LOOSECUT: INTERACTIVE IMAGE SEGMENTATION WITH LOOSELY BOUNDED BOXES

Hongkai Yu[1], *Youjie Zhou*[1], *Hui Qian*[2], *Min Xian*[3], *and Song Wang*[1]

[1]University of South Carolina, SC [2]Zhejiang University, Hangzhou [3]Utah State University, UT

ABSTRACT

One popular approach to interactively segment an object of interest from an image is to annotate a bounding box that covers the object, followed by a binary labeling. However, the existing algorithms for such interactive image segmentation prefer a bounding box that tightly encloses the object. This increases the annotation burden, and prevents these algorithms from utilizing automatically detected bounding boxes. In this paper, we develop a new LooseCut algorithm that can handle cases where the bounding box only loosely covers the object. We propose a new Markov Random Fields (MRF) model for segmentation with loosely bounded boxes, including an additional energy term to encourage consistent labeling of similar-appearance pixels and a global similarity constraint to better distinguish the foreground and background. This MRF model is then solved by an iterated max-flow algorithm. We evaluate LooseCut in three public image datasets, and show its better performance against several state-of-the-art methods when increasing the bounding-box size.

***Index Terms*—** Interactive image segmentation; Graph cut; Loosely bounded box

1. INTRODUCTION

Accurately segmenting an object of interest from an image with convenient human interactions plays a central role in image/video editing, object detection and tracking [1, 2, 3, 4]. One widely used interaction is to annotate a bounding box around the object. On one hand, this input bounding box provides the spatial location of the object. On the other hand, based on the image information within and outside this bounding box, we can have an initial estimation of the appearance models of the foreground and background, with which a binary labeling can be performed to achieve a refined segmentation of the foreground and background [5, 6, 7, 8, 9, 10, 11, 12, 13, 14].

However, due to the complexity of the object boundary and appearance, most of the existing methods of this kind prefer the input bounding box to tightly enclose the foreground object. An example is shown in Fig. 1, where the widely used GrabCut [6] algorithm fails when the bounding box does not tightly cover the foreground object. The preference of a tight bounding box increases the burden of the human interaction, and moreover it prevents these algorithms from utilizing automatically generated bounding boxes, such as boxes from object proposals [15, 16, 17], that are usually not guaranteed to tightly cover the foreground object.

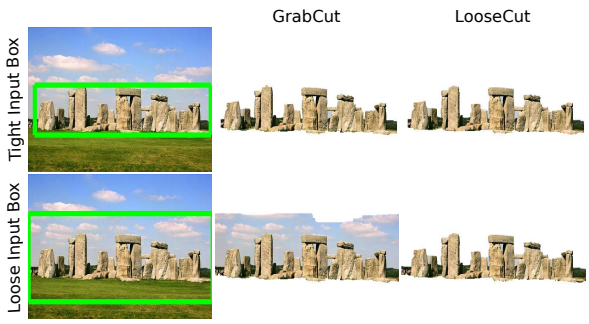

Fig. 1. Sample results from GrabCut and the proposed LooseCut with tightly and loosely bounded boxes.

The existing algorithms fail on loosely bounded boxes mainly because of the inaccurate appearance model of the foreground estimated in this case. Many existing algorithms [6] follow an iterative procedure to refine the segmentation. With a loosely bounded box, the initially estimated appearance model of the foreground is inaccurate and it is more likely to see the segmentation being trapped to a local optimum before getting to the global optimum.

Following the classic GrabCut [6], we develop a new LooseCut algorithm that can better segment the foreground object with loosely bounded boxes. Specifically, in LooseCut we propose two strategies to perform robust segmentation with loosely bounded boxes. First, we propose to add a *label consistency term* to the GrabCut energy function to encourage the consistent labeling to the similar-appearance pixels, either adjacent or non-adjacent. Second, we propose to enforce a *global similarity constraint* in the iterative segmentation procedure to explicitly emphasize the appearance difference between the foreground and background models. Considering the additional energy term and the constraint, we develop an iterated max-flow algorithm for image segmentation. In the experiments, we evaluate the proposed LooseCut algorithm using bounding boxes with varying looseness and compare its performance against several state-of-the-art interactive image segmentation algorithms.

2. PROPOSED APPROACH

In this section, we first briefly review the classic GrabCut algorithm and then introduce the proposed LooseCut algorithm.

2.1. GrabCut

GrabCut [6] performs a binary labeling to each pixel using an MRF model. Let $X = \{x_i\}_{i=1}^{n}$ be the binary labels at each pixel i, where $x_i = 1$ for foreground and $x_i = 0$ for background. Let $\theta = (M_f, M_b)$ denotes the appearance models including foreground Gaussian Mixture Model (GMM) M_f and background GMM M_b. GrabCut seeks an optimal labeling that minimizes

$$E_{GC}(X, \theta) = \sum_i D(x_i, \theta) + \sum_{i,j \in \mathcal{N}} V(x_i, x_j), \quad (1)$$

where $\mathcal{N}$ defines a pixel neighboring system, e.g., 4 or 8 neighbor connectivity. The unary term $D(x_i, \theta)$ measures the cost of labeling pixel i as foreground or background based on the appearance models θ. The pairwise term $V(x_i, x_j)$ enables the smoothness of the labels by penalizing discontinuity among the neighboring pixels with different labels. Max-flow algorithm [18] is usually used for solving this MRF optimization problem. GrabCut takes the following steps to achieve the binary image segmentation with an input bounding box: (Step 1) Estimating initial appearance models θ, using the pixels inside and outside the bounding box respectively. (Step 2) Based on the current appearance models θ, quantizing the foreground and background likelihood of each pixel and using it to define the unary term $D(x_i, \theta)$. Solving for the optimal labeling that minimizes Eq. (1). (Step 3) Based on the obtained labeling X, refining θ and going back to (Step 2). Repeating this procedure until convergence.

2.2. MRF Model for LooseCut

Following the MRF model used in GrabCut, the proposed LooseCut takes the following MRF energy function:

$$E(X, \theta) = E_{GC}(X, \theta) + \beta E_{LC}(X), \quad (2)$$

where E_{GC} is the GrabCut energy given in Eq. (1), and E_{LC} is an energy term for encouraging label consistency, weighted by $\beta > 0$. In minimizing Eq. (2), we also enforce a global similarity constraint to better estimate θ and distinguish the foreground and background. In the following, we elaborate on the label consistency term $E_{LC}(X)$ and the global similarity constraint.

2.3. Label Consistency Term E_{LC}

To encourage the label consistency of the similar-appearance pixels, either adjacent or non-adjacent, we first cluster all the image pixels using a recent superpixel algorithm [19] that preserves both feature and spatial consistency. Following a K-means style procedure, this clustering algorithm partitions the image into a set of compact superpixels and each resulting cluster is made up of one or more superpixels.

Let C_k indicates the cluster k, and pixels belonging to C_k should be encouraged to be given the same label. To accomplish this, we set a cluster label x_{C_k} (taking values 0 or 1) for each cluster C_k and define the label-consistency energy term as

$$E_{LC}(\mathbf{X}) = \sum_k \sum_{i \in C_k} \phi(x_i \neq x_{C_k}), \quad (3)$$

where $\phi(\cdot)$ is an indicator function taking 1 or 0 for true or false argument. In the proposed algorithm, we will solve for both the pixel labels and cluster labels simultaneously in the MRF optimization.

2.4. Global Similarity Constraint

In this section, we define the proposed global similarity constraint. Let M_f have K_f Gaussian components M_f^i with means μ_f^i, $i = 1, 2, \cdots, K_f$ and M_b have K_b Gaussian components M_b^j with means μ_b^j, $j = 1, 2, \cdots, K_b$. For each Gaussian component M_f^i in the foreground GMM M_f, we first find its nearest Gaussian component $M_b^{j(i)}$ in M_b as

$$j(i) = \arg \min_{j \in \{1,\ldots,K_b\}} \left| \mu_f^i - \mu_b^j \right|. \quad (4)$$

With this, we can define the similarity between the Gaussian component M_f^i and the entire background GMM M_b as

$$S(M_f^i, M_b) = \frac{1}{\left| \mu_f^i - \mu_b^{j(i)} \right|}, \quad (5)$$

which is the inverse of the mean difference between M_f^i and its nearest Gaussian component in the background GMM. Then, we define the global similarity function Sim as

$$Sim(M_f, M_b) = \sum_{i=1}^{K_f} S(M_f^i, M_b). \quad (6)$$

Similar definition for GMM distance could be found in our previous work [13]. In the MRF minimization to be discussed in the next section, we will enforce the global similarity $Sim(M_f, M_b)$ to be low (smaller than a threshold) in the step of estimating θ.

2.5. Optimization

In this section, we propose an algorithm to find the optimal binary labeling that minimizes the energy function defined in Eq. (2), subject to the global similarity constraint. Specifically, in each iteration, we first fix the labeling X and optimize over θ by enforcing the global similarity constraint on

$Sim(M_f, M_b)$. After that, we fix θ and find an optimal X that minimizes $E(X, \theta)$. These two steps of optimization is repeated alternately until convergence or a preset maximum number of iterations is reached. As an initialization, we use the input bounding box to define a binary labeling X in iteration 0 (inside box: label 1, outside of box: label 0). In the following, we elaborate on these two optimization steps.

Fixing X and Optimizing over θ: With fixed binary labeling X, we can estimate θ using a standard EM-based clustering algorithm: All the pixels with label 1 are taken for computing the foreground GMM M_f and all the pixels with label 0 are used for computing the background GMM M_b. We intentionally select K_f and K_b such that $K = K_f - K_b > 0$ since some background components are mixed to the foreground for the initial X defined by a loosely bounded box. For the obtained M_f and M_b, we examine whether the global similarity constraint is satisfied, i.e, $Sim(M_f, M_b) \leq \delta$ or not. If this constraint is satisfied, we take the resulting θ and continue to the next step of optimization. If this constraint is not satisfied, we further refine M_f using the following algorithm:

1. Calculate the similarity $S(M_f^i, M_b)$ between each Gaussian component of M_f and M_b, by following Eq. (5) and identify the K Gaussian components of M_f with the largest similarity to M_b.

2. Among these K components, if any one, say M_f^i, does not satisfy $S(M_f^i, M_b) \leq \delta$, we delete it from M_f.

3. After all the deletions, we use the remaining Gaussian components to construct an updated M_f.

This algorithm will ensure the updated M_f and M_b satisfy the global similarity constraint.

Fixing θ and Optimizing over X: Inspired by [20] and [7], we build an undirect graph with auxiliary nodes to find an optimal X that minimizes the energy $E(X, \theta)$. In this graph, each pixel is represented by a node. For each pixel cluster C_k, we construct an auxiliary node A_k to represent it. Edges are constructed to link the auxiliary node A_k and the nodes that represent the pixels in C_k, with the edge weight β as used in Eq. (2). An example of the constructed graph is shown in Fig. 2, where pink nodes v_1, v_5, and v_6 represent three pixels in a same cluster, which is represented by the auxiliary node A_1. All the nodes in blue represent another cluster. With a fixed θ, we use the max-flow algorithm [18] on this graph to seek an optimal X that minimizes the energy $E(X, \theta)$.

The graph constructed as in Fig. 2 is similar to the graph constructed in OneCut [7]. However, OneCut seeks to minimize the L_1-distance based histogram overlap between the foreground and background. This is different from the goal of the proposed algorithm: we seek better label consistency of the pixels in the same cluster by using this graph structure.

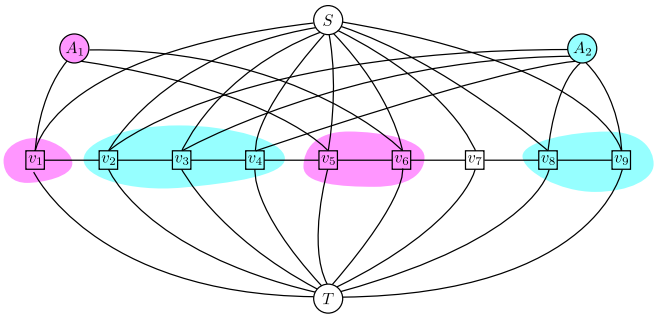

Fig. 2. Graph construction for the step of optimizing over X with a fixed θ. v_i's are the nodes for pixels and A_i's are the auxiliary nodes. S and T are the source and sink nodes. Same color nodes represent a cluster.

We will compare with OneCut in the later experiments. The full LooseCut algorithm is summarized in Algorithm 1.

Algorithm 1 LooseCut

Input: Image I, bounding box B, # of clusters N
Output: Binary labeling X to pixels in I

1: Construct N superpixel-based clusters using [19].
2: Create initial labeling X using box B.
3: **repeat**
4: Based on the current labeling X, estimate and update θ by enforcing $Sim(M_f, M_b) \leq \delta$.
5: Construct the graph using the updated θ with N auxiliary nodes as shown in Fig. 2.
6: Apply the max-flow algorithm [18] to update labeling X by minimizing $E(X, \theta)$.
7: **until** Convergence or maximum iterations reached

3. EXPERIMENTS

We conduct experiments on three widely used image datasets – the GrabCut dataset [6], the Weizmann dataset [21, 22], and the iCoseg dataset [23], and compare its performance against several state-of-the-art interactive image segmentation methods, including GrabCut [6], OneCut [7], MILCut [9], and pPBC [8]. As in [9, 7, 10], we use *Error Rate* to evaluate interactive image segmentation by counting the percentage of misclassified pixels inside the bounding box. We also take the pixel-wise *F-measure* as an evaluation metric, combining precision and recall in terms of ground-truth segmentation. For the number of Gaussian components in GMMs, we set $K_b = 5$ and $K_f = 6$. As discussed in Section 2.5, $K = K_f - K_b = 1$. To enforce the global similarity constraint, we delete $K = 1$ component in M_f. The number of clusters (auxiliary nodes in graph) is set to $N = 16$. For the LooseCut energy defined in Eq. (2), we consistently set $\beta = 0.01$. The unary term and binary term in Eq. (2) are the same as in [6] and RGB color features are used to construct

Methods	$L = 0\%$		$L = 120\%$		$L = 240\%$		$L = 600\%$	
	F-measure	Error Rate(%)	F-measure	Error Rate(%)	F-measure	Error Rate(%)	F-measure	Error Rate(%)
GrabCut	0.916	7.4	0.858	10.1	0.816	12.6	0.788	13.7
OneCut	**0.923**	6.6	0.853	8.7	0.785	9.9	0.706	13.7
pPBC	0.910	7.5	0.844	9.1	0.827	9.4	0.783	12.3
MILCut	-	**3.6**	-	-	-	-	-	-
LooseCut	0.882	7.9	**0.867**	**5.8**	**0.844**	**6.9**	**0.826**	**6.8**

Table 1. Segmentation performance on GrabCut dataset with bounding boxes of different looseness.

the GMMs. We set $\delta = 0.02$ in deleting the foreground GMM component to enforce the global similarity constraint. There are two stop conditions for the iteration: the maximum number of 10 iterations is reached or the change of the labeling results is small ($< 0.1\%$). For all the comparison methods, we follow their default settings in their codes.

We construct bounding boxes with different looseness and examine the resulting segmentation. We compute the fit box to the ground-truth foreground and slightly dilate it by 10 pixels along four directions (left, right, up, and down). We take it as the baseline bounding box with 0% looseness and then keep dilating this bounding box uniformly along four directions to generate a series of looser bounding boxes – a box with a looseness L (in percentage) indicates its area increases by L against the baseline bounding box. A bounding box will be cropped when any of its sides reaches the image perimeter.

GrabCut dataset [6] consists of 50 images. Nine of them contain multiple objects while the ground truth is only annotated on a single object, e.g., ground truth only labels one person but there are two people in the loosely bounded box. Such images are not applicable to test the performance with loosely bounded boxes, so we use the remaining 41 images in our experiments. From Weizmann dataset [21, 22], we pick a subset of 45 images for testing, by discarding the images where the baseline bounding box almost covers the full image (cannot have looser bounding boxes). For the same reason, from iCoseg dataset [23], we select a subset of 45 images for our experiment.

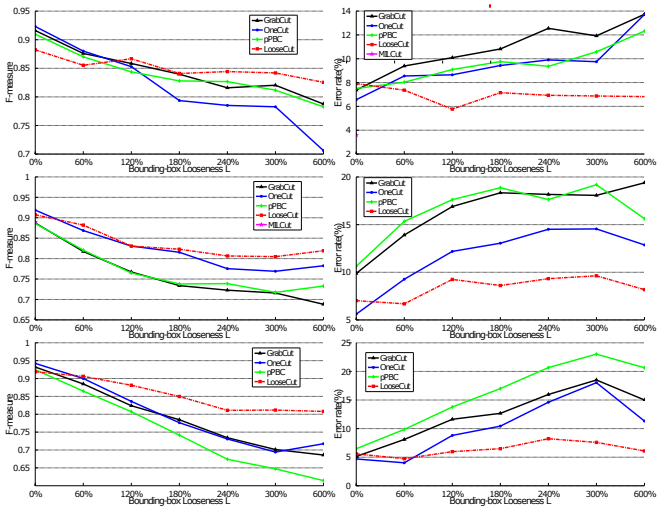

Fig. 3. Segmentation performance on datasets of GrabCut(top), Weizmann(middle) and iCoseg(bottom).

Experimental results are summarized in Fig. 3. The segmentation performance degrades when the bounding-box looseness increases for all the methods. LooseCut shows a slower performance degradation than the comparison methods. *With increasing looseness, e.g., $L \geq 120\%$, LooseCut shows higher F-measure and lower Error Rate than all the comparison methods.* Table 1 reports the values of F-measure and Error Rate of segmentation with varying-looseness bounding boxes on GrabCut dataset. Since MILCut's code is not publicly available, we only report its results with the baseline bounding boxes from its original paper. Although MILCut performs best when $L = 0\%$, MILCut will fail with a loosely bounded box because it cannot guarantee to generate desirable positive bags along the sweeping lines in a loosely bounded box for its multiple instance learning. Sample segmentation results, with the input bounding boxes of different looseness, are shown in Fig. 4.

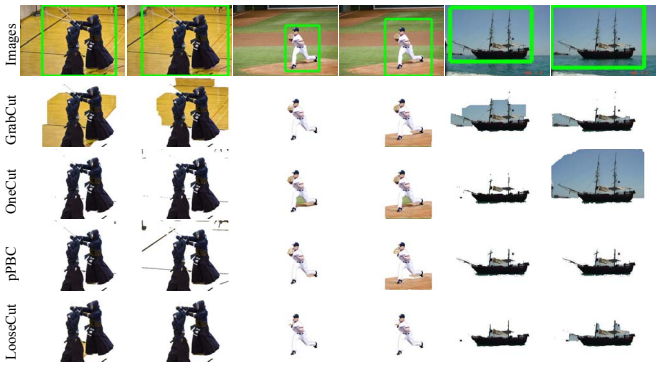

Fig. 4. Sample results of interactive image segmentation.

4. CONCLUSION

This paper proposed a new algorithm for interactive image segmentation with loosely bounded boxes. Specifically, we introduced a label consistency term and a global similarity constraint into the MRF model used in GrabCut and then developed an iterative algorithm to solve the new MRF optimization. Experiments results showed the effectiveness of LooseCut against several comparison algorithms in handling loosely bounded boxes.

Acknowledgment This work was supported in part by UES Inc./AFRL-S-901-486-002, NSF-1658987, NSFC-61672376 and NCPTT-P16AP00373. Thanks to Yuewei Lin, Dazhou Guo, Kang Zheng, Kareem Abdelfatah for insightful discussions and suggestions.

5. REFERENCES

[1] Jie Feng, Brian Price, Scott Cohen, and Shih-Fu Chang, "Interactive segmentation on rgbd images via cue selection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[2] Meng Jian and Cheolkon Jung, "Interactive image segmentation using adaptive constraint propagation," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1301–1311, 2016.

[3] D. Guo, J. Fridriksson, P. Fillmore, C. Rorden, H. Yu, K. Zheng, and S. Wang, "Automated lesion detection on mri scans using combined unsupervised and supervised methods," *BMC medical imaging*, vol. 15, no. 1, pp. 50, 2015.

[4] H. Yu, Y. Zhou, J. Simmons, C. P Przybyla, Y. Lin, X. Fan, Y. Mi, and S. Wang, "Groupwise tracking of crowded similar-appearance targets from low-continuity image sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 952–960.

[5] Leo Grady, Marie-Pierre Jolly, and Aaron Seitz, "Segmentation from a box," in *IEEE International Conference on Computer Vision*, 2011, pp. 367–374.

[6] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

[7] M. Tang, L. Gorelick, O. Veksler, and Y. Boykov, "Grabcut in one cut," in *IEEE International Conference on Computer Vision*, 2013, pp. 1769–1776.

[8] M. Tang, I.B. Ayed, and Y. Boykov, "Pseudo-bound optimization for binary energies," in *European Conference on Computer Vision*, 2014, pp. 691–707.

[9] J. Wu, Y. Zhao, J-Y Zhu, S. Luo, and Z. Tu, "Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 256–263.

[10] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *IEEE International Conference on Computer Vision*, 2009, pp. 277–284.

[11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[12] Stephen Gould, "Max-margin learning for lower linear envelope potentials in binary markov random fields," in *International Conference on Machine Learning*, 2011, pp. 193–200.

[13] H. Yu, M. Xian, and X. Qi, "Unsupervised cosegmentation based on a new global gmm constraint in mrf," in *IEEE International Conference on Image Processing*, 2014, pp. 4412–4416.

[14] H. Yu and X. Qi, "Unsupervised cosegmentation based on superpixel matching and fastgrabcut," in *IEEE International Conference on Multimedia and Expo*, 2014, pp. 1–6.

[15] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 73–80.

[16] C.L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*, pp. 391–405. 2014.

[17] Y. Zhou, H. Yu, and S. Wang, "Feature sampling strategies for action recognition," *CoRR*, vol. abs/1501.06993, 2015.

[18] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *IEEE International Conference on Computer Vision*, 2001, pp. 105–112.

[19] Y. Zhou, L. Ju, and S. Wang, "Multiscale superpixels and supervoxels based on hierarchical edge-weighted centroidal voronoi tessellation," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3834–3845, 2015.

[20] P. Kohli, L. Ladicky, and P.H.S. Torr, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[21] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[22] E. Borenstein and S. Ullman, "Class-specific, top-down segmentation," in *European Conference on Computer Vision*, pp. 109–122. 2002.

[23] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3169–3176.