Object Segmentation based on Saliency Extraction and Bounding Box

Jian Ma, Lei Huang, Bo Yin, Fengfu Yin

Jian Ma
College of Environmental Science and Engineering
Ocean University of China,
Qingdao, China, 266100
yinff3@126.com

Bo Yin
College of Information Science and Engineering,
Ocean University of China,
Qingdao, China, 266100
ybfirst@126.com

Lei Huang
College of Information Science and Engineering
Ocean University of China,
Qingdao, China, 266100
jasonhuangsc@gmail.com

Fengfu Yin
Eco-design and manufacturing Institute,
Haier R&D Department,
Qingdao, China, 266100
yinff3@126.com

***Abstract*—Object segmentation is desirable in many practical applications, e.g., object classification. However, due to various object appearances and shapes, confusing backgrounds, object segmentation in an effective way is still a challenging issue. In this paper, a novel algorithm of object segmentation based on saliency extraction and bounding boxes is proposed. The segmentation performance is significantly improved by introducing saliency extraction into the segmentation scheme. Firstly, bounding boxes are acquired by object detection algorithms, foreground and background model is constructed using bounding boxes. Then, saliency extraction procedure is introduced, and adaptive weights for each pixel are computed based on the saliency extraction. Finally, undirected graph which incorporates the adaptive weights for each pixel is constructed and graph cuts is implemented to obtain the segmentation results. Comprehensive and comparative experiments demonstrate that our proposed algorithm has achieved promising performance over a challenging public available dataset.**

Keywords - object segmentation; saliency extraction; graph cuts; bounding box; adaptive weight

I. INTRODUCTION

Object segmentation in static images is an important and challenging issue for understanding images. Many approaches have been proposed to solve this problem. Interactive methods focus on image segmentation with prior foreground and background seeds which are often labeled manually. These methods can achieve better segmentation results, but are not suitable for practical applications. To avoid the interactive operation, researchers proposed automatic segmentation approaches which use object detection techniques to get the rough object regions instead of manual labels. However, due to various object appearances and shapes, confusing background, object segmentation remains a challenging problem.

There is plenty of previous work related to object segmentation. Generally, interactive methods[1][2][3] can achieve promising results, but are not suitable for practical applications that require automatic operations. Therefore, automatic scheme for object segmentation is proposed. The automatic methods use specified foreground and background regions as input. In general, the form of specified foreground regions is bounding box[4][5]. The segmentation algorithm based on energy minimization is used to get the result with bounding boxes. To get more accurate segmentation results with the detected bounding box, Lempitsky et al.[6] adopted the graph cuts algorithm with the constraint that the desired segmentation should have parts that are sufficiently close to each of the sides of the bounding box. This is reasonable under the condition that the bounding box is accurate. Yang et al.[7] developed the adaptive edgelet features as the unary term of Conditional Random Field Model to exploit the spatial coherence of labels of neighboring pixels. These methods employ appearance models for the foreground and background which are estimated through object detection algorithms and achieve segmentation by solving energy minimization problems.

However, All the above methods which using bounding box as input, suppose that all the pixels in the bounding box can give equal contribution for the foreground model construction. This will lead false foreground prior, especially when the structure of the object is not compact. Therefore, in this paper, we propose a novel object segmentation method based on saliency extraction and bounding boxes, with the saliency of the pixels, we give each pixel a specific weights for constructing the foreground and background models.

Our contribution is that we propose a novel object segmentation method which introduce saliency extraction into the bounding box-based scheme. With the saliency extraction results, adaptive weights for each pixels is

computed for constructing the undirected graph for graph cuts method.

II. OBJECT SEGMENTATION ALGORITHM

A. System Overview

The framework of the proposed method is illustrated in Figure 1. Generally, our method consists of three stages. First, foreground and background model is constructed using bounding boxes. Second, adaptive weights for each pixel are computed based on the saliency extraction.

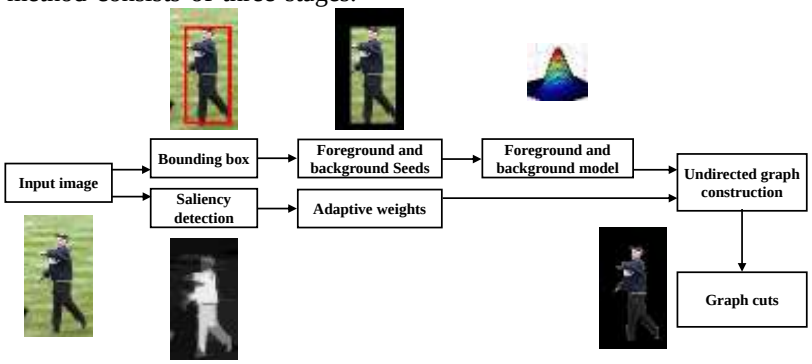

Figure1. Framework of the proposed method.

Finally, undirected graph is constructed and graph cuts is implemented to obtain the segmentation results.The details are presented below.

B. Foreground and background model construction

In order to obtain foreground and background model, object detection method is applied. In our scheme, we use the part based methods to get the bounding boxes[5][8]. Then, the foreground and background model is constructed by Gaussian mixture model (GMM). The probability density function for GMM of an observation $\mathbf{x}$ can be written as

$$p(\mathbf{x};\phi)=\sum_{i=1}^{g}\pi_i \cdot p_i(\mathbf{x};\theta)=\sum_{i=1}^{g}\pi_i \cdot \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\} \quad (1)$$

where $p_i(\mathbf{x};\theta)$ is the $p.d.f.$ corresponding to the ith Gaussian model G_i, $\mathbf{x}$ is a RGB color vector, θ consists of the elements of the mean vectors $\mathbf{u}_i$ and the covariance matrices Σ_i, d is the dimension of vector $\mathbf{x}$, and $d=3$ in our algorithm, $\phi=(\pi^T,\theta^T)^T$, π_i is the mixing parameter, $\sum_{i=1}^{g}\pi_i=1$ and $\pi_i \geq 0$, g is the number of the Gaussian models for construction of the GMM.

C. Adaptive weights determination

In order to use the prior appearance for foreground and background model construction in an effective way, we introduce the adaptive weights for each pixel. Image saliency is one of the hot research issues in image processing domain. Cheng et al.[9] proposed a fast regional contrast based saliency extraction algorithm, and had got promising results. In this paper, we introduce it into our segmentation scheme. The image saliency of regional contrast based saliency extraction can be represented as follows:

$$S(r_k)=\sum_{r_k \neq r_i}\exp(-D_s(r_k,r_i)/\sigma_s^2)\omega(r_i)D_r(r_k,r_i), \quad (2)$$

where $D_r(r_1,r_2)=\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}f(c_1,i)f(c_2,j)D(c_{1,i},c_{2,j})$, $D(c_{1,i},c_{2,j})$ is the color distance metric between pixels $c_{1,i}$ and $c_{2,j}$, $D_r(r_1,r_2)$ is the color distance between regions r_1 and r_2. $D_s(r_k,r_i)$ is the spatial distance between regions r_k and r_i, σ_s controls the strength of spatial weighting. Larger values of σ_s reduce the strength of spatial weighting. $f(c_k,i)$ is the frequency of the *i-th* color $c_{k,i}$ among all n_k colors in the k-th region r_k with $k=\{1,2\}$. The details of the saliency extraction algorithm can refer to Ref. [9]. With the saliency extraction algorithm, the saliency map is acquired. Figure 2 give some examples, the color of the pixel means the saliency, the whiter, the more salient. With the saliency results, we compute the adaptive weights for each pixel as follows:

$$\varphi_{(i,j)}=\begin{cases}1+\lceil(s_{(i,j)}-\beta)/(10*\beta)\rceil, & \text{if } s_{(i,j)}>\beta \\ 1-\lceil|(s_{(i,j)}-\beta)/(10*\beta)|\rceil, & \text{if } s_{(i,j)}<\beta\end{cases} \quad (3)$$

$$\varphi_{(i,j)}=1+\lceil(s_{(i,j)}-\beta)/(10*\beta)\rceil, \quad \text{if } s_{(i,j)}>\beta \quad (4)$$

where $\varphi_{(i,j)}$ is the adaptive weights for each pixel at location (i,j). $s_{(i,j)}$ is the saliency of pixel (i,j). β is the threshold that gives 95% recall rate for the training images, and is chosen empirically. For pixels in the bounding boxes, we use Eq.(3), and for pixels outside the bounding boxes, we use Eq.(4).

Figure 2. Samples of saliency extraction. The first line is the input images, and the second line is the corresponding saliency extraction results.

D. Undirected graph construction and graph cuts

Graph cuts are famous methods which have been successfully used for seeded image segmentation. Representing the image as an array $z=(z_1,...z_n,...z_N)'$ with z_n corresponding to the color or grey value of pixel n, the undirected graph $G=(V,E)$ is constructed with the image pixels as the nodes(V) and the neighborhood relationship between pixels(e.g. 4-neighborhood) as edge(E). There are also two specially designated terminal nodes " F "and" B "that represent "foreground" and "background" labels. Edges between pixels are called neighborhood links (n-links) and edges connecting pixels and terminal nodes are called terminal links (t-links). Then, the image segmentation corresponds to a nodes partitioning in the graph G. Defining an array of "opacity" values $\alpha=\{\alpha_1,...\alpha_n,...\alpha_N\}$ for all pixels, where $\alpha_n\in\{0,1\}$ with 0 for the background and 1 for the foreground, the image segmentation also can be expressed as a solution for inferring the unknown variables α from a given image z. Finally, the global optimal solution of α is solved by minimizing a Gibbs energy function $E(\alpha,\theta,z)$ as follows:

$$E(\alpha,\theta,z)=U(\alpha,\theta,z)+V(\alpha,z) \tag{5}$$

where

$U(\alpha,\theta,z)=\sum_n -\log(p(z_n;\alpha_n))$,

$V(\alpha,z)=\gamma\sum_{(m,n)\in C}B_{\{m,n\}}[\alpha_n\neq\alpha_m]$. $U(\alpha,\theta,z)$ is the region term, which defines the cost of t-links. $V(\alpha,z)$ is the boundary term, which defines the cost of n-links.

The traditional graph cuts method compute the cost of t-links with the same weight for all the pixels. In our scheme, we incorporate the adaptive weights for each pixel into the cost computation of t-links. Then, $E(\alpha,\theta,z)$ can be defined as:

$$E(\alpha,\theta,z)=\Phi\cdot U(\alpha,\theta,z)+V(\alpha,z) \tag{6}$$

where $\Phi=[\phi_1,\phi_2,...\phi_n...\phi_N]$.

With the adaptive weights and the foreground and background models, the undirected graph can be constructed. Then the segmentation results is obtained through the graph cuts method. In our scheme, we use the optimized version in Ref.[2].

III. EXPERIMENTAL RESULTS

In this section, we conduct comprehensive evaluations of our method. The dataset, baseline algorithms and evaluation metrics are described first.

A. Dataset, baseline algorithms, evaluation metrics

To evaluate the effectiveness of our method, we test it on a challenging public datasets: Parse dataset from[10]. The parse dataset contains 305 images of full body with a wide variety of activities ranging from standing and walking to dancing and performing exercises. The dataset include a standard train/test split. And we use the 205 testing images to evaluate our method. We compared our method with the optimized version of graph cuts method in[2](Grabcut) and saliency based segmentation method[9]. These two methods are denoted as Grabcut-Boundingbox and Grabcut-Saliency. The Grabcut-Boundingbox method using the bounding box as input, the same with ours. The Grabcut-Saliency method use the saliency extraction results as the input.

F-metric as follows is used to evaluate the performance of our method, which is similar with Ref. [11]:

$$F\text{-}metric=\frac{\sum_n(r_s(n)\cap r_g(n))}{\sum_n(r_s(n)\cup r_g(n))} \tag{7}$$

where r_s and r_g denote the segmented binary body and ground truth respectively, n is the pixel and the operators $\cap$ and $\cup$ perform pixel-wise AND and OR, respectively.

B. Comparison with other methods

As described in Section 2.3, the threshold β is obtained from the training images. A larger β means there are more pixels assigned as important ones for foreground model construction. In our experiments, $\beta=40$, which is determined by analyzing the threshold that gives 95% recall rate for the training images.

Figure 3 plots the comparison with baseline methods, where the red line represents our method, and the yellow and blue lines correspond to the Grabcut-Bounding box[2] and Grabcut-saliency[9] methods respectively. From the results we can see that our method performs best among the compared methods.

To quantitative illustrate the performance of our method, the mean and standard deviation of *F-metric* which have been used in the previous work [11] is employed. Table 1 provides the values of the each compared methods. From the table, we can see that our method outperforms the other methods by more than 5%.

Beyond these quantitative comparisons, we highlight the qualitative improvement in Figure 4. From Figure 4, we can see our method can get more accurate segmentation results than other methods.

In our experiments, we find that, the Grabcut-Boundingbox method gives the worst performance for that using rectangle region as prior foreground directly may contain background which will lead to a bad segmentation. In contrast, our method is more robust and promising since we introduced adaptive weights for each pixel for undirected graph construction, which can alleviate the influence from the background regions in the bounding boxes.

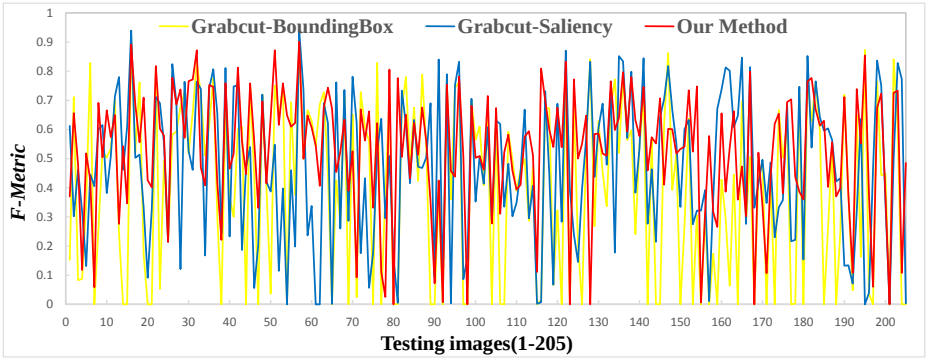

Figure 3. Comparison of our method with Grabcut-BoundingBox and Grabcut-Saliency.

TABLE 1. COMPARISON OF OUR METHOD WITH OTHER METHODS.

Method	Our Method	Grabcut-BoundingBox	Grabcut-Saliency
Mean	0.5272	0.4028	0.4695
Std. dev.	0.2099	0.2926	0.2510

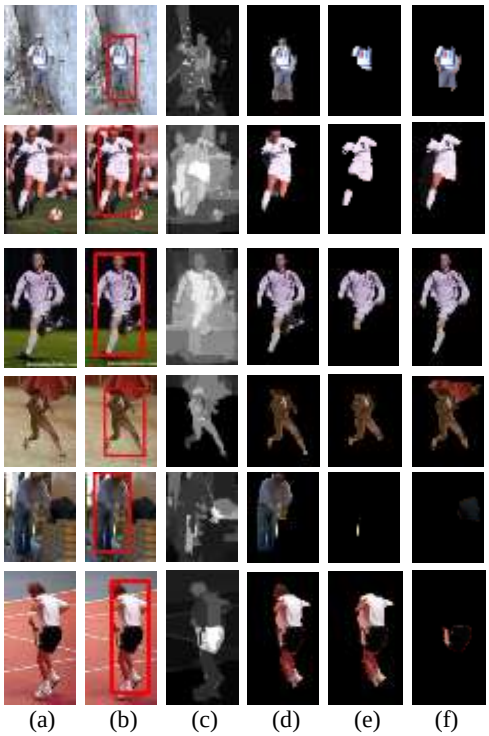

Figure 4. Performance comparisons for different methods. (a) Input images, (b) Bounding boxes, (c) Saliency extraction results, (d)-(f) are segmentation results. (d) Our method, (e) Grabcut-Boundingbox, (f) Grabcut-Saliency.

IV. CONCLUSIONS AND FUTURE WORK

We have proposed a novel method of object segmentation. Unlike the other methods, we give each pixel an adaptive weights by introducing the saliency extraction algorithm. The adaptive weights is used in the undirected graph construction. Based on the undirected graph, graph cuts is executed to get the segmentation results. Comprehensive experiments show that our method can get promising results. In the future, we plan to employ more information, e,g, shape information, for bounding box-based object segmentation.

ACKNOWLEDGEMENT

This work was supported by the Fundamental Research Funds for the Central Universities under Grant 201413021; by the Science and Technology Fund Planning Project of Qingdao under Grant 11-2-1-16-hy; by the National Nature Science Foundation of China under Grant 61202208.

REFERENCES

[1] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *Proc. of IEEE International Conference on Computer Vision*, 2001, vol. 1, pp. 105–112.

[2] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, pp. 309–314, 2004.

[3] A. Denecke, H. Wersing, J. J. Steil, and E. Körner, "Online figure–ground segmentation with adaptive metrics in generalized LVQ," *Neurocomputing*, vol. 72, no. 7, pp. 1470–1482, 2009.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893 vol. 1.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models,"

IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

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

[7] B. Yang, C. Huang, and R. Nevatia, "Segmentation of objects in a detection window by Nonparametric Inhomogeneous CRFs," *Computer Vision and Image Understanding*, vol. 115, no. 11, pp. 1473–1482, 2011.

[8] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1385–1392.

[9] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 409–416.

[10] D. Ramanan, "Learning to parse images of articulated bodies," *Advances in Neural Information Processing Systems*, vol. 19, p. 1129, 2007.

[11] S. Li, H. Lu, and L. Zhang, "Arbitrary body segmentation in static images," *Pattern Recognition*, vol. 45, no. 9, pp. 3402–3413, Sep. 2012.