A SYSTEMATIC PERFORMANCE EVALUATION OF INTERACTIVE IMAGE SEGMENTATION METHODS BASED ON SIMULATED USER INTERACTION

Emmanouil Moschidis, Jim Graham

Imaging Science and Biomedical Engineering, School of Cancer and Enabling Sciences, Stopford Building, The University of Manchester, Oxford Road, Manchester M13 9PT.
Emmanouil.Moschidis@postgrad.manchester.ac.uk, Jim.Graham@manchester.ac.uk

ABSTRACT

In this paper we report on the results of a systematic performance evaluation of three efficient image segmentation algorithms, namely Graph-Cuts, Random-Walker and Grow-Cut. The evaluation focuses on their function as the computational part of an interactive segmentation system. The implications caused by the human involvement in the overall process are avoided by simulating two different patterns of user interaction. The methods are evaluated with respect to accuracy, precision, efficiency and parameter sensitivity on three dimensional medical images. The results provide useful insight regarding the algorithmic performance of the selected techniques and the effect of the identified patterns of user interaction on the segmentation outcome.

Index Terms— Performance Evaluation, Interactive Image Segmentation, Simulated User Interaction, 3-D Medical Images, Graph-Cuts, Random-Walker, Grow-Cut.

1. INTRODUCTION

Human experts, such as radiologists, often segment large three dimensional (3-D) images as part of their clinical routine. While manual segmentation is tedious, subjective and inefficient, automatic methods rarely work as well as is required. Interactive image segmentation methods may be employed as a useful alternative. The recent advent of different efficient computational methods offers a large variety of candidates for the segmentation task.

Interactive segmentation techniques generally consist of three components: the Graphical User Interface (GUI), the interactive part and the computational part [1]. In this study we focus on the computational part, assessing three popular algorithms, namely Graph-Cuts [7], Random-Walker [8] and Grow-Cut [9]. Graph-Cuts is a graph-based method that considers an image as a flow network represented by a graph. The segmentation outcome is obtained when the edges of the graph carry the maximum possible flow. This also provides the cut on the graph, with the minimum cost among all candidates, which separates the seeds provided by the user. Random-Walker is a graph-based method that calculates the probability that a random walk, which initiates from any voxel of an image, will reach a seed specified by the user, given the bias that it cannot cross high image gradients. Finally, Grow-Cut uses iteratively heuristic rules based on cellular automata that define the labels in a voxel neighborhood. When the algorithm converges the final labeling provides the segmentation outcome.

These interactive algorithms share a widely used pattern of interaction, the brush-strokes. In the context of such an interaction, the expert selects certain groups of pixels that belong to a specific class by drawing scribbles with the mouse or a pointing device. In case of binary segmentation the classes are two, foreground and background. In this study we evaluate the three methods with respect to their ability to perform binary segmentation on 3-D medical images, given a controlled user input.

The brush as an interactive tool enables the user to mark large groups of pixels, which provides the algorithms with sufficient input to create plausible segmentations. However, for the purposes of an evaluation study this input should be minimal and controlled. This allows for a better observation of the algorithmic response of the assessed methods by revealing potential failure modes. Therefore, instead of brush-strokes we provide the algorithms with a small number of seeds (individual voxels). Furthermore, in order to focus the evaluation on the computational part of the methods, we provide this input via simulated user interaction. This allows for the exclusion of inconsistent human interaction from the evaluation framework.

The rest of the paper is organized as follows: In section 2 we present the evaluation framework of our study and the metrics associated with it. In section 3 we present our experiments and results, followed by a discussion in section 4.

2. THE EVALUATION FRAMEWORK

The evaluation framework of our study is thoroughly described in [6]. In this section we summarize its main aspects, which are the simulation of the user interaction and the metrics used for assessing the accuracy, the precision and the efficiency of the algorithms.

In our simulation we identify two main patterns of interaction as illustrated in Fig. 1. In the first pattern, the user is selecting seeds that are spread throughout the volume of the anatomy of interest and the background, simulating a general specification of foreground and background; we refer to these as Volume Seeds. In the second, the user selects seeds one voxel away on either side of the ground truth boundary of the anatomy of interest. This is intended to simulate careful local specification of the boundary, and is referred to here as Surface Seeds.

In the context of our study we use two sets of real medical images and two synthetic ones. Specifically these are: five 3-D (83×80×104) brain MR images, in which the task is to segment the brain ventricles, five 3-D (30×40×29) prostate MR images, in which the task is to segment the whole prostate and two simulated 3-D (227×172×11) brain MR images, one T1 and one Inversion
Recovery Turbo Spin Echo (IRTSE), in which the task is to separate the gray matter from the white matter. The ventricle and prostate datasets were chosen to represent a range of difficulty in segmentation. The ventricle images are relatively straightforward to segment, while the prostate images are challenging as they show considerable internal variation in gray level. The synthetic data provide known ground truth for a segmentation task (gray/white matter separation), in which manual delineation is likely to be unreliable, while they demonstrate realistic tissue properties. The surrogate of truth of the real images is provided by manual delineation of the anatomy of interest by one expert. The ground truth of the synthetic data was provided by the forward model of the segmentation algorithm described in [5].

Fig. 1: An illustration of the Volume Seeds (left) and the Surface Seeds selection pattern (right) on a coronal slice of a brain MR image. The foreground and background seeds are colored yellow and blue respectively.

The accuracy of segmentation, with respect to the surrogate ground truth, is evaluated as follows: all the voxels are classified into true and false positives (TP, FP) and true and false negatives (TN, FN). The metric for segmentation accuracy (A) is defined as:

\[
A = 100 \times \frac{TP + TN}{TP + TN + FP + FN} \%
\]  

(1)

The precision of the methods is assessed by measuring the effect of perturbation of the input seeds (see 3.2) on the segmentation result. For every pair of segmentations \( V_{S1}, V_{S2} \) the Tanimoto coefficient (Tc) [4] is calculated as:

\[
Tc = \frac{V_{S1} \cap V_{S2}}{V_{S1} \cup V_{S2}} = \frac{TP}{TP + FP + FN}
\]  

(2)

The efficiency of the interactive segmentation methods is related to the speed and the cognitive load of the overall process. The amount of interaction required for a plausible segmentation, the extent of the precision required during the input provision by the user and the computational speed are the three aspects of a method's efficiency. The first one, 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 aspect, an efficient algorithm does not demand precise pictorial input from the user (low cognitive load). Therefore, a repeatable method is also efficient. Lastly, the computational efficiency is usually reported in seconds.

In section 3.4 we investigate the sensitivity of these methods to operating parameters. The comparisons in sections 3.1-3.3 use optimal parameter settings.

Fig. 2: An axial slice, its ground truth and a 3-D segmentation example of a prostate MR image (top left), a brain MR image (bottom left), a T1 (top right) and an IRTSE (bottom right) synthetic brain MR image. The anatomy to be segmented is the prostate, the brain ventricles and the gray matter, respectively.

3. EXPERIMENTS AND RESULTS

3.1 Accuracy

In order to assess the ability of the algorithms to provide accurate segmentation outcomes, we varied the number of input seeds, both foreground and background, from 1 to 30. For each number of seeds 30 different random initializations were used. The results were averaged over the ventricles and prostate image sets. The first two rows of fig. 3 depict the results of the experiments for the identified patterns of interaction. Overall the graphs show that the accuracy of the segmentation outcome improves as the number of seeds increases. Also, Volume Seeds selection provided better results than Surface Seeds selection. In addition, apart from the prostate experiment in which Random-Walker demonstrated slightly higher accuracy, Graph-Cuts produced the most accurate results, generally reaching its best performance with a smaller number of input seeds than required for the other methods. Visual inspection of the segmentation volumes also revealed that Grow-Cut is prone to leakages when seeds are placed close to partial volumes, which leads to low accuracy due to increased number of false positive. Random-Walker on the other hand demonstrates a resistance to leakages, which generally is a desirable property. However, when seeds are placed in partial volumes, the algorithm tends to segment the partial volumes as a separate tissue, which reduces its accuracy due to increased number of false negatives.

3.2 Precision

In this part of the study we performed a variable seed displacement of 2 of one initialization per number of seeds, where \( i \in [0, 8] \); when \( i = 0 \) the seed is only displaced to its immediate neighbor, whereas when \( i = 8 \) it is displaced by 256 positions. All the possible values of \( i \) correspond to 9 perturbations and equal number of segmentation outcomes. All the possible pairs (36 in total) of segmentations, were compared with respect to their overlap (eq. 2). For each seed initialization the results were averaged. The last two rows of Fig. 3 depict the results of the experiments. Volume Seeds selection provided better results than Surface Seeds selection in this experiment as well. Graph-Cuts demonstrated high robustness to imprecise input on most datasets. Random-Walker performed slightly better on the prostate dataset and overall was quite robust to imprecise input, although the input required for reaching results comparable to Graph-Cuts was higher. Grow-Cut demonstrated limited robustness to alterations of the seed placement.
Fig. 3: Accuracy (a-h) and Precision (i-p) for Graph-Cuts, Random-Walker and Grow-Cut as a function of the number of input seeds on Ventriciles (1st column), Prostate (2nd column), T1 (3rd column) and T2 (4th column) simulated images for Volume Seeds (1st and 3rd row) and Surface Seeds (2nd and 4th row). The error bars represent the ± 1.96 × standard error of the mean.

3.3 Efficiency

In terms of computational speed, among the implementations that we possess, Graph-Cuts is the most computationally efficient method, whereas Random-Walker is the most computationally expensive. Grow-Cut is slower than the former, but much faster than the latter. Random Walker's and Grow-Cut's shortest execution times were around 10 times and 1.5 times longer than Graph-Cuts' execution time respectively.

In terms of cognitive load, all algorithms can equally provide plausible results even with a few input seeds. Also, Grow-Cut's limited robustness to imprecise input signifies that some seed initializations are better than others. Consequently, the user should spend more time and effort, in order to provide the algorithm with “good" input seeds or to correct inaccurate segmentation suggestions. Therefore, its cognitive load is higher than the other two methods, which demonstrate high robustness to imprecise input.

3.4 Sensitivity to parameters

Since Grow-Cut is a non parametric method, this section concerns only Graph-Cuts and Random-Walker. Graph-Cuts was used without a regional term. Therefore, the only free parameter of both methods is the one that controls their Gaussian weighting function. This function assigns certain weights in the weighted graph based on the intensity differences of neighboring voxels. The weighted graph provides an enhanced data representation; therefore its construction can be seen as an implicit preprocessing step.

In order to assess how σ and β affect the segmentation outcome, we varied them for 30 different random initializations of 30 input seeds. A step of 5 was selected for the alteration of σ, whereas the equivalent step for β was chosen as 50 (2σ²). The results (fig. 4) show that Graph-Cuts demonstrates slightly higher sensitivity than Random-Walker to its parameter. Small changes of σ can cause great changes to the segmentation accuracy for Graph-Cuts, whereas Random-Walker responds more smoothly to alterations of
Values for $\sigma$ and $\beta$ that provide the best possible accuracy are found within the intervals [5,50] and [50,500] respectively. It was also seen that changes of the value of the free parameter cause alteration to the time required for the completion of the segmentation task. However, only in Random-Walker was a strong correlation between $\beta$ and the time required for the segmentation task observed. Generally in our experiments higher $\beta$ values caused longer execution times.

![Graph-Cuts and Random-Walker](image)

**Fig. 4:** Accuracy of Graph-Cuts (a,c) and Random-Walker (b,d) segmentations for different values of their free parameter for Volume Seeds (a,b) and Surface Seeds (c,d) on the datasets used in our study.

In Random-Walker's 3-D implementation, we used the preconditioned conjugate gradients method. Since iterative solvers give a solution within a desired limit of accuracy, we observed the effect of the solver's accuracy on the segmentation outcome. For values $10^{-5}$ the segmentation outcome remains constant, whereas for values less than $10^{-5}$ it degrades heavily.

### 4. DISCUSSION

The experiments performed in our study led to a couple of interesting observations; a major observation is the fact that generally Volume Seeds selection enables the algorithms to provide more accurate and more repeatable segmentations than the Surface Seeds selection. We believe that this happens due to the partial volume effects which are prominent in medical images. If the user selects only voxels close to the boundaries of the anatomy of interest, s/he may select partial volume voxels, which may be recognized as a third separate tissue. Therefore, information from the internal of the organ to be segmented is important for the segmentation task.

In the context of our evaluation, Graph-Cuts proved to be more efficient than its competitors in terms of computational speed and cognitive load. Random-Walker demonstrated robustness against leakages towards partial volumes, which should not be overlooked, since this feature can be valuable in cases where strong edges are not present, a situation which would suggest a failure mode for Graph-Cuts that detects the lowest cost surface in a 3-D fashion. Grow-Cut demonstrated higher sensitivity to imprecise input than the other methods, which leads to higher user cognitive load.

Lastly, the implicit preprocessing step of the graph-based methods, which incorporates the construction of the weighted graph is important for both Graph-Cuts and Random-Walker. Random-Walker is however more robust to the alteration of its parameter.

### 5. ACKNOWLEDGMENTS

We would like to thank Leo Grady for source code and useful insight regarding the implementation of Random-Walker in 3-D and Neil Thacker for the synthetic images used in our study. The project is funded by the Biotechnology and Biological Sciences Research Council (BBSRC).

### 6. REFERENCES


