A general framework for evaluating interactive image segmentation algorithms

Bingjie Jiang* and Tonwei Ren

Springer-Verlag, Computer Science Editorial,
Tiergartenstr. 17, 69121 Heidelberg, Germany
{alfred.hofmann,ursula.barth,ingrid.haas,frank.holzwarth,
anna.kramer,leonie.kunz,christine.reiss,nicole.sator,
erika.siebert-cole,peter.strasser,lncs}@springer.com
http://www.springer.com/lncs

Abstract. The abstract should summarize the contents of the paper and should contain at least 70 and at most 150 words. It should be written using the *abstract* environment.

Keywords: We would like to encourage you to list your keywords within the abstract section

1 Introduction

Interactive image segmentation has been extensively studied in the latest decade. Many state-of-the-art algorithms in this field have been proposed, starting from Boycov et. al[3],followed by Grabcut[9],Random Walker[5],Bai and Sapiro [2] and [6]. However, when it comes to the evaluation of these algorithms, the comparison can hardly be objective due to different human interferences. As is often the case, interactive image segmentation algorithms are tested upon user scribbles provided by the specific author. In this way, the performance of segmentation result could heavily depend on certain batch of seeds selection, rendering the result not convincing enough when compared with other algorithms.

This paper deals with the problem of evaluating interactive segmentation algorithms in an objective and comprehensive way. The contribution of this paper includes: (1) Analysis of differences between user labels in a quantitative way. By clustering colors in the image, we figured out the correspondence between user labels and color clusters. (2)...

The remainder of this paper is organized as follows:...

* Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

2 Related Work

******

3 User-interaction differences

3.1 Dataset design

The dataset contains 96 images from publicly available Berkeley Segmentation Dataset[7]. These images are selected so that each of them contains at least one obvious object which could be unambiguously explained to participants.These images are representative of some major challenges of image segmentation, including fuzzy boundary, complex texture and complex lighting conditions. Ground truths are precisely hand-labeled for each image in order to avoid any bias.

3.2 Experiment

In this section we will discuss the design of experiment. We use the software provided by The K-Space Segmentation Tool Set, [8]. Screenshots of the tool are shown in Figure 3.2. The 5 participants are all students from computer science background but have limited knowledge in interactive image segmentation. Each participant was given a clear guidance and enough time to familiarize themselves and become proficient with the software that would be used for the experiment. Sample markers were also provided in avoid of misunderstanding . Then in real experiment, each participants are provided with 96 images and the corresponding ground-truth which tells exactly which object to extract. However, we hide the segmentation result from user so that they will not realize if they have provided a "good" mark or not. We also confined the time for labeling each image. In this way, we manage to (1)limit the effort of participants to draw scribbles in consideration of real-life application.(2)obtain the most natural response of users rather than inputs guided by segmentation result.

3.3 User-interaction differences

In our person-oriented experiments, great differences among user labels were observed. It comes to us instinctively that different people tend to consider different part of foreground and background object as salient. Under this guidance, we processed the 5 marker files of each image out of the dataset and calculated the pair-wise intersection degree. The result is shown in table 1and table 2. The intersection degree is defined as follows:

$$intersection\ of\ marked\ pixels / union\ of\ marked\ pixels$$

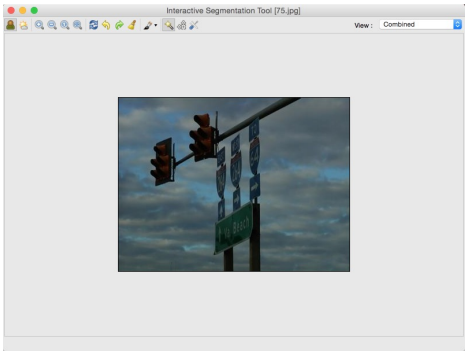


Fig. 1. The screenshot of k-space segmentation tool

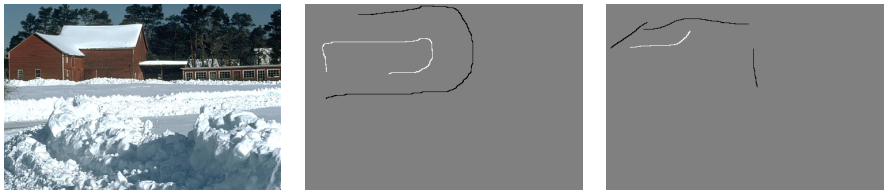
Table 1 visualizes the value on foreground and Table 2 the background. We make the following observations: (1)The average value of intersection/union between two people is rather low, ranging from 1.6592% to 3.5195% in foreground and 0.0905% to 0.9700% in background, which indicates that markers of different people have little common in common.Figure 2 shows an example of user markers which share no common pixel (2) People share less similarities in background markers than foreground. It is because the foreground objects are confined to a certain area while the backgrounds are more extensive.

	seed 1	seed 2	seed 3	seed 4	seed 5
seed 1	100%	2.3768%	3.5195%	2.0146%	1.6592%
seed 2	2.3768%	100%	2.7080%	2.1684%	2.2459%
seed 3	3.5195%	2.7080%	100%	2.0485%	2.0102%
seed 4	2.0146%	2.1684%	2.0485%	100%	1.8441%
seed 5	1.6592%	2.2459%	2.0102%	1.8441%	100%

Table 1. intersection degree of foreground

	seed 1	seed 2	seed 3	seed 4	seed 5
seed 1	100%	0.8700%	0.8600%	0.0905%	0.6300%
seed 2	0.8700%	100%	0.9700%	0.2100%	0.7300%
seed 3	0.8600%	0.9700%	100%	0.1600%	0.6800%
seed 4	0.0905%	0.2100%	0.1600%	100%	0.6700%
seed 5	0.6300%	0.7300%	0.6800%	0.6700%	100%

Table 2. intersection degree of background



(a) source image (b) label 1 (c) label 2

Fig. 2. Two labels of the same image

3.4 Segmentation differences

Based on the above result, we took a step further by examine the segmentation results of different markers.The four interactive segmentation algorithms [3][5][2][6]in Table 3 was used. Based on the 96*5*4 segmentation results, we have found out that due to the variance between different markers, the segmentation results tend to be quite different, too. We achieved here by calculating percent of pixels which was segmented out as foreground simultaneously by 1 person, two people, three people ,four people and 5 people out of all the segmented pixels. The result is shown in Table 4 5 6 7 . We have observed that: (1) The effectiveness of different markers as input of interactive segmentation algorithms vary from person to person. It could also be observed that the seed made by participant 1 have a better performance than others in terms of average, maximum and minimum, which means there exists kind of "professional lables" [4] (2) On average level, the result reaches at most 66.760% in table 5 made by participant 1. While the maximum ratio could be high as 99.050%, the minimum ratio was also low as 1.060%. This shows that when applied to segmentation algorithms, the variance of markers could further lead to quite great difference in segmentation results.

Algorithm	Description
BJ	Boykov Jolly Graph cut
GSC	Boykov Jolly with Geodesic Star-Convexity
SP	Bai and Sapiro
RW	Random Walker

Table 3. intersection degree of background

4 Automatic user interaction simulation

4.1 Quantitative of images

In interactive image segmentation,users are expected labels some key features of the image foreground and background. The effectiveness of these labels is greatly

intersection scale	1	2	3	4	5
average	66.360%	55.090%	52.300%	50.710%	49.620%
max	99.020%	98.580%	98.370%	98.200%	98.060%
min	21.060%	1.240%	1.180%	1.160%	1.140%
variance	0.0557	0.0883	0.0897	0.0899	0.0898
coefficient of variance	0.3556	0.5394	0.5727	0.5913	0.6039

Table 4. bj

intersection scale	1	2	3	4	5
average	66.760%	55.490%	52.490%	50.600%	49.170%
max	99.040%	98.500%	98.170%	97.900%	97.660%
min	21.130%	1.230%	1.100%	1.080%	1.060%
variance	0.0518	0.0809	0.0813	0.0812	0.0812
coefficient of variance	0.3409	0.5126	0.5432	0.5632	0.5795

Table 5. gsc

intersection scale	1	2	3	4	5
average	56.390%	41.030%	36.740%	34.350%	32.660%
max	96.040%	94.290%	93.470%	92.850%	92.260%
min	22.950%	2.920%	2.270%	2.030%	1.910%
variance	0.0286	0.0431	0.0418	0.0405	0.0393
coefficient of variance	0.2999	0.5060	0.5565	0.5859	0.6070

Table 6. sp

intersection scale	1	2	3	4	5
average	63.180%	49.860%	45.720%	43.180%	41.410%
max	98.660%	97.910%	97.430%	97.010%	96.600%
min	23.010%	3.190%	2.580%	2.190%	2.000%
variance	0.0295	0.0440	0.0432	0.0433	0.0436
coefficient of variance	0.2719	0.4207	0.4546	0.4819	0.5042

Table 7. rw

related to their coverage of foreground and background key components. So we decide to simulate markers by simulating their coverage of key pixel clusters. In our method, we used the simple linear iterative clustering (SLIC) method [1] implemented by VLFeat open source library[10]. The SLIC algorithm clusters pixels in the combined five-dimensional color and image plane space which brings compact, nearly uniform superpixels. We then decompose all the superpixels into groups according to the quantitative feature of each superpixel. In this way, we could analyze user markers in superpixel level and superpixel group level.

4.2 Quantitative simulation of user labels

Instead of continuous user scribbles, our simulation is targeted at generating several points to represent key features in the image. So our first step was to validate the effectiveness of points in segmentation as compared to continuous lines. After figuring out the superpixel groups which were covered by user markers, we randomly select one superpixel from these groups and then randomly select one point from that superpixel, in this way, we formed a point set which contains points from all groups covered by a certain artificial marker. We then applied five sets which were generated the same way to segmentation algorithms. In terms of evaluation of segmentation accuracy, we define the following criteria:

$$\begin{aligned} F_{gt} &= \text{foreground in ground truth} \\ F_s &= \text{foreground in segmentation} \\ \text{recall} &= \frac{F_{gt} \cap F_s}{F_{gt}} \\ \text{precision} &= \frac{F_{gt} \cap F_s}{F_{gt}} \end{aligned} \tag{1}$$

The accuracy evaluation of five batches of markers and five batches of auto-generated seeds applied to the four algorithms are shown in the following tables(To be done).

	seed 1	seed 2	seed 3	seed 4	seed 5
Precision	0.8595	0.8423	0.8449	0.8543	0.7474
Recall	0.8822	0.9158	0.9154	0.9303	0.9108
	simu-seed 1	simu-seed 2	simu-seed 3	simu-seed 4	simu-seed 5
Precision	0.7659	0.7850	0.7500	0.7327	0.7997
Recall	0.7101	0.7339	0.7780	0.7740	0.7337

Table 8. bj

5

	seed 1	seed 2	seed 3	seed 4	seed 5
Precision	0.8671	0.8619	0.8560	0.8747	0.7562
Recall	0.8794	0.9177	0.9137	0.9183	0.9129
	simu-seed 1	simu-seed 2	simu-seed 3	simu-seed 4	simu-seed 5
Precision	0.7885	0.7823	0.7733	0.7443	0.7856
Recall	0.7371	0.7366	0.7939	0.7813	0.7403

Table 9. gsc

	seed 1	seed 2	seed 3	seed 4	seed 5
Precision	0.7591	0.7672	0.7672	0.8496	0.7057
Recall	0.8837	0.9109	0.9019	0.9279	0.8759
	simu-seed 1	simu-seed 2	simu-seed 3	simu-seed 4	simu-seed 5
Precision	0.7716	0.7781	0.7556	0.7716	0.7766
Recall	0.8347	0.8477	0.8568	0.8515	0.8438

Table 10. rw

	seed 1	seed 2	seed 3	seed 4	seed 5
Precision	0.6471	0.6567	0.6534	0.7069	0.6811
Recall	0.9306	0.9501	0.9454	0.8991	0.9015
	simu-seed 1	simu-seed 2	simu-seed 3	simu-seed 4	simu-seed 5
Precision	0.6722	0.6995	0.6802	0.6631	0.7153
Recall	0.9251	0.9221	0.9331	0.9249	0.9189

Table 11. sp

	all-group-seed-1	all-group-seed-2	all-group-seed-3
Precision	63.180%	49.860%	45.720%
Recall	98.660%	97.910%	97.430%

Table 12. rw

Bibliography

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels. Technical report, 2010.

[2] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[3] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 105–112. IEEE, 2001.

[4] Yu Fu, Jian Cheng, Zhenglong Li, and Hanqing Lu. Saliency cuts: An automatic approach to object segmentation. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[5] Leo Grady. Random walks for image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1768–1783, 2006.

[6] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman. Geodesic star convexity for interactive image segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3129–3136. IEEE, 2010.

[7] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001.

[8] Kevin McGuinness and Noel E O'Connor. The k-space segmentation tool set. 2008.

[9] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, 2004.

[10] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.