Interactive Segmentation with Recommendation of Most Informative Regions

Canxiang Yan, Dan Wang, Shiguang Shan, and Xilin Chen

Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences(CAS), Institute of Computing Technology, Beijing 100190,China
{canxiang.yan,dan.wang,shiguang.shan,xilin.chen}@vipl.ict.ac.cn

Abstract. Compared to automatic segmentation, interactive segmentation is a flexible method to separate the interesting object from background. However, satisfactory results may not be achieved even with lots of interactions since user's operation may not provide enough information to decide the labels of ambiguous regions. To deal with this problem, we present an interactive segmentation approach based on active learning scheme, which can automatically recommend the most informative regions to guide the user interactions. Our method employs a two-step strategy. Firstly, based on initial user interactions, it adopts active learning to iteratively select the most crucial regions and query the oracle for their true labels. In the second step, we minimize an energy function, which combines low-level features extracted from total interactions, to segment the object. Experimental results demonstrate our method can achieve high segmentation accuracy within desirable interactions.

Keywords: active learning, interactive, segmentation, energy minimization

1 Introduction

Interactive segmentation has gained great concerns in the field of computer vision. A lot of methods [3,5,6,10,12,13,14,15,17] have been proposed for interactive segmentation. They describe an interactive operation between an annotator and a machine to achieve segmentation results. More specifically, given an input image with single or multiple interesting objects, the oracle is queried for labeling limited pixels contained in foreground objects and backgrounds. Then this prior information is utilized to obtain the full segmentation results of desirable objects.

In this paper, we strive to address above-mentioned problems, via presenting a novel interactive segmentation algorithm based on active learning scheme [1,7,18]. The overview of the method is illustrated in Figure 1. Given an input image, a user is asked to give initial interactions by providing scribbles on objects (red lines) and backgrounds (blue lines). Then, the active learning method is employed to iteratively recommend a series of candidate regions (denoted as green regions) for guiding interactions. In each iteration, only one crucial region with greatest information is chosen for users to receive further annotation. we

define region entropy(RE), distance ratio(DR) and neighbour similarity(NS) to describe regions, then use combined information map to find most crucial region in each iteration. After T iterations, we combine feature maps into our graphical CRF model, and minimize an energy function to obtain final segmentation result.

The rest of our paper is organized as follows. Sec. 2 reviews related work. Sec. 3 presents our method. Sec. 4 is our active learning algorithm. Sec. 5 discusses experimental results. Finally, Sec. 6 concludes the paper.

2 Related Work

2.1 Interactive Segmentation

The current literatures about interactive segmentations [10,12,14,17]are roughly classified into two categories: 1) The boundary-based methods, like active contour models [10] and intelligent scissors [14], use an adaptive curve to fit the object's boundary to pop out target object. 2) Graph-based methods, such as GrabCut [17] and Lazy Snapping [12], formulate interactive segmentation into an energy minimization problem. More recently, some methods extend the previous efforts to improve the performance of interactive segmentation. Delong et al. [5] introduce a two-level MRF to model object appearance. In comparison, our target is to make user interactions more effective and reduce user effort.

2.2 Active Learning

The process of active learning is defined as querying the oracle for the true label of input data. It has been widely used in image Retrieval [8], image classification [16], object categorization [9], object segmentation [1,7,18]. A. Fathi et al. [7] propose an incremental self-training approach by iteratively labeling the least uncertain frame for video segmentation. In [18], active learning is used to address the substantial gap between segmentation accuracy of fully and weakly supervised methods. Instead, our paper proposes an active learning method to iteratively select most crucial regions to guide the user interactions. The contributions of this paper are mainly summarized as follows: 1).we use active learning in our interactive segmentation framework to recommend most crucial regions. This method can achieve high segmentation accuracy within desirable interactions. 2).we define region entropy, distance ratio and neighbor similarity, which are helpful to recommend the most crucial regions.

3 Our Method

We formulate the interactive segmentation problem as a binary labeling problem. Denote $L=\{a_p\}$ is the final result; for each component a_p , if pixel p belongs to foreground objects, $a_p = 1$; otherwise, $a_p = 0$. Given an input observation image I, we build a graph $G = (V, E)$, whose vertex corresponds to pixel and edges

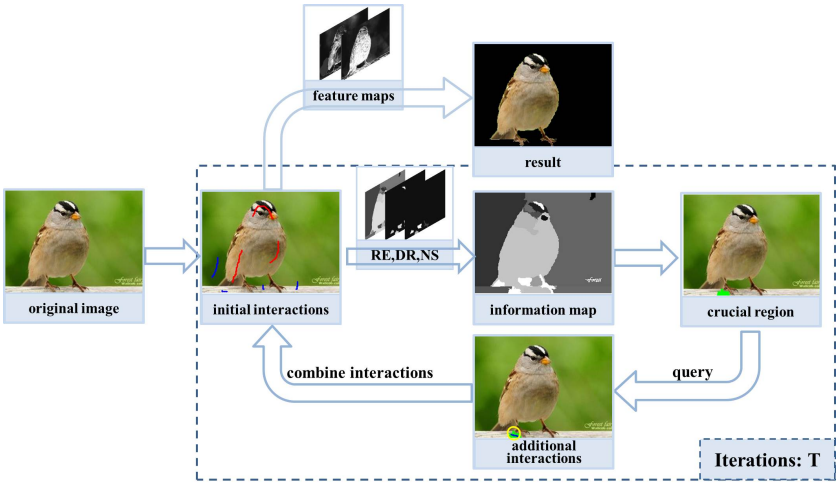

Fig. 1. an overview of our method.

are defined in an 8-connected neighborhood system. Our objective is to infer the optimal result L^* that leads to the minimization of the following Gibbs energy function:

$$E(L|I, A_0) = \sum_{p \in V} \sum_{m=1}^{M} \lambda_m F_m(a_p|I, A_0) + \sum_{(p',p) \in E} S(a_p, a_{p'}|I, A_0), \tag{1}$$

where A_0 is user interactions;F_m is the local-based energy function trained based on low-level appearance features and λ_m is the associated weights. We use a linear combination of M local-based energy function. $S(\cdot)$ denotes the pairwise function to smooth the label configuration of L. Immediately below, we elaborate these two energy functions, respectively.

Specifically, F_m is a local-based function, written as:

$$F_m(a_p) = \begin{cases} f_m(p) & \text{if } a_p = 1; \\ 1 - f_m(p) & \text{if } a_p = 0. \end{cases} \tag{2}$$

$S(a_p, a'_p)$ penalizes the disagreement between adjacent pixels. We define it as:

$$S(a_p, a_{p'}) = l(a_p \neq a_{p'}) \cdot exp(-\beta d^2_{pp'}), \tag{3}$$

3.1 Energy Minimization Based on Feature Maps

Suppose that user interactions A_0 are acquired, then we extract two low-level feature maps: 1) An object and background Gaussian Mixture Model, which are learned from object pixels and background pixels in A_0, respectively. 2) Color spatiality extracts spatiality from the color model and becomes a global feature to describe the object.

Fig. 2. information map for crucial regions. (a), (b), (c) are region entropy, distance ratio, and neighbor similarity respectively. (d) is final information map.

Minimizing Energy Function. We use Graph-Cuts algorithm [2,11] to minimize the energy function in our graphical CRF model. In our graph $G = (V, E)$ over all pixels, each node $v \in V$, is connected to the source s and the sink t and their adjacent nodes, respectively. Based on graph theory, each configuration $L = \{a_p\}$ is equal to a cut of graph G, which makes each node only connect to one of the two terminal nodes. Thus, for obtaining the optimal segmentation result L^*, our target is to find a cut which has minimum cost. So far, our energy minimization problem is equal to minimum cut of the G.

4 Active learning for most informative regions

Active learning starts when the initial user interactions are finished. Firstly, we over-segment the image I into homogeneous regions with irregular size and shape using mean shift technique [4]. After initial user interactions, some object regions and background regions are initialized respectively. The image I then can be represented as a set of all irregular regions from over-segmentation: $I = \{(A_{r_1}, r_1), (A_{r_2}, r_2), ..., (A_{r_N}, r_N)\}$, where N is the number of regions; In each pair (A_{r_i}, r_i), the label A_{r_i} takes a value from $\{A_u, A_{obj}, A_{bg}\}$ which means that the initial annotation of region r_i is unknown region, object or background. Using homogeneous regions in our method can improve the computational efficiency and allow us to compute complex statistics (e.g. texture and color).

The active learning method is conducted iteratively. In t^{th} iteration, the following three steps are performed one by one: 1) it learns region-based features: region entropy, distance ratio and neighbor similarity based on previous interactions and combines them to form an information map. 2) Then most informative regions are recommended to user; additional interactions are made on those regions, and then are combined with previous interactions as the user inputs of the $(t+1)^{th}$ iteration. After T iterations, the total interactions A_0 are obtained. The whole process is presented below.

4.1 Information map

We use three maps to describe how informative a region is: region entropy, distance ratio and neighbor similarity.

Region entropy We define initial foreground object pixels: $P_{obj} = \{p | A_{r_i} = A_{obj}, p \in r_i \text{ and } r_i \in I\}$, and use raw RGB color of P_{obj} to learn GMM_{obj}.

Similarly, we use initial background pixels to get GMM_{bg}. The components of GMM_{obj} and GMM_{bg} are 5 and 8 respectively. Then, for each pixel p_{ik} in region r_i, we normalize its object and background likelihoods from $GMMs$ to get a two-class distribution and calculate its Shannon entropy $H(p_{ik})$. The region entropy then can be written as:

$$M_{RE}(r_i) = \frac{1}{k}\sum_{k=1}^{K} H(p_{ik}), \tag{4}$$

where K is the number of pixels included in the region r_i, $M_{RE}(r_i)$ is the average entropy of region r_i. Region entropy is defined based on a simple idea that the label of a region may be ambiguous to decide if the appearance of the region is mixed by foreground object and background or its appearance has large differences to both. Shannon entropy can measure the uncertainty of a pixel label a_p and the average entropy $M_{RE}(r_i)$ of pixels in a region will be high if the total uncertainty is great. Thus, the more ambiguous a region is, the higher entropy it gets. We then get an entropy map M_{RE} by normalizing $M_{RE}(r_i)$ over all regions in image I to the range $[0, 1]$, shown in Figure 2(a). A region which has complex appearance at the bird body is bright, while another region which has consistent appearance at the bird head is darker than the former.

Distance ratio First we define the set of initial foreground regions $O = \{r_i | A_{r_i} = A_{obj}, r_i \in I\}$. For a region r' which has the initial user annotation A_u, we calculate Chi-square distance between the histogram of RGB color of r_i and r':

$$D(r_i, r') = \frac{1}{2}\sum_{k} \frac{(H^k_{r_i} - H^k_{r'})^2}{H^k_{r_i} - H^k_{r'}} \tag{5}$$

where H^k is the k^{th} bin of the color histogram; we quantify RGB colors to $5 \times 5 \times 5$, thus H has 125 bins.We then define the minimum distance from O as:

$$D^*(O, r') = min_{r_i \in O} D(r_i, r'). \tag{6}$$

Similarly, the distance $D^*(B, r')$ from r' to background regions B is defined. The distance ratio of the region r' is written as:

$$M_{DR}(r') = \frac{1}{|D^*(O, r') - D^*(B, r')| + 1}. \tag{7}$$

We then get a ratio map M_{DR} which is also normalized to $[0, 1]$. As shown in Figure 2(b), the unknown region at the tail of the bird is crucial because it is both similar to the wood with background label, which the bird stands on, and some labelled parts of the bird body.

Neighbor similarity Appearance information extracted from initial object regions O is often not effective enough to decide the whole object. There are two key reasons: 1) O only contains some parts of an object; 2) each part of the desired object may have different appearance. With this knowledge, we use neighbor similarity to find crucial regions which are adjacent to object regions O

and have different appearance to O. Denote an unknown region r'' is adjacent to *object* regions O', where O' is a subset of O and consists of labelled regions. We calculate the Chi-square distance of RGB histogram $D(O', r'')$ Equ. (5). Then neighbor similarity can be written as:

$$M_{NS}(r'') = w \cdot D(O', r''), \tag{8}$$

where w is a distance weight with the range $[0.5, 2]$:

$$w = \begin{cases} 0.5 & \text{if } R(O', r'') \in (0, 0.5]; \\ R(O', r'') & \text{if } R(O', r'') \in (0.5, 2]; \\ 2 & \text{if } R(O', r'') \in (2, +\infty]. \end{cases} \tag{9}$$

where $R(O', r'')$ is the area ratio between O' and r''. In Figure 2(c), the small regions near the bird paw are more crucial.

4.2 Interactions

Then we combine those maps using same weights to get the final information map $\gamma : \gamma(r_i) = M_{RE}(r_i) + M_{DR}(r_i) + M_{NS}(r_i)$, for $i = 1, 2, , N$. The most crucial region is the one which has the highest score in γ; the brightness of regions in Figure 2(d) is proportional to their score.

After the above operations, the method queries the oracle for the true label of the most crucial region. Then this additional interaction on crucial region is combined with initial interactions. Denote the additional interaction in the t^{th} iteration is A^t, thus the updated initial interactions in the $(t+1)^{th}$ iteration are $A_0 = A_0 \cup A^t$. After T iterations, the total interactions A_0 is the combined set of initial user interactions and all additional user interactions.

5 Experiments

In our experiments, we present results of our method and compare them with the state of the art algorithm [17]. All the experiments use unified parameters: we set the associated weights λ_1 and λ_2 to $\{0.5, 0.5\}$ in energy function Eq.(1). The parameters of mean shift over-segmentation code are set to $\{sigmaS = 9, sigmaR = 9, minRegion = 100\}$. Only one crucial region is recommended in each iteration. Then we use min-cut algorithm [2] to get an optimal solution of our energy minimum problem.

We show several experimental results in Figure 3. The desired object has great appearance variances. Given an input image, a user marks on the foreground object(red scribbles) and background(blue scribbles) respectively; the second column are the result of initial interactions. All 5 recommended crucial regions are overlapped onto the original image and highlighted with green color. The additional interactions are made in yellow circles. The last column shows the optimal results by graph-cuts algorithm. Under the parameter setting of over-segmentation algorithm, there are about 50 regions per image with 260×200

Fig. 3. Experimental results of our method. From the left to right: 1) input images; 2) images overlapped by initial interactions; 3) all recommended crucial regions with their true labels in yellow circles; 4) segmentation results. (Best viewed in color)

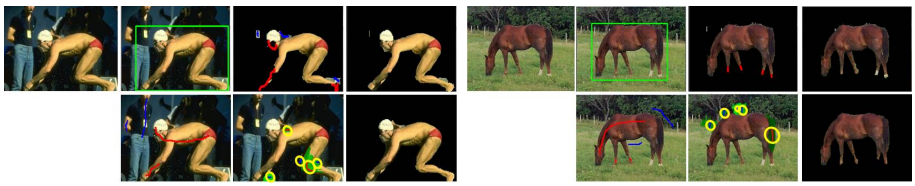

Fig. 4. Comparison of our segmentation results with GrabCut. Our method in the second row is initialized by object and background scribbles while GrabCut in the first row is initialized by a bounding box. Results are highlighted with black background.

resolution and the method only recommends about 10% regions to guide the user. The experiment on a variety of images shows that it can lead to significant improvements in segmentation accuracy within limited iterations.

Figure 4 is the comparison of our segmentation result with GrabCut. GrabCut starts with an initial bounding box. Then the iteratively scribbles are marked by the user until obtaining the desirable result. Instead, most crucial regions are recommended iteratively in our method; segmentation results of GrabCut and our method show that 1) the most crucial regions are effective to guide the user. 2) By recommending only 10% regions, we can get comparable segmentation results.

6 Conclusion

We present an active learning method for interactive segmentation. First, it iteratively selects the most crucial regions. In each iteration, region entropy, distance ratio and neighbor similarity are learned to form information map, and our method recommends the most crucial region from this map to guide user. Second, we combine color model and color spatiality feature maps acquired from

interactions into our model, and minimize associate energy function to obtain the final segmentation result. Experimental results show that our method is effective to achieve high segmentation accuracy in limited interactions. Our future work will focus on the learning of associate weights to make them object-specific.

Acknowledgments. This paper is partially supported by National Basic Research Program of China (973 Program) under contract 2009CB320902; Natural Science Foundation of China (NSFC) under contracts Nos. 60833013 and No. 60832004.

References

1. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: icoseg: Interactive co-segmentation with intelligent scribble guidance. In: CVPR. pp. 3169–3176 (2010)
2. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. TPAMI 26(9), 1124–1137 (2004)
3. Brunne, G., Chittajallu, D., Kurkure, U., Kakadiaris, I.: Patch-cuts: A graph-based image segmentation method using patch features and spatial relations. In: BMVC (2010)
4. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. TPAMI 24(5), 603–619 (2002)
5. Delong, A., Gorelick, L., Schmidt, F., Veksler, O., Boykov, Y.: Interactive segmentation with super-labels. In: EMMCVPR. pp. 147–162 (2011)
6. Ding, L., Yilmaz, A.: Enhancing interactive image segmentation with automatic label set augmentation. In: ECCV. pp. 575–588 (2010)
7. Fathi, A., Balcan, M., Ren, X., Rehg, J.: Combining self training and active learning for video segmentation. In: BMVC. pp. 78–1 (2011)
8. Hoi, S., Jin, R., Zhu, J., Lyu, M.: Semi-supervised svm batch mode active learning for image retrieval. In: CVPR. pp. 1–7 (2008)
9. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with gaussian processes for object categorization. In: ICCV. pp. 1–8 (2007)
10. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. IJCV 1(4), 321–331 (1988)
11. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? TPAMI 26(2), 147–159 (2004)
12. Li, Y., Sun, J., Tang, C., Shum, H.: Lazy snapping. TOG 23(3), 303–308 (2004)
13. Mishra, A., Aloimonos, Y., Fah, C.: Active segmentation with fixation. In: ICCV. pp. 468–475 (2009)
14. Mortensen, E., Barrett, W.: Intelligent scissors for image composition. In: CGIT. pp. 191–198 (1995)
15. Ning, J., Zhang, L., Zhang, D., Wu, C.: Interactive image segmentation by maximal similarity based region merging. Pattern Recognition 43(2), 445–456 (2010)
16. Qi, G., Hua, X., Rui, Y., Tang, J., Zhang, H.: Two-dimensional active learning for image classification. In: CVPR. pp. 1–8 (2008)
17. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. TOG 23(3), 309–314 (2004)
18. Vezhnevets, A., Buhmann, J., Ferrari, V.: Active learning for semantic segmentation with expected change. In: CVPR (2012)