A MODEL FOR SIMULATING USER INTERACTION IN HIERARCHICAL SEGMENTATION

Bruno Klava and Nina S. T. Hirata

Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo, Brazil

ABSTRACT

Evaluation of interaction related aspects is an important step for comparing interactive segmentation algorithms and also for the development of new and better algorithms. However, as experiments with users are not simple, simulations with robot users are becoming a common practice for evaluating marker based interactive segmentation methods. We propose a novel interaction model for hierarchy based segmentation methods. Simulations of the model with some proposed policies are compared to experiments with real users in order to evaluate and validate the proposed model.

***Index Terms*—** interactive segmentation, hierarchical segmentation, user interaction simulation

1. INTRODUCTION

Interactive image segmentation is an intermediary approach between automatic and manual approaches, coping with the inability of automatic approaches to generate correct results and also with the excessive workload of purely manual approaches. In an interactive approach, users guide the segmentation process through a set of interaction operations. A group of such operations consists in placing markers on the objects or their contours, which are then used as region seeds or as anchor points to delineate region contours [1, 2, 3]. Another group of operations consists in merging or splitting regions to achieve a desired result [4], common in hierarchical segmentation, where an image partition in a stack of nested partitions can be quickly selected and individual regions can be merged or split in order to refine the selected partition [5, 6, 7, 8]. Hierarchies are suitable when segmentations with different levels of details are desired or when there are too many objects to be segmented.

An important characteristic in interactive segmentation is to allow users to achieve the desired segmentation results with minimum effort. Thus, both segmentation results [9, 10] and issues related to user effort [11, 12] should be taken into consideration for the evaluation of these methods. As experiments involving users are not simple, a frequently used approach to evaluate interactive methods is to perform simulations with robot users that mimic real users. While using robot users allows automatic evaluation of interactive segmentation methods, it requires a previous validation of the robot interaction model.

In [11] a model for simulating user interaction for segmentation algorithms based on markers is proposed and validated. In the proposed model, markers are successively generated according to a probabilistic distribution that is based on the segmentation error at each step. The use of robot users to evaluate marker based interactive segmentation systems and algorithms seems to be already a common practice in the literature [13, 14].

With respect to hierarchical segmentation, most hierarchies usually do not contain partitions that are close to the desired segmentation and therefore additional region splitting and merging operations are often used. To the best of our knowledge, there are no reports on interaction models for hierarchy based segmentation. One of the contributions of this work, presented in Section 2, is an interaction model based on policies that determine at each step of the segmentation process which regions are to be merged or split.

To validate the proposed model, we performed experiments with users as detailed in Section 3. Although several aspects related to interaction are important, investigation of any cognitive aspects of interaction is beyond the scope of this paper. We are interested in the sequence of operations performed by the users. However, explicit comparison between sequences of operations from different simulations are not practical because they vary greatly in terms of number and types of operations (due to variation in image content and differences among users). Thus, in order to relate model and user simulations, we compute quality measures for each of the partitions generated during the simulation, and represent each simulation as a time series that expresses the evolution of the measure throughout the segmentation process. Although the evolution of these measures can not model the sequence of operations performed by the users, we assume that in a certain degree it reflects the choices made by the users. Thus by analyzing the evolution profile, we may correlate distinct simulations. We analyze and discuss the simulations from this perspective in Section 4, and present our conclusions in Section 5.

Thanks to FAPESP (2009/16852-0, 2011/50761-2), CNPq (484572/2013-0), CAPES, NAP eScience - PRP/USP for funding.

2. INTERACTION MODEL

A hierarchy in the context of image segmentation is a nested set of partitions of an image, with the finest partition at the bottom and the coarser one, consisting of the partition with only one region (the whole image domain), at the top. A partition at a given level t is built by merging two adjacent regions of the partition at level $t - 1$. These types of hierarchies can be represented by a binary tree, where the leaves correspond to the atomic regions in the finest partition, and an internal node corresponds to the region obtained by merging the regions corresponding to its child nodes.

Given a hierarchy of partitions represented as a binary tree, we assume that one can apply the following operations [5, 7]:

- **global threshold operation:** selects a level of the hierarchy (in watershed hierarchies, it corresponds to selecting a threshold for the metric used in the hierarchy construction);
- **local splitting:** splits a selected region in the two regions that were merged when building the hierarchy;
- **local merging:** merges the selected region with all the regions under its parent region in the hierarchy;
- **manual merging:** allows the user to select adjacent regions to be merged, independent of the structure imposed by the metric used to build the original hierarchy.

Note that the local merging operation, contrary to the manual merging that allows users to specify which regions should be merged, is a blind one in the sense that a user has no ways to know to which of the adjacent regions the selected region will be merged. Hence, in our model the local merging operation is not considered. Regarding the threshold operation, it acts globally in the hierarchical structure and can undo the effect of previous operations. Thus a thresholding should be used only at the beginning of the process. Then, in our model we consider that the threshold operation can be used at most once, as the first operation in the segmentation process.

Supposing that the contour of a desired segmentation (hereafter, ground truth) is a subset of the contours in the finest partition, at any step of the segmentation process one of the following holds for each region: (1) it corresponds to a region of the ground truth, so it should not be affected by posterior operations, (2) it intersects with more than one region of the ground truth, so it should be split in a posterior operation, and (3) it is properly contained in a single region of the ground truth so it should be merged afterwards with adjacent regions also in the same condition.

Thus, in each step of the simulation, the following operations can be considered: (i) any region of type 2 can be split into two regions by using the local splitting operation and (ii) any two or more regions of type 3 that are adjacent to each other and contained in the same ground truth region can be merged together using the manual merging operation. The simulation ends when there are only regions of type 1, that is, when the current partition corresponds to the ground truth.

To fully define the simulation model, we propose three policies to establish the order in which the local splitting and the manual merging operations are performed:

- **Policy 1:** perform all the necessary splittings before any mergings. In this case, the number of merging operations is minimized, as it is necessary to perform at most N of these operations, where N is the number of regions in the ground truth. Note that the interaction effort of each one of these operations can be high, as it could be necessary to select a lot of small regions to be merged together.
- **Policy 2:** perform a manual merging as soon as it is possible, executing a local splitting only when there are no regions that can be merged together. In this case, the number of performed merging operations is maximized. Nevertheless, the interaction effort of each one of these operation tends to be small.
- **Policy 3:** if a merging is possible, execute it according to a probability function $\mathcal{P}$ given by the ratio between the size (in pixels) of the regions to be merged and the total image size. This probability function privileges the formation of more general structures, postponing the refinement of details, in a coarse to fine strategy.

Policies 1 and 2 can be viewed as particular cases of policy 3, for $\mathcal{P} = 0$ and $\mathcal{P} = 1$, respectively. Policy 3 is a trade-off between policies 1 and 2 and should better represent the overall behavior of the users, which is not deterministic.

3. SIMULATIONS

To cover diverse levels of segmentation complexity, a set of 20 segmentation tasks was designed, considering images ranging from microscopy to natural scenes photography. The ground truths were created so as to contain only contours present in the finest partition of the hierarchy, to guarantee that it can always be achieved in the hierarchy. Moreover, the ground truth was not generated using the hierarchy to avoid it being easily found. Besides that, image sizes were restricted to up to 800×600 pixels, sufficient to require several steps of interaction but not overly time demanding. The images are available online [15].

A particular hierarchy may or may not favor a specific segmentation goal depending on how it was built, and it is possible that in some cases a segmentation can be straightforwardly accomplished using a marker based approach (as is the case of some tasks in our dataset). Nevertheless, since the goal of the experiment is not to evaluate hierarchies, but aspects related to user interactions on a hierarchy, we used watershed hierarchies built based on the volume criterion [16] in all the tasks.

3.1. User experiments

The user experiments were performed using the segmentation tool `SegmentIt` [17], that implements the hierarchical operations listed in Section 2, with the following features added to support the experiments: **(a)** a video demonstrating its usage and a test image so that volunteers could test and get acquainted with the hierarchical operations; **(b)** an online scheduling system that selects a segmentation task aiming to have each task executed by a similar number of volunteers; **(c)** the possibility for the volunteer to pause the experiment after executing one segmentation task and then continuing to execute other tasks later; **(d)** no time restriction imposed to execute each task; **(e)** automatic sending of the data related to the execution to a server for posterior analysis after each segmentation task is executed; **(f)** a textual description of the segmentation result to be achieved. A thumbnail of the ground truth is also made available during the experiment to clear up any doubts about the textual description (showing the ground truth does not compromise the experiment, as the interest is not to evaluate the final quality measures achieved by the volunteers, but how it varies throughout the process; also, showing a thumbnail of the ground truth allows the volunteers to apply any desired operations until he/she is confident about having obtained a partition close to the ground truth, and this avoids the need to determine a stop criterion for the robot users, for which the simulation process is run until the actual ground truth is achieved).

Volunteers were invited to participate in the experiments and they were instructed to segment as much images as they desired, not being required to perform all the 20 tasks. A total of 15 volunteers collaborated with the experiment, resulting in 200 segmentation tasks being executed.

3.2. Robot user experiments

For simulating each of the policies described earlier, an important issue is how to establish the order in which regions are processed under each policy. Users usually adopt a strategy such as segmenting one object at a time, following a coarse to fine approach, or other strategies that depend on image content. Hence, the order in which regions are selected to be split or merged is not arbitrary. For the model simulation, we need to consider a processing order that reflects this non arbitrary behavior. For that end, regions of type 2 (those that should be split) are processed following the order of a queue. Using a queue, regions will somehow be processed in the order they appear in the partition and also those spatially close to a previously processed region tend to be processed earlier. For the regions of type 3 (those to be manually merged), we use a hash map structure that imposes no ordering. In fact, under policy 1, the order in which merges are performed is irrelevant since the merge operations are performed in the last steps only; under policy 2, a region is merged as soon as it is classified as a region of type 3; and under policy 3, the regions to be merged are determined by the probability function.

Policies 1 and 2 were simulated once for each task, and policy 3, which is probabilistic, was simulated 10 times for each task to reflect distinct users. Since the result of a threshold operation is equivalent to the result of a sequence of local operations, thresholding was not included in the robot user simulations.

4. ANALYSIS AND DISCUSSION

As quality measures of the generated partitions, we computed accuracy (Rand index), precision, and recall, three measures that are commonly used for comparing partitions [18]. In the context of image segmentation, these measures are usually computed examining pairs of pixels [10] whereas here we compute them for pairs of atomic regions.

These measures were computed for each simulation. Thus, the total number of time series for each measure (accuracy, precision and recall) is 440, corresponding to 200 executions of segmentation tasks by human users plus 240 simulations with robot users. Among these 240 robot user simulations, 40 correspond to simulations using the deterministic policies 1 and 2 (one for each of the 20 tasks), and 200 correspond to 10 simulations using policy 3 for each of the 20 tasks.

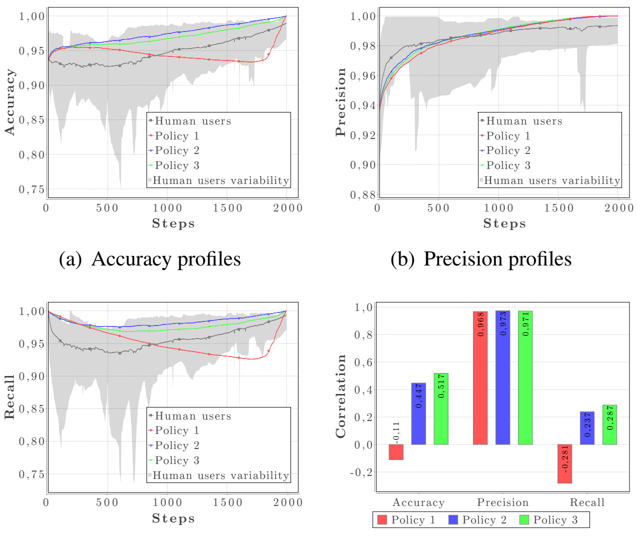
Since the time series have different lengths, we first expanded them linearly to have the same number of points. Then, for each measure, the expanded time series were separated in four groups (one for each of the three policies of robot users plus one for human users) and then averaged by group for each of the measures, generating the four profiles for each measure used in the analysis.

In the experiments, some operations that are not included in the robot user simulations (such as thresholding, undo or redo operations) were available for the users and they are present in the user time series. As these operations may result in characteristics not present in the robot time series, we considered filtering these operations from the user time series. However, as the only difference between filtered and non filtered time series were small fluctuations that did not affect the general profile evolution, in the analysis we consider user time series with no filtering.

4.1. Model evaluation

The obtained profiles are shown in the charts (a) - (c) of Figure 1. Note that the robot users always reach the maximum values for the measures, as the final result is always equal to the ground truth partition. For this reason the robot users usually perform a higher number of operations than the human users.

The accuracy and recall profiles of policy 1 are the ones that most differ from the others. This is due to the fact that under policy 1 all manual merging operations are performed



(a) Accuracy profiles

(b) Precision profiles

(c) Recall profiles

(d) Correlation

Fig. 1. Profiles computed from the experiments relative to each proposed policy and to the human users, for each of the three measures considered: (a) accuracy, (b) precision and (c) recall. The correlation of each measure between human user profiles and each proposed policy profiles are shown in (d).

at the end of the process, causing a high number of false negatives (regions that are disconnected in the current partition but should be connected, according to the ground truth) and a low number of true positives (regions connected in the current partition and in the ground truth) before the mergings are performed. For this same reason, these time series ascend abruptly at the end of the process. All the other profiles have a smoother behavior. In particular, policies 2 and 3 are visually similar to each other, and human user profiles also have an overall evolution that is similar to those two policies profiles.

To quantify the similarities, we computed the correlation (Pearson product-moment correlation coefficient) [19] between robot user and human user profiles, which is shown in Figure 1(d). The correlation between the precision profile of each of the three policies and the human user profile are very high. On the other hand, the correlation for accuracy and recall profiles are not that strong. However, notice that the correlation is positive for policies 2 and 3, whereas it is negative for policy 1.

Although the correlations between the polices profiles and the human users profiles are not that strong, note that the policies profiles falls mostly within the variability of the human users time series, shown in the filled areas of the charts (a) - (c) of Figure 1. The bounds of these areas, for each of the measures, are defined by the user average time series, computed for each of the users that took part in the experiments.

Regarding the probability function $\mathcal{P}$ used in policy 3, we simulated it for a range of constant values to examine whether the operations order matter. Ranging $\mathcal{P}$ from 0 to 1 makes the resultant profile shift from policy 1 to policy 2 profiles, as shown in the chart of Figure 2. Although the correlations of the constant probabilities profiles were close to the correlations of the profile of policy 3 [15], the proposed function $\mathcal{P}$ is more interesting since it depends on the image content.

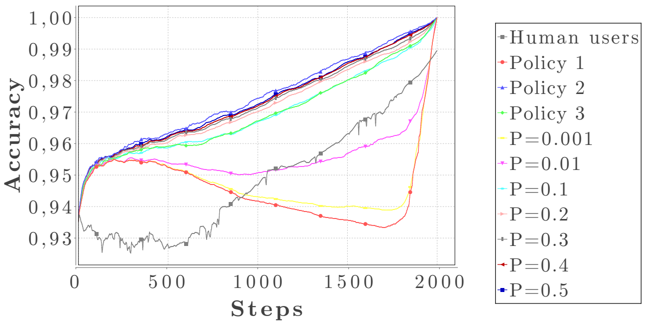


Fig. 2. Policy 3 profiles in a range of constant values for $\mathcal{P}$.

5. CONCLUSIONS

We have presented three policies to model user interaction with respect to the sequence of performed operations in a hierarchical segmentation context. After analyzing the segmentation accuracy, precision and recall profiles of the three policies and also of the users, we conclude that policies 2 and 3 present similar performance and are better related to user profiles than policy 1. Although not strongly correlated to human users, profiles of policies 2 an 3 falls within the human variability range and therefore could be used to automatically evaluate more complex experimental scenarios (for instance, experiments involving high resolution images, or on large set of images, or with different types of hierarchies). We consider that policy 3 is preferable because it is non deterministic, better reflecting the nature of choices made by human users during an interactive segmentation process. To convert these profiles to an estimation of user effort, a function that takes into consideration, for instance, the number and type of operations employed, could be computed.

To perform the experiments for evaluating the proposed model, we have used a tool that implements watershed hierarchies with the described operations. However, the model can be used with other types of hierarchies and other interaction operations can be included. The proposed benchmark, using the same quality measures or with new measures, remains valid for new scenarios.

We believe that the proposed model will contribute to the development of algorithms for building hierarchies that are semantically more expressive, and therefore that will require less interaction effort.

6. REFERENCES

[1] S. Beucher and F. Meyer, *Mathematical Morphology in Image Processing*, chapter The Morphological Approach to Segmentation: The Watershed Transformation, pp. 433–481, Marcel Dekker, 1993.

[2] Eric N. Mortensen and William A. Barrett, “Interactive segmentation with intelligent scissors,” *Graphical Models and Image Processing*, vol. 60, no. 5, pp. 349 – 384, 1998.

[3] Andrew Blake, Carsten Rother, M. Brown, Patrick Pérez, and Philip H. S. Torr, “Interactive image segmentation using an adaptive GMMRF model,” in *ECCV (1)*, Tomás Pajdla and Jiri Matas, Eds. 2004, vol. 3021 of *Lecture Notes in Computer Science*, pp. 428–441, Springer.

[4] Jifeng Ning, Lei Zhang, David Zhang, and Chengke Wu, “Interactive image segmentation by maximal similarity based region merging,” *Pattern Recogn.*, vol. 43, no. 2, pp. 445–456, Feb. 2010.

[5] F. Zanoguera, B. Marcotegui, and F. Meyer, “A toolbox for interactive segmentation based on nested partitions,” in *Proceedings of the International Conference on Image Processing*, 1999, vol. 1, pp. 21–25.

[6] Joshua E. Cates, Ross T. Whitaker, and Greg M. Jones, “Case study: an evaluation of user-assisted hierarchical watershed segmentation,” *Medical Image Analysis*, vol. 9, no. 6, pp. 566–578, 2005.

[7] Bruno Klava and Nina S. T. Hirata, “Interactive image segmentation with integrated use of the markers and the hierarchical watershed approaches,” in *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, February 2009, vol. 1, pp. 186–193.

[8] Michael Gerstmayer, Yll Haxhimusa, and Walter G. Kropatsch, “Hierarchical interactive image segmentation using irregular pyramids,” in *Proceedings of the 8th international conference on Graph-based representations in pattern recognition*. 2011, GbRPR’11, pp. 245–254, Springer-Verlag.

[9] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert, “Toward objective evaluation of image segmentation algorithms.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 929–944, 2007.

[10] Jordi Pont-Tuset and Ferran Marqués, “Measures and meta-measures for the supervised evaluation of image segmentation,” in *CVPR*, 2013, pp. 2131–2138.

[11] Kevin McGuinness and Noel E. O’Connor, “Toward automated evaluation of interactive segmentation,” *Comput. Vis. Image Underst.*, vol. 115, no. 6, pp. 868–884, June 2011.

[12] Pushmeet Kohli, Hannes Nickisch, Carsten Rother, and Christoph Rhemann, “User-centric learning and evaluation of interactive segmentation systems,” *Int. J. Comput. Vision*, vol. 100, no. 3, pp. 261–274, 2012.

[13] Emmanouil Moschidis and Jim Graham, “A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction,” in *Proceedings of the 2010 IEEE International Conference on Biomedical imaging: from Nano to Macro*, 2010, ISBI’10, pp. 928–931.

[14] Yibiao Zhao, Xiaohan Nie, Yanbiao Duan, Yaping Huang, and Siwei Luo, “A benchmark for interactive image segmentation algorithms,” in *Person-Oriented Vision (POV), 2011 IEEE Workshop on*, 2011, pp. 33–38.

[15] Bruno Klava and Nina S. T. Hirata, “A model for simulating user interaction in hierarchical segmentation - supplementary material,” `http://www.vision.ime.usp.br/~klava/hierarchical-segmentation-model/`, 2014.

[16] Fernand Meyer, “Hierarchies of partitions and morphological segmentation,” in *Scale-Space ’01: Proceedings of the Third International Conference on Scale-Space and Morphology in Computer Vision*, 2001, pp. 161–182.

[17] Bruno Klava and Nina S. T. Hirata, “SegmentIt - Interactive image segmentation tool,” `http://segmentit.sourceforge.net/`.

[18] Marina Meilă, “Comparing clusterings – an information based distance,” *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873 – 895, 2007.

[19] J. L. Rodgers and W. A. Nicewander, “Thirteen ways to look at the correlation coefficient,” *The American Statistician*, vol. 42, pp. 59–66, 1988.