Multi-instance Object Segmentation with Occlusion Handling

Yi-Ting Chen[1], Xiaokai Liu[1,2], Ming-Hsuan Yang[1]
[1]University of California at Merced. [2]Dalian University of Technology.

In this paper, we present a multi-instance object segmentation algorithm to tackle occlusions. As an object is split into two parts by an occluder, it is nearly impossible to group the two separate regions into an instance by purely bottom-up schemes. To address this problem, we propose to incorporate top-down category specific reasoning and shape prediction through exemplars into an intuitive energy minimization framework.

We start by finding the occluding regions, i.e., the overlap between two instances. For example, the overlap between the person and motorbike gives the occluding region, i.e., leg of the person, in Figure 1. To find these regions, we need to parse and categorize the two overlapping instances. Recently, Hariharan *et al.* [3] propose a simultaneous detection and segmentation (SDS) algorithm that shows a significant improvement in the segmentation classification task. This classification capability provides us a powerful top-down category specific reasoning to tackle occlusions. Then, we use categorized segmentation hypotheses obtained by SDS to infer occluding regions by checking if two of the top-scoring categorized segmentation proposals are overlapped. If they overlap, we record this occluding region into the occluding region set.

On the other hand, the classification capability are used to generate category specific likelihood maps and to find the corresponding category specific exemplar sets to better estimate the shape of objects. These category specific likelihood maps are used to indicate the location of objects in different category in a probabilistic manner. As the bottom-up segmentations tend to undershoot (e.g., missing parts of an object) and overshoot (e.g., containing background clutter), we enhance these segments through the non-parametric, data-driven shape predictor based on the chamfer matching [4]. The overview of the exemplar-based shape prediction is shown in Figure 2.

The inferred occluded regions, shape predictions and class-specific likelihood maps are formulated into an energy minimization framework to obtain the desired segmentation candidates (e.g., Figure 1(d)). Let y_p denote the label of a pixel p in an image and $\mathbf{y}$ denotes a vector of all y_p. The energy function given the foreground-specific appearance model $\mathcal{A}_i$ is defined as

$$E(\mathbf{y};\mathcal{A}_i) = \sum_{p\in\mathcal{P}} U_p(y_p;\mathcal{A}_i) + \sum_{p,q\in\mathcal{N}} V_{p,q}(y_p,y_q)\,, \tag{1}$$

where $\mathcal{P}$ denotes all pixels in an image, $\mathcal{N}$ denotes pairs of adjacent pixels, $U_p(\cdot)$ is the unary term and $V_{p,q}(\cdot)$ is the pairwise term. Our unary term $U_p(\cdot)$ is the linear combination of several terms and is written as

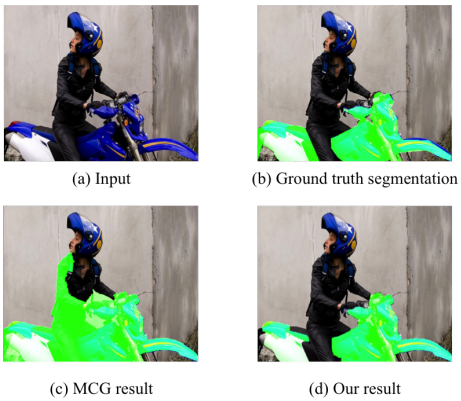
$$\begin{aligned} U_p(y_p;\mathcal{A}_i) \quad = \quad & -\alpha_{\mathcal{A}_i}\log \mathrm{p}(y_p;c_p,\mathcal{A}_i) - \alpha_{\mathcal{O}}\log \mathrm{p}(y_p;\mathcal{O}) \\ & -\alpha_{\mathcal{P}_{c_j}}\log \mathrm{p}(y_p;\mathcal{P}_{c_j})\,. \end{aligned} \tag{2}$$

For the pairwise term $V_{p,q}(y_p,y_q)$, we follow the definition as Grabcut [5].

The first potential $p(y_p;c_p,\mathcal{A}_i)$ evaluates how likely a pixel of color c_p is to take label y_p based on a foreground-specific appearance model $\mathcal{A}_i$. As in [5], an appearance model $\mathcal{A}_i$ consists of two Gaussian mixture models, foreground and background. Each foreground-specific appearance model $\mathcal{A}_i$ is initialized using a foreground mask f_i.

The second potential $p(y_p;\mathcal{O})$ accounts for the occlusion handling in the proposed energy minimization framework where $\mathcal{O}$ denotes the occluding set. Given a foreground mask f_i and its score $s_{f_i}^{c_j}$ in class c_j, we check the corresponding score of the region $f_i\backslash\mathcal{O}^*$, which is removing one of the possible occluding region $\mathcal{O}^*$ in $\mathcal{O}$ from the foreground mask f_i. When $s_{f_i\backslash\mathcal{O}^*}^{c_j} > s_{f_i}^{c_j}$, that means the pixel p in the occluding region $\mathcal{O}^*$ is discouraged to be associated with the foreground mask f_i. In this case, we penalize the energy of the occluding regions by adding the penalization when the pixel location p is foreground. When the pixel location p is background, the energy of the occluding regions is subtracted with the penalization.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.



(a) Input (b) Ground truth segmentation (c) MCG result (d) Our result

Figure 1: **Segmentation quality comparison.** Given an image (a), our method (d) can handle occlusions caused by the leg of the person while MCG [1] (c) includes the leg of the person as part of the motorbike. Moreover, the segment in (c) is classified as a *bicycle* using class-specific classifiers whereas our segment can be classified correctly as a *motorbike*.

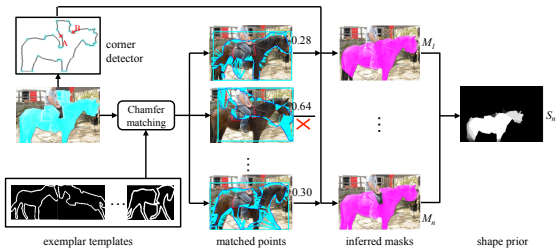


Figure 2: Overview of the exemplar-based shape predictor. This figure shows an example that the shape predictor uses the top-down class-specific shape information to remove the overshooting on the back of the horse.

The third potential $p(y_p;\mathcal{P}_{c_j})$ corresponds to one of the class-specific likelihood map $\mathcal{P}_{c_j}$. Because of the probabilistic nature of class-specific likelihood map $\mathcal{P}_{c_j}$, we set the third potential as $p(y_p;\mathcal{P}_{c_j}) = \mathcal{P}_{c_j}^{y_p}(1-\mathcal{P}_{c_j}^{1-y_p})$. Finally, we iteratively minimize the energy function (1) as in [5]. Parameters of the foreground-specific appearance model will keep updating in each iteration until the energy function converges.

The implementation and evaluation of the proposed algorithm is described in the paper in detail. We demonstrate the effectiveness of the proposed algorithm by comparing with SDS on the challenging PASCAL VOC segmentation dataset [2]. The experimental results show that the proposed algorithm achieves favorable performance. Moreover, the results suggest that high quality segmentations improve the detection accuracy significantly.

[1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge (VOC). *IJCV*, 88(2):303–338, 2010.

[3] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014.

[4] Ming-Yu Liu, Oncel Tuzel, Ashok Veeraraghavan, and Rama Chellappa. Fast directional chamfer matching. In *CVPR*, 2010.

[5] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.