J. Vis. Commun. Image R. 33 (2015) 179–192

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Cooperative object search and segmentation in Internet images ☆

Bai Yang [a], Huimin Yu [a,b,*], Yi Xie [a]

[a] Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
[b] The State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 12 March 2015
Accepted 20 September 2015
Available online 28 September 2015

Keywords:
Cosegmentation
Segmentation
Object search
Region matching
Combinatorial optimization
Structural properties
Iterative segmentation
Hierarchical search

ABSTRACT

We propose a combined approach for object search and segmentation in realistic Internet image collections. According to a query object, our goal is to locate and segment out those objects of interest. Our approach mainly includes two modules: the hierarchical discriminative region matching method and the iterative object segmentation algorithm. The hierarchical matching method is proposed to perform a hierarchical search to localize the seed-regions for segmentation. Then the iterative segmentation algorithm searches the optimal solution for the final segmentation, with the constraints from structural properties and seed-regions. These two modules work cooperatively because the seed-regions serve as constraints for segmentation and are also verified by segmentation results. Unlike existing search and segmentation approaches, our method produces accurate segmentation results and ignores noise images (images not containing the object of interest). The experimental results validate the advantages of our method on several benchmark datasets.

1. Introduction

We consider the task of searching and segmenting a specific object in general Webuser photo sets. Typically, the specific object is determined by user labeling in one photo image, and the goal is to search through the entire photo set to detect the specific object and label each pixel in the set of photo images according to whether it belongs to the specific object. This problem is a great challenge because it concerns a realistic scenario in Internet images, and solving will be useful for improving efficiency in image searching, image synthesis and object recognition.

Our task is a challenging problem that combines aspects of visual object search and segmentation. Visual object search, whose goal is to accurately locate the target object in image sets, remains a great difficulty. This is due to the fact that the target objects usually vary significantly in color, scale and viewpoint from the query object, and occupy only a small portion of an image with cluttered background. In our task, we aim to locate the objects of interest for segmentation. The dominant strategy for object localization has been multi-scale scanning. After a classifier is trained to distinguish objects from non-objects, a sliding window is swept across the image to find the best object positions [1–5]. Later, various speed-ups [4,6,7] have been presented. However, there exist several limitations about this strategy: (1) as the total number of images and windows to be evaluated in an exhaustive search is growing, the computational expense of searching all windows is tremendous; (2) it gives imprecise detection results for non-boxy objects because a bounding box may include "extra" features which may actually mislead the detector; and (3) training a classifier is impractical when applied to our task for realistic photo sets.

Moreover, we aim to segment out those objects of interest from realistic Internet image collections, according to the query object labeled by the user. For object segmentation, the traditional figure-ground segmentation is not able to address such challenging realistic images. Fortunately, the user provides the query as supervisory information for segmentation and that makes this problem somewhat relevant to the cases of supervised cosegmentation [8–17], which jointly segments multiple images containing the common objects based on some hints provided by the user. However, existing classical cosegmentation algorithms still suffer from some limitations when applied to realistic Webuser photo sets. In most existing cosegmentation methods, the energy functions proposed were based on the assumption that the object of interest occurs in every single image, without explicitly considering the cases in which the photo sets also contain many noise images, i.e., images that do not contain the object of interest at all. Additionally, most existing cosegmentation methods were shown to work well for some datasets whose images contain salient and

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author at: Institute of Information and Communication Engineering, Dept. of Information Science & Electronic Engineering, Zhejiang University, No. 38 Zheda Road, Hangzhou 310027, China.
E-mail addresses: yangbai@zju.edu.cn (B. Yang), yhm2005@zju.edu.cn (H. Yu), yixie@zju.edu.cn (Y. Xie).

http://dx.doi.org/10.1016/j.jvcir.2015.09.011
1047-3203/

similar foreground objects. Thus, it is easy to segment out those foreground objects from backgrounds based on each single image. In contrast, for Internet image collections such as photo sets, the images are more diverse and more complicated. Significant variations in posture, illumination, and view point in addition to texture, shape and scale of the object of interest pose great difficulties for existing cosegmentation techniques.

Another idea combining the object discovery and segmentation was proposed to address the noisy images for Internet images. For example, Rubinstein et al. [15] propose to use dense correspondence to discover and segment out common objects from Internet images, and Wang et al. [17] proposed a co-training framework to jointly segment and recognize categorized objects from noisy web image collection. However, these methods were presented for a different goal with ours, and when applied to our task they still suffer from some limitations. On one hand, these methods focus on objects of a certain category, while our goal is to search and segment out a specific object that is most similar to the query. Thus, we need to evaluate the "unique features" from the query object instead of class-specific features. On the other hand, general Webuser photo image may contain multiple different foreground objects, which pose great difficulties for these methods.

To address these limitations, we propose to perform a cooperative object search and segmentation. Even if the two tasks are different, they are obviously related. If a visual object search method can precisely localize the target object within an image, it becomes easy to obtain a good segmentation result. Inversely, knowing accurate segmentation mask of the target object can help to verify the search result which may be incorrect at the beginning. We use regions as the basic elements of our approach because they have some pleasant properties: (1) each region is represented by a rich set of image cues [18] (color, texture and shape), which permit the use of the more powerful yet expensive bag-of-words features [19] and (2) they can avoid including extra clutter features outside the region. Our framework mainly includes two modules: the hierarchical discriminative region matching method and iterative seed-region based object segmentation. Starting by producing a robust bag of regions for each image based on the over-segmentation method [20], the hierarchical discriminative region matching method then evaluates "the most discriminative region" and searches seed-regions through the entire image group. Lastly, we perform the iterative object segmentation based on these seed-regions. Additionally, a verification process is introduced for seed-region verification according to the segmentation results, thus making these two modules work in a mutually rewarding way.

The purpose of using the hierarchical discriminative region matching method is to localize the object of interest in a large image collection. We assume that not all regions belonging to an object are discriminative, i.e., that a sufficiently discriminative region can better represent a specific object. Thus, based on the hierarchical over-segmentation algorithm of [20], our method only needs to search the seed-regions through the entire image dataset according to the discriminative region, which not only greatly reduces the number of location to consider but also provides a novel way to detect objects of any shape.

For region-based object detection and segmentation, it remains a difficult problem to obtain the final segmentation result from numerous over-segmented regions. Several methods have been proposed. The branch-and-bound framework [21] was proposed for object localization. In [22], Russakovsky and Ng introduced a Steiner tree based approach to select object candidate regions. To detect non-boxy objects, a maximum-weight connected subgraph technique [23] was proposed. Moreover, Gunhee and Xing [24] proposed a combinatorial auction style optimization algorithm to solve a cosegmentation problem. Motivated by these works, we propose the iterative seed-region based object segmentation algorithm. The algorithm alternates between two sub-modules: the foreground modeling and the region assignment. Starting with the manual annotations of the specific object, the foreground model of the object of interest is built, and then we use this model to search for the optimal solution for the final segmentation from numerous over-segmented regions, with constraints provided from a maximum spanning tree and the seed-regions. After each round of object segmentation through the entire dataset, the images with incorrect seed-regions are removed, and the foreground model is rebuilt on the original annotations and the segmentation results of the rest of images. These steps are repeated until convergence (i.e., until the segmentation results are unchanged).

We test our method on a newly created dataset FlickrMFC [24] and the standard benchmark iCoseg dataset [14]. The experimental results in Section 6 show that our method successfully solves the object search and segmentation problem in challenging and realistic image collections. Moreover, the segmentation accuracies show that our method outperforms several previous methods in both rigid and non-rigid categories.

In summary, we make the following contributions: (1) we present a combined approach to cooperatively perform the object search and segmentation for realistic Internet images; (2) rather than the traditional exhaustive search approaches, we propose the use of the hierarchical discriminative region matching method, which is well-suited for both rigid and non-rigid objects; (3) the iterative seed-region based object segmentation algorithm is proposed to search for the optimal solution for the final segmentation result; and (4) we propose a novel foreground model encoding the color and texture information for the object of interest. The foreground model can be updated during each iteration step.

2. Related work

2.1. Object search

Sliding window detection methods [1,3,4] are classical techniques which perform exhaustive search over the image. However, the huge search space makes this method computationally expensive. Part-based object localization methods proposed by Felzenszwalb [2] are related to the sliding window techniques. These methods also perform an exhaustive search, but they search for objects and object parts, and obtain a good detection performance. Lampert et al. [21] developed a branch-and-bound technique to search for the optimal window within an image, but the computational cost is still large.

Rather than an exhaustive search, in this paper, we propose to use hierarchical discriminative region matching method to localize the seed-regions for the specific object. Combined with the iterative seed-region based object segmentation, our goal is to perform full-object segmentation in entire image group.

2.2. Object segmentation

This problem is somewhat related to the cases of supervised cosegmentation and interactive segmentation [25–30] in segmentation step. Cosegmentation was first introduced by Rother et al. [8], who proposed a novel energy function which uses histogram matching to measure dissimilarities among the two foregrounds to simultaneously segment the *common* objects in two different images. Since then, several works have been presented to improve and refine this problem, including both the unsupervised [12,15,31,32] and supervised settings [14,33,34].