Simulation of User Interaction for Performance Evaluation of Interactive Image Segmentation Methods

Emmanouil Moschidis[1] and Jim Graham[2]

Imaging Science and Biomedical Engineering, School of Cancer and Imaging Sciences, Stopford Building, The University of Manchester, Oxford Road, Manchester M13 9PT.

Abstract. Interactive image segmentation is often employed in the context of medical image analysis, as an alternative to automatic and manual image segmentation. In the last decade it has attracted a lot of attention due to the advent of efficient algorithms able to perform in interactive speed even for large three dimensional (3D) images. However, the human integration in the segmentation process restricts its repeatability and impedes its objective evaluation. Also, it inhibits the identification of the intrinsic properties of the algorithms. In this paper we report on a framework for performance evaluation of interactive image segmentation techniques, which is based on simulated user interaction. This allows for the construction of reproducible and tractable experiments, which can form the basis of a systematic and objective performance evaluation framework for interactive segmentation methods. We demonstrate quantitative results using three interactive segmentation algorithms from the literature.

1 Introduction

In the context of medical image analysis, interactive image segmentation methods are often preferred to automatic or manual ones. Automatic methods usually fail to produce results that meet the expectations of a human, whereas manual approaches incorporate tedious marking-up sessions with increased cognitive load. Interactive image segmentation appears to be the best compromise between automatic and manual methods due to its properties; it provides the user with enough control over the entire process so that s/he can achieve an arbitrary segmentation that meets her/his expectations, while it demands less effort than the manual approach.

An interactive image segmentation method consists of the following components: the computational part, the interactive part and the graphical user interface (GUI) [1]. The GUI is the component that accepts the user's guidance for action, often via visual programming components (text-boxes, drop-down menus, sliders etc.), also termed controls, or via direct image clicks (pictorial input). The interactive part translates the input given via the GUI into parameters that can be used by the computational part. Finally, the computational part uses a specific model to encode the information available in an image (e.g. graph, MRF etc.). Once the required parameters are provided by the interactive part, it performs calculations that lead to a segmentation outcome.

In interactive segmentation of medical images, an expert (usually a radiologist) is steering the segmentation process. The acceptance or rejection of the segmentation outcome depends solely on his perception of the result with respect to the actual anatomical structure, termed ground truth. This interaction, however, must be minimal, in order to allow for efficient analysis of the datasets in terms of time and effort (speed and cognitive load). The most successful interaction pattern in the literature is the provision of pictorial input via brush strokes[5-8]. This pattern of interaction is intuitive, fast, easy, applicable to three dimensions (3D) and gives the user the ability to achieve arbitrary segmentations. In the context of such an interaction, the user is using a brush to mark specific groups of pixels (voxels in 3D), also termed as seeds, of an image as belonging to a specific class. For binary segmentation these classes are only two, foreground and background. In this study we focus on the task of binary segmentation in 3D medical images.

Two main problems arise, when evaluating interactive segmentation methods with the brush strokes as their interaction pattern: the lack of repeatability and the excessive input provision. The lack of repeatability is due to the human integration in the segmentation process. The excessive input provision is due to the nature of brush strokes, which provide a large number of input seeds that make most methods perform well. These two problems impede the objective performance evaluation of various aspects of interactive segmentation techniques. We argue that the objectiveness of such an evaluation framework can be guaranteed by simulated automation of the user interaction and assessment of the performance of the selected methods with small number of input seeds.

The automation of the variety of cognitive actions performed by the user is of course a difficult task. It is hard to automate the causal relationships associated with every action in the segmentation process. Moreover, it is debatable

1 Emmanouil.Moschidis@postgrad.manchester.ac.uk
2 Jim.Graham@manchester.ac.uk

what a human considers as a normal or natural interaction. However, having a simulation that mimics realistically the human interaction is of great importance, since the experiments that are based on it can provide insight regarding the true interaction as well. In section 2 we propose a set of patterns for achieving this goal. In section 3 we present the evaluation framework used in our study and the metrics associated with it. In section 4 we report some preliminary results from the use of the suggested framework for the performance evaluation of three recent interactive segmentation methods in 3D, the GraphCuts [5,6], the RandomWalker [7] and the GrowCut [8]. This is also the first time that implementations of the latter two algorithms are reported in 3D. GraphCuts is a graph-based method that treats a graph as a flow network. The cut on the graph that separates the foreground seeds from the background seeds is given by the saturated edges in the network with the use of the MinCut/MaxFlow theorem. Random Walker is also a graph-based method, which calculates the probability that a random walk, which initiates from an undefined (unclassified) pixel, will reach a seed point, given the bias that it cannot cross high gradient edges. GrowCut is based on the Cellular Automata model. It considers all pixels in an image as living cells with certain label and strength. Each cell is attacked by its neighbours and in case the attacking force is higher than its strength, the cell changes its label to that of the attacking cell.

2 Simulated User Interaction

We identify two main patterns of user interaction, the random clicks and the careful seed selection. In the first case the user is selecting quickly and rather randomly foreground and background seeds. In the second one, s/he is trying to carefully select seeds that best represent the foreground and the background. We believe that in the former case the seeds have equal probability to come from any place within the 3D volume, whereas in the latter case most of the seeds will be selected from places near to the object boundaries. Based on this hypothesis, the strategy we follow for automating the user interaction is the following:

Random Clicks: In order to simulate this pattern of user interaction, we save the indices of the foreground and background seeds defined in the ground truth. Consequently we randomly select seeds (indices) that are uniformly spread within these vectors as background and foreground seeds. That way, these seeds will be relatively uniformly spread throughout the ground truth background and foreground volume.

Careful Seed Selection: In order to simulate this pattern of user interaction, we identify the seeds from the ground truth that belong to the surface of the object of interest and those lying one voxel outside its surface (outer boundary). The seed indices are saved in two vectors. Consequently, we randomly select seeds (indices) that are uniformly spread within the outer boundary vector for background and the object surface vector for foreground.

Variability of User Interaction: In order to accommodate the variability that characterises the human interaction, we displace the foreground and background seeds, as determined by the ground-truth, in random directions by fixed distances. In order to accommodate both low and high variability we select a variable displacement of 2^i , where $i \in [0,8]$; when i=0 the seed is only displaced to its immediate neighbour, whereas when i=8 it is displaced by 256 positions. It is obvious that for low values of i we simulate an interaction with low imprecision, whereas with high values of i we simulate an interaction with high imprecision.

3 Evaluation Framework

Taking into account the evaluation frameworks that exist in the literature [2-4] we suggest an evaluation framework for interactive segmentation that is summarised in the following components:

Ground Truth: In this study we use real medical images, since they provide more challenges for a segmentation algorithm than artificial ones. More specifically we use a 3D (83×80×104) MR brain image and the task is to segment the brain ventricles. The surrogate of truth is given by manual delineation of the anatomy of interest by one expert. Of course, ideally one would like to have multiple segmentations from multiple experts. However, this segmentation is used as a relative evaluation estimate, acknowledging the fact that this may not be the ideal surrogate of ground truth.

Accuracy: In order to assess the accuracy of the methods, we evaluate the segmentation outcome provided by the computational part of the algorithms for a specified number of input seeds. The result is compared against the surrogate of ground truth and the correctly and misclassified voxels are identified. The confusion matrix is then created, in which voxels are divided into true and false positives (TP, FP) and true and false negatives (TN, FN). The metric for segmentation accuracy is defined as:

$$\text{Accuracy} = 100 \times \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|} \ \% \qquad (1)$$

In addition, a 3D distance transform is used to provide the maximum and the mean distance from the boundary points estimated by the segmentation algorithm to the boundaries provided in the ground truth [3]. This is considered as information complementary to the accuracy metric.

Repeatability: In order to assess the repeatability, we assess the effect of the pertubation of the input seeds on the segmentation result. For a selected number of input seeds, the seeds are pertubed by a variable measure. For every pair of segmentation V_{S1} ,V_{S2} the relative overlap (RelOv), known as Tanimoto coefficient [4], is then calculated as:

$$\text{RelOv} = \frac{V_{S1} \cap V_{S2}}{V_{S1} \cup V_{S2}} = \frac{|TP|}{|TP|+|FP|+|FN|} \qquad (2)$$

Efficiency: In order to assess the efficiency of the interactive segmentation methods, which is related to the speed and the cognitive load of the overall segmentation process, the following questions are posed: "how much interaction does the method require for a plausible segmentation?", "how precise must the user be during the input provision?" and finally "how fast does the computational part process the input and provide results?". The amount of interaction that is required, till the user achieves the desirable outcome, is calculated in terms of clicks or input seeds. The fewer input seeds (low cognitive load) that are required for the segmentation task, the higher is the efficiency of the algorithm. With respect to the second question, an efficient algorithm does not demand precise pictorial input from the user (low cognitive load), in order to deliver a plausible segmentation. Therefore, a repeatable method is also efficient. Finally, the speed of the segmentation's provision by the computational part of the interactive segmentation method is reported in seconds.

4 Experiments and Results

Accuracy: In order to assess the accuracy of the computational part of the three selected interactive segmentation methods, we varied the number of input seeds, both foreground and background, from 1 to 30. Figure 1 depicts the effect of the alteration of the number of seeds on the accuracy of the computational part of the algorithms along with the observed variation for 30 different random seed initialisations per seed number, for the simulation of both cases of user interaction. Figure 2 depicts the maximum observed voxel-based distance of the surface of the segmentation's outcome from the ground truth surface along with its variation. The same seed initialisation was used for all the algorithms. At this point, it is worth recalling that the measured accuracy is considered with respect to the surrogate of truth, which is provided by human annotation and which may well incorporate errors. Finally, in order to assess the performance of each algorithm in terms of speed (computational efficiency) the time that was required for the completion of the segmentation task was recorded. The average elapsed time for each experiment is shown in table 1.

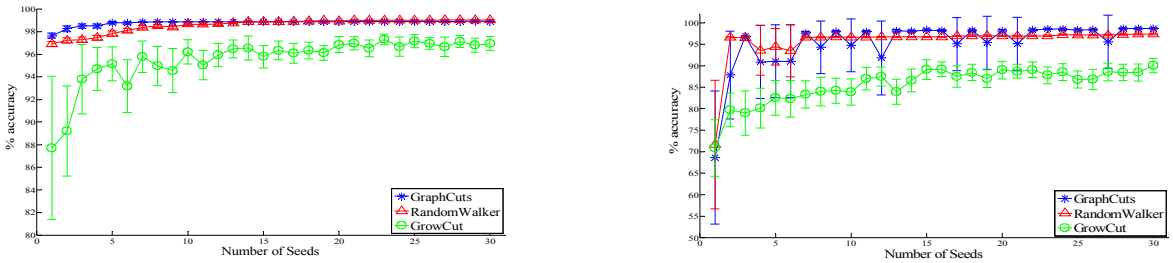

Figure 1: Diagrams of accuracy for variable number of seeds for seeds with "Random Clicks" (left) and "Careful Seed Selection" (right). The error bars represent the ±1.96 × (standard error) of the mean.

It can be seen in figure 1 that even for a small number of seeds, the accuracy of GraphCuts and RandomWalker is high. One should not interpret a high score in accuracy as an almost perfect segmentation. The overall score is biased towards large values by the large number of true negatives (correctly classified background voxels). When the number of seeds is increased, the accuracy of these two algorithms remains steady. GrowCut demonstrates larger variations and inferior accuracy to them. Lastly in the case of Careful Seed Selection the algorithms perform worse, possibly because the surface of the organs does not comprise a good seeds candidate due to partial volume effects. This will be further discussed in the repeatability assessment section.

The diagrams in figure 2 complement the information depicted in figure 1, as they provide the maximum distance of the segmented object surface from the ground truth surface. RandomWalker gives boundaries that demonstrate small distance from the ground truth boundaries, whereas GraphCuts provides some voxels as foreground, which have high distance from the ground truth boundaries. GrowCut provides a lot of erroneous segmentations that contribute to low accuracy but also to large maximum distance from the ground truth. This is due to leakages of the foreground in difficult areas, where background input seeds are not present and therefore the leakage cannot be prevented.

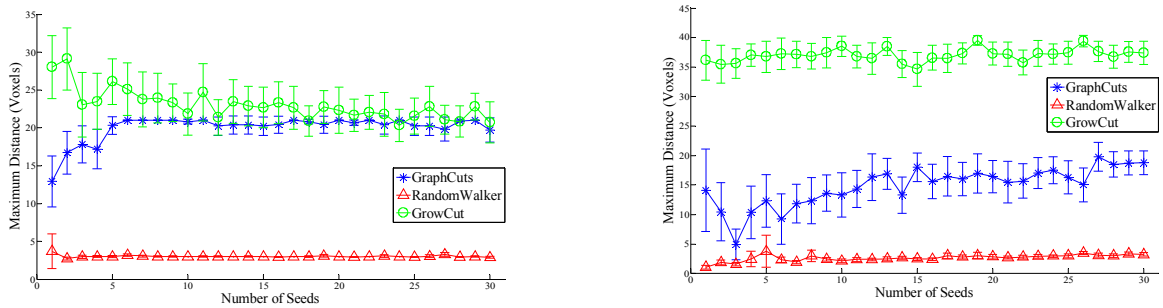

Figure 2: Diagrams of maximum distance from the ground truth surface for variable number of seeds with "Random Clicks" (left) and "Careful Seed Selection" (right). The error bars represent the ±1.96 × (standard error) of the mean.

Repeatability: During this experiment, the number of seeds was increased from 1 to 5 and then it was further increased using a step of 5. Then, for each number of seeds, the initial selected seeds were pertubed, as described in section 2, in order to allow for the accommodation of the user variability. Finally the relative overlap of the resulting pairs (36 pairs for each case of different number of seeds) was calculated according to equation 2. The results of this experiment are depicted in figure 3.

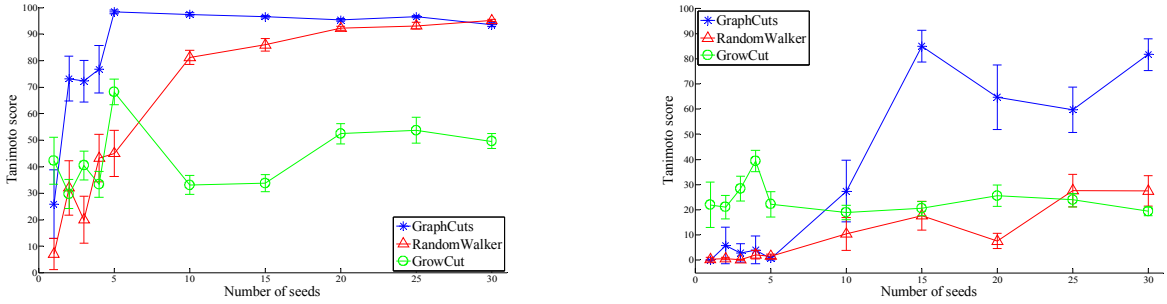

Figure 3: Diagrams of the overlap measure (tanimoto score) for variable number of seeds with "Random Clicks" (left) and "Careful Seed Selection" (right). The error bars represent the ±1.96 × (standard error) of the mean.

GraphCuts demonstrates very high repeatability even for a small number of seeds. Its performance is lower for very low number of seeds but increases when the latter increases (left diagram of figure 3). RandomWalker demonstrates a similar performance, although the minimum number of seeds demanded for repeatable results is higher. Also, although its performance increases with the number of seeds, it only becomes comparable to GraphCuts when 30 seeds are selected. The performance of GrowCut is lower than the other two methods. The overlap score is low (50%), even for a high number of seeds (20-30). This is due to uncontrollable region growing, when the pertubed seeds are placed to positions that promote leakage (partial volume effects). An interesting observation is the slight degradation of GraphCuts' performance when the number of seeds increase beyond 5. This may be caused by "invalid" seeds that belong to the surrogate of ground truth but not to the actual object of interest. The right diagram in figure 3, may support this argument, as it rather suggests that the algorithms fail to demonstrate a repeatable performance. Nevertheless, even in these circumstances, GraphCuts seems to achieve relatively repeatable results for seeds between 15 and 30. This failure mode probably occurs because some of the surface (boundary) voxels, which are suggested as foreground voxels by the surrogate of truth, may be background pixels in reality or just boundary voxels that do not share voxel intensity similarities with the inner part of the object. This could also explain the fluctuations of the performance of GraphCuts in the left diagram of figure 3; some of the seeds also come from the ground truth surface. If this surface consists of a number of "invalid" seeds, these seeds will affect negatively the segmentation. However, this would happen in a rather unpredictable (random) fashion, since the selection of seeds is a random process.

Efficiency: The performed experiments provide enough information to answer the questions regarding the algorithmic efficiency (section 3). Figure 1 shows that both GraphCuts and RandomWalker can achieve plausible segmentation, even with low number of input seeds. In addition, GraphCuts does not demand precise seed placement. In fact, the resulting segmentation is the same even for large variations of the seed placement. RandomWalker seems to possess this property as well but for a higher number of input seeds. GrowCut is not able to cope with alterations of the seed placement. Also, from the obtained results, it is questionable whether there is a critical number of seeds that will guarantee a good segmentation. Therefore, the cognitive load required by GrowCut is higher than the other two methods, since some seed initialisations are better than others. As a consequence, the user should spend more time and effort, in order to provide the algorithm with "good" input seeds or to correct inaccurate segmentation suggestions by the computational part of the technique. In terms of computational speed, among the implementations that we possess, GraphCuts is the most computationally efficient method, whereas RandomWalker is the most computationally expensive. GrowCut is slower than the former and faster than the latter. The time that was required by each method was relatively constant over a range of numbers of seeds. Table 1 summarises the recorded average segmentation time of the methods during the two variations of the accuracy experiment.

Method	**Average Time ± 1.96 × standard error (secs)**	**Average Time ± 1.96 × standard error (secs)**
GraphCuts	13.1 ± 2.7	28.3 ± 12.0
RandomWalker	975.6 ± 9.0	1013.6 ± 12.3
GrowCut	82.4 ± 5.5	91.5 ± 3.0

Table 1: Average time required by each algorithm for the segmentation task of the accuracy experiment with "Random Clicks" (middle) and "Careful Seed Selection" (right).

5 Conclusions

The experimental results presented in the previous sections, show that the suggested evaluation framework can assist towards the assessment of interactive segmentation algorithms in 3D. The obtained results are tractable and reproducible. Also, the simulated interaction seems to cover the most significant variations of human interaction. The experiments that have already been performed, provided useful information regarding the performance of the tested techniques with respect to accuracy, repeatability and efficiency. GraphCuts proved to be the most efficient method among the three that were assessed for the specific task assigned in our study. From the experiments performed, it was shown that the seed selection exactly from the object's surface is problematic. In order to verify this statement, the same experiments will be repeated by excluding the surface from being a candidate for foreground seed selection. This work will continue with further experiments that will reveal the different characteristics of these three methods.

Acknowledgements

This work is funded by the Biotechnology and Biological Sciences Research Council (BBSRC).

References

1. S.D. Olabarriaga & A.W.M. Smeulders "Interaction in the segmentation of medical images: A survey", *Medical Image Analysis*, vol.5, no.2, pp. 127-142, 2001.
2. J.K. Udupa et al. "A framework for evaluating image segmentation algorithms", *Computerized Medical Imaging and Graphics*, vol.30, no.2, pp.75-87, 2006.
3. G. Gerig, M. Jomier & M. Chakos "Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation", *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp.516-523, 2001.
4. D.W. Shattuck et al. "Online resource for validation of brain segmentation methods", *NeuroImage*, vol.45, no.2, pp.431-439, 2009.
5. Y.Y.Boykov & M.P. Jolly "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images", In *Proceeding of the Eighth IEEE International Conference on Computer Vision (ICCV)*, vol. I, pp.105-112, 2001.
6. Y. Boykov & V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 26, no. 9, pp. 1124-1137, 2004
7. L. Grady "Random walks for image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28, no. 11, pp. 1768-1783, 2006.
8. V. Vezhnevets & V. Konouchine ""Grow-Cut" - Interactive Multi-Label N-D Image Segmentation by Cellular Automata", In *Fifteenth International Conference on Computer Graphics and Applications (Graphicon-2005)*, Novosibirsk Akademgorodok, Russia, June 20-24, 2005.