequate for validation. In cases where experts are not certain about the true segmentation, this uncertainty should be incorporated into the validation methodology. Unfortunately, it is often infeasible to acquire reference segmentations by a substantial number of experts. Even large validation initiatives such as LIDC [1] collected only four segmentations per case. Most individual researchers do not have access to more than one or two experts. A common restriction, however, is that experts are usually asked to draw a single contour as their best estimate of the true segmentation. Variability is then measured in terms of the differences between the best estimates of multiple experts. An aspect that is mostly disregarded is the uncertainty of *each individual* expert. Before drawing a contour, each reader has to make two decisions: where to draw the most probable boundary within an often blurred margin and whether or not to include ambiguous regions which may or may not be part of the object.

The hypothesis of our work is that the variability between multiple experts can in part be reproduced by a smaller number of experts, if they are given a tool to express their uncertainty. Such a tool will be presented and evaluated in this paper. The evaluation uses the same data as our previous study [3] and compares the results of three users with the new tool to those of ten users drawing conventional contours. Although we focus on a particular problem, liver tumor segmentation in CT, the methodology is easily generalised.

2 Related work

A related approach was presented by Restif [4]. He introduced a framework called *Comets* that allows a single user to create a probabilistic reference segmentation. It was specifically developed for 2d cytometry images where blurred boundaries and connected objects are common problems. The user draws the most probable outline and adds inner and outer limit pixels which are definitely inside or outside the object, but as close to the border as possible. From this input a confidence map is computed by setting 0 on the drawn outline, ± 1 on the limit pixels and interpolating on all other pixels.

As compared to Restif's work, this article presents three additional contributions. First, the focus will be on 3d images. While transferring the concept to 3d is straightforward in principle, efficiency becomes an issue when contours have to be drawn in each slice. The concept of limit pixels may not be intuitive for all users and it might take some time to define them on all slices. Therefore, we opted for a simpler and more efficient interaction based on contours. Second, Comets does not distinguish statistical and semantic uncertainty but covers both by a single method and blends them together in the confidence map. For validation purposes, however, it is advantageous to separate these two aspects. This is done explicitly in our new tool. Finally, Restif does not compare Comets to other ways of generating reference segmentations. Since our work was motivated by the goal to reduce the number of necessary experts without losing information, we conducted a user study to evaluate this.



(a) (b) (c) (d)

Figure 1: Illustration of the workflow of the new tool and the results it produces. (a) User-drawn contour (yellow) and inner and outer contours (green and red) automatically constructed from the radius of the circle. (b) Probability map. (c) Additional region with confidence 0.5 (blue). (d) Probability map.

3 Workflow

With our tool, implemented in MeVisLab [5], segmentation is done in two phases. In the first phase, the most probable contour is drawn. The statistical uncertainty is modeled by a rim around this contour. The inner boundary of the rim delineates all voxels which are definitely part of the tumor. Analogously, all voxels outside the outer boundary definitely belong to the background. The width of the uncertainty rim is set by the user before drawing the contour. For simplicity, this setting is applied globally on each slice, but can be adapted locally afterwards. The current width is visualized as the diameter of a circle displayed at the cursor position and can be changed by turning the mouse wheel (Figure 1(a)).

Once the user has finished drawing, the inner and outer contours are generated by applying a distance transform to the user-defined contours and adding or subtracting the uncertainty radius. These contours are displayed and can be edited. Although in many cases a global uncertainty radius is reasonable, there are cases where a different value should be set locally. For example, a tumor may have a blurred boundary to the liver parenchyma, but a clearly defined one to a structure outside the liver. Editing is achieved by drawing new partial contours which are inserted into the existing ones.

Now the contours are transformed into a probability map (Figure 1(b)). Voxels are assigned a value of 1 if they are inside the inner contour and 0 if they are outside the outer contour. Between the contours, probabilities are linearly interpolated. Note that, unlike Restif [4], the values are limited to $[0, 1]$ and do not decrease further outside the outer contour.

In the optional second phase, additional regions can be outlined and assigned a confidence of belonging to the tumor (Figure 1(c)). For these regions, no uncertainty margin is defined because that seemed to be too confusing for users, although technically it would not be a problem. Regions are included in the probability map by using the maximum of the value assigned in the first phase and the confidence set by the user (Figure 1(d)). Alternatively, the results of the two phases can be stored separately for further analysis.

4 Evaluation

Our new tool was evaluated in a study with three experts (one radiologist and two radiology technicians) and the same 13 liver tumors that were used in our previous study [3]. Four example tumors are shown in the top row of Figure 2.

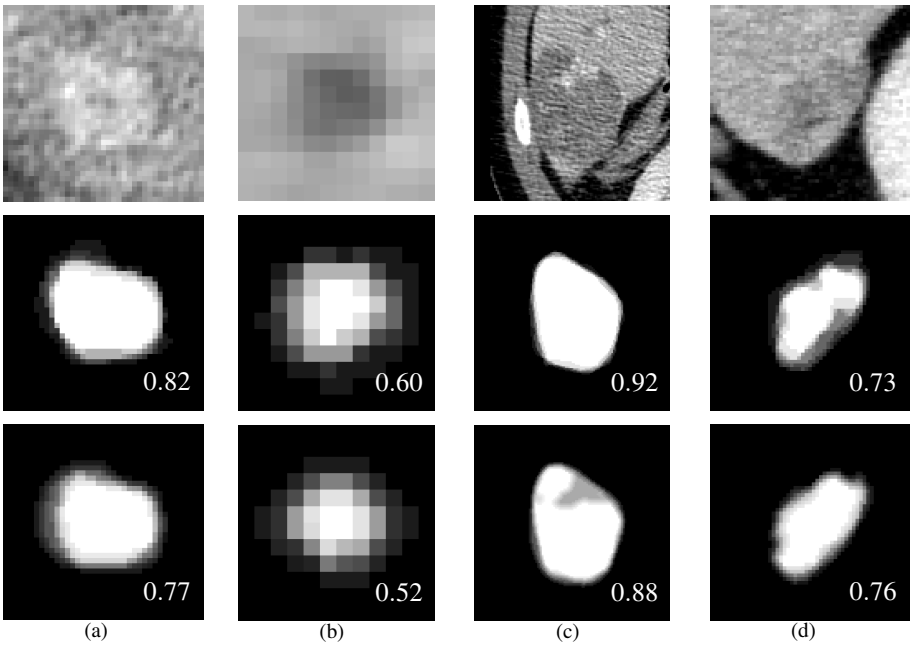


Figure 2: Four example tumors from the study. Top row: Original images. Middle row: Averaged probability maps created by ten experts drawing conventional contours [3]. Bottom row: Average probability maps created by the three study participants with the new tool. Additionally, the fuzzy self-overlap as defined in Section 4 is given.

The usage of the features offered by the tool varied across the participants. Readers 1 and 2 adapted the uncertainty width in each case, whereas Reader 3 always used the same value (in voxels). Reader 3 also did not draw any additional regions. The two others added three and eight regions, respectively, to eight of the 13 tumors.

We compared the new results to our earlier ones and found a high visual similarity for many of the tumors. The middle and bottom rows of Figure 2 show some examples. The chosen uncertainty widths correspond well to the statistical uncertainty among ten experts as illustrated by tumors (a) and (b). Still, some interesting effects can be seen. In tumor (c), for instance, a region was left out by one of the three readers although it had been included by all ten readers in the earlier study. For tumor (d), on the other hand, there was slightly more variability among ten readers than could be reproduced by three.

For a more quantitative analysis, we define a metric that captures the variability encoded in a probabilistic segmentation. It is based on the fuzzy volume overlap, where the volume of a segmentation is the sum of the probabilities of all voxels, with intersection and union being defined by the voxel-wise minimum and maximum [2]. The fuzzy overlap of two segmentations compares two aspects, the mean segmentations and the spread of probabilities around them. Applying the fuzzy overlap to a probabilistic segmentation and its own mean segmentation, defined by thresholding at 0.5, measures the variability as desired. We call this the *fuzzy self-overlap*. It is 1 for a binary segmentation and gets lower the more the probabilities are spread. Figure 2 gives these values for the example tumors.

Figure 3 compares the variability in averaged segmentations created from the ten conven-

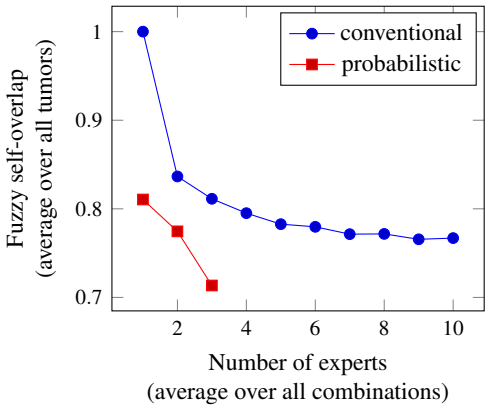


Figure 3: Variability in combined segmentations by different numbers of experts, using conventional and probabilistic expert segmentations. The lower the fuzzy self-overlap, the higher the variability.

tional segmentations of our earlier study and from the three probabilistic ones of the present study. In the plot, it is clearly visible that with the new tool more information can be acquired using fewer experts. One expert using the new tool could replace three experts drawing conventional contours. Together, the three experts in our study generated more variability than ten in the previous study.

After the study, the participants were interviewed. They said that they felt unfamiliar with expressing their uncertainty because usually they have to make a crisp decision. While, however, the uncertainty width was adopted easily, the readers had difficulties defining additional regions and quantifying their confidence. This shows that users need some training to get used to the new way of thinking the tool requires. The reader who achieved the best results was already interviewed in the development phase and probably had the best understanding of the concepts at the time of the study.

5 Discussion

The motivation for this work was to be able to reduce the number of experts needed for a validation study without losing information and without increasing the workload per expert too much. A basic decision was made to separate statistical and semantic uncertainties explicitly, both for reducing the effort and for making it available for further analysis. In the study, the statistical uncertainties were captured well at virtually no additional cost because the uncertainty width was set very quickly. A possible disadvantage of the conceptual separation, however, is the fact that users typically decide to add a confidence region in the first phase, but have to wait for the second phase before they can actually draw it. This requires a high concentration and memory capacity and might be a reason why not many confidence regions were added. A workflow that allows alternating the two phases on each slice might improve this. As a further improvement, one might think about not just adding, but also subtracting confidence regions from the initial segmentation. This might be more intuitive than leaving out regions with a very high confidence in the first phase and adding them later.

The results of the study show that using the new tool expert uncertainty can be recovered with a lower number of experts as compared to conventional contours. This was confirmed both visually and quantitatively. It is interesting to see that in some cases confidence regions were used that have no correspondence among ten experts. This shows that the explicit cap-

turing of uncertainty can actually gather additional information compared to just averaging over a large number of segmentations. But on the other hand, there are also some cases where the complete variation cannot be reproduced with a lower number of readers. In Figure 2, tumors (c) and (d) illustrate this duality.

The processing time was not measured, but from our observations during the study it can be said that the new methods allows a considerable reduction of efforts. Assuming that segmentation took 25 % longer than pure outlining, which is a very conservative estimation since confidence regions are typically small and cover only a couple of slices, the overall person time was still reduced by almost two thirds.

Future work is necessary to investigate how these probability maps can be used for algorithm validation. Since they are not inherently binary, many common approaches are not directly applicable. Some widely used metrics like the volume overlap can be easily generalized for probabilistic segmentations, whereas for surface distances there is no obvious solution and different proposals have been made. Crum et al. [2] discuss their application in medical image analysis. They focus, however, on the case where the algorithm result is probabilistic rather than the reference segmentation. Further experiments should provide insight into how suitable these methods are for validation. Also, common methods are not able to make use of the explicit distinction between statistical and semantic uncertainty. The additional information that is becoming available calls for a completely new validation paradigm that works not only on (a set of) random expert delineations, but builds up knowledge about plausible and implausible segmentations.

We believe that it is important to work towards more meaningful and reliable validation of segmentation algorithms. This article is a first step that shows how this can be achieved with limited expert efforts.

Acknowledgment

We thank Christiane Engel and Gülsen Yanc for participating in the study. This work was funded in part by the European Regional Development Fund.

References

[1] S. Armato III, G. McLennan, L. Bidaut et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011.

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

[3] J. H. Moltz, S. Braunewell, J. Rühaak et al. Analysis of variability in manual liver tumor delineation in CT scans. In *IEEE ISBI*, pages 1974–1977, 2011.

[4] C. Restif. Revisiting the evaluation of segmentation results: Introducing confidence maps. In *MICCAI*, pages 588–595, 2007.

[5] F. Ritter, T. Boskamp, A. Homeyer et al. Medical image analysis: A visual approach. *IEEE Pulse*, 2(6):60–70, 2011.