Neurocomputing 135 (2014) 107–116

Contents lists available at ScienceDirect

Neurocomputing

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

Object cosegmentation by nonrigid mapping

Zhao Liu, Jianke Zhu*, Jiajun Bu, Chun Chen

College of Computer Science, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 11 June 2013
Received in revised form
8 November 2013
Accepted 12 December 2013
Communicated by Steven Hoi
Available online 30 January 2014

Keywords:
Image segmentation
Cosegmentation
Nonrigid mapping
Deformable model

ABSTRACT

Image segmentation is an important research topic in image processing and computer vision. Recently, cosegmentation has received more and more attention. Although lots of research efforts have already studied this problem in the case of single object, there still lacks the deep investigation on multiple objects cosegmentation. In this paper, we try to attack this challenge by transferring the foreground segmentations using nonrigid mapping. We present a framework, in which we first take advantage of deformable part models to detect the foreground regions across the images, and the segmentation is formulated as an energy minimization problem on pixel labeling. We have conducted a set of experiments on the FlickrMFC dataset and iCoseg dataset. The experimental results demonstrate that our proposed approach outperforms the state-of-the-art methods.

1. Introduction

Image segmentation is an important research topic in image processing and computer vision. Recently, cosegmentation has received more and more attention. Different from the traditional segmentation methods, cosegmentation mainly focuses on the problem of handling clusters of images with similar foreground, which is more efficient in practice.

Most of the current existing cosegmentation studies aim at solving the problem with single object. However, images containing multiple foreground regions are common in the real-world applications. And it is necessary to investigate the image cosegmentation with multiple objects. Different from the single object cosegmentation, multiple foreground objects cosegmentation is a more difficult problem. Related challenges mainly come from the variation of different foreground types, the occlusions between the target regions and different target poses in many images. The existing cosegmentation methods have made big progress in recent years in this field [1,2]. However, most of these methods have two common problems. First, these methods ignore the process of identifying different objects first, which always cause unclear judgements of the foreground regions. Second, the new existing methods cannot recognize the same object when it perform a nonrigid transformation between different images, and this problem is conventional in real world applications.

To tackle the above issues, in this paper, we propose an optimal framework to solve the multiple targets cosegmentation problem, we use several deformable part models to detect different target foreground regions in an image set. The reason we use deformable part model is that it is not only able to represent different kinds of foreground regions but also be able to capture the variation of these regions between different images. In addition, since the deformable model has a multiple level structure, we can trace the vote of each level of the foreground objects to improve the object retrieval process. Moreover, we use a mixed knowledge transfer learning process in our framework, besides the traditional rigid mapping process used in previous methods, we add an extra term to map the nonrigid object. By adding the nonrigid mapping, our method is able to recognize the same objects with high variations in different images.

Our contribution can be generalized to three points: (1) a mixed knowledge mapping method in estimating our foreground mask; (2) the detect stage using the deformable part models; (3) several discriminative features in representing our detected window, where our system is able to retrieve the accurate mask from the many candidate training regions.

We have conducted performance evaluation on the benchmark datasets including FlickrMFC and iCoseg. Our experimental results show that our proposed method outperforms the state-of-art-art methods around 10% on average.

2. Related work

Image segmentation is a long-studied computer vision problem, useful in extracting the crucial image regions, it is always used as a pre-process in tracking and human pose estimation. The previous image segmentation methods, such as [3,4], need users'

* Corresponding author.
E-mail addresses: liuzhao@zju.edu.cn (Z. Liu), jkzhu@zju.edu.cn (J. Zhu), bjj@zju.edu.cn (J. Bu), chenc@zju.edu.cn (C. Chen).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.050

interactions during the segmentation process. Though can get more accurate results, these methods are constrained in real applications because they are complex and inefficient. In recent years, model-based segmentation methods have been developed rapidly, two typical examples are the graphical model [5,6] and the region model [7,8]. However, these methods are restricted in the single image segmentation.

Cosegmentation is not a novel, but fast developing field especially in recent years. Early in the year of 2006, Rother [9] has introduced the definition of it and put forward the premier solution. Early cosegmentation techniques, like in [10–12], formulate the cosegmentation as an energy minimization problem, and most of them use MRF as the smoothness term in each image; these methods have a general shortage, that is, they only focus on the single target cosegmentation, and most of them are restricted in the images with background easy to distinguish.

In recent years, researchers have paid more attention to multi-view cosegmentation. Rubio et al. [13] present an unsupervised method for segmenting object from multiple images, the main idea of their work is to establish the correspondence between the similar superpixel regions in the images and to use the Gaussian Mixture Model (GMM) to predict the distribution of the pixels. Also, Kowdle et al. [14] present a multi-view object cosegmentation framework, the main point of this method is the energy model established with appearance and stereo cues. Graph is a useful data structure in image detection and recognition, it is also used by researchers in multiple foreground cosegmentation. In [1], Kim and Eric proposed an iterative cosegmentation algorithm, in which the foreground and background regions are constructed with an adjacent graph, and dynamical programming is used to infer the optimal tree from the graph. GrabCut is widely used as the energy inference in cosegmentation. As a typical example, Batra et al. [15] propose an interactive method in which they use GMM as the initialization term for predicting the foreground and background regions. Kuettel et al. [2,16] propose a supervised foreground segmentation algorithm, in this method the authors desire a transfer learning process to automatically predict the bounding windows in which the foreground may exist, and they use these windows to initialize GrabCut algorithm. Though useful in single target segmentation, this method cannot handle the multiple targets segmentation well since the detection algorithm [17] it used only works on regions in the images but not on the accurate object. The difference between our method and the previous method is that in our method we use deformable part models to replace the foreground model, thus it is able to segment the object both on the image level and on the object part level, moreover, we use a nonrigid transfer mapping in our method, which helps us to capture the highly variation of targets in images. Deformable part models, such as pictorial model [18–20] or multi-scale deformable model [21,22], are widely used in computer vision research especially in recent years. Until now, these kinds of models are mostly used in detection and pose estimation field, a well-known work is Felzenszwalb et al.'s multi-scale detection [23]. More recently, Azizpour and Laptev [24] use strongly supervised data to improve the detection performance. However, the use of this model in the segmentation field had not been explored until recent 2 years. Thandiackal [25] proposes the framework based on deformable shape model for binary segmentation, the main idea is an appearance interaction process. Also, Li et al. [26] segment the human body from the images by using a two-scale superpixel based deformable model, in this work the authors use graph cut to segment each body part. Different from the above methods which only work on typical type of object, our method can be applied to images containing multiple objects with different types.

3. Object cosegmentation by nonrigid mapping

3.1. Problem formulation

The segmentation process can be considered as a binary labeling problem over all the pixels in the image sets. For each image, define the pixels in a vector $V = \{v_1, v_2, \dots v_n\}$, and we have trained the corresponding deformable part models $T = \{t_1, t_2, \dots, t_n\}$, the potential equation has a normal form

$$E(X,T) = \sum_{v_p \in V} \sum_{t_i \in T} \phi(v_p; t_i) + \sum_{v_p, v_q \in E} \psi(v_p, v_q) \tag{1}$$

The first sum represents the probability a pixel v_p which will take a binary label l_p (1 be the foreground and 0 be the background), and the second sum is the smoothness term penalizing the neighboring pixels for taking different values. Here E represents the neighboring relationship between two pixels. We simply compute the smoothness term as follows:

$$\psi(v_p, v_q) = \gamma\, dis(v_p, v_q)^{-1} [p \neq q] \exp(-\beta |l_p - l_q|^2) \tag{2}$$

similar to most of the previous methods [2,27], the smoothness term represents the color difference between the neighboring pixels, for each pixel we take its 8 adjacent pixels into consideration. In our work, we mainly focus on expanding the unary term. The unary potential in our model is a sum of different responses which can be written as

$$\phi(x_p; T) = -\log A(l_p; v_p, T) - \log L(l_p; v_p, T) - \log M(l_p; v_p) \tag{3}$$

The potentials $\log A(l_p; v_p, T)$ and $\log L(l_p; v_p, T)$ evaluate the probability a pixel v_p to take the label l_p according to the appearance term obtained from color space and the location term obtained from segmentation mapping. We compute the appearance term as Daniel et al. did in [2]. For the location term, we first collect the votes from all the body parts, then add them as a result. Additionally, we add a nonrigid term $M(l_p; v_p)$, this term is for evaluating the probability a pixel being labeled as foreground, according to the nonrigid mapping prior described in Section 3.4. For a single image, it has the following form:

$$M(l_p; v_p) = M(v_p) v_p + (1 - M(v_p))(1 - v_p) \tag{4}$$

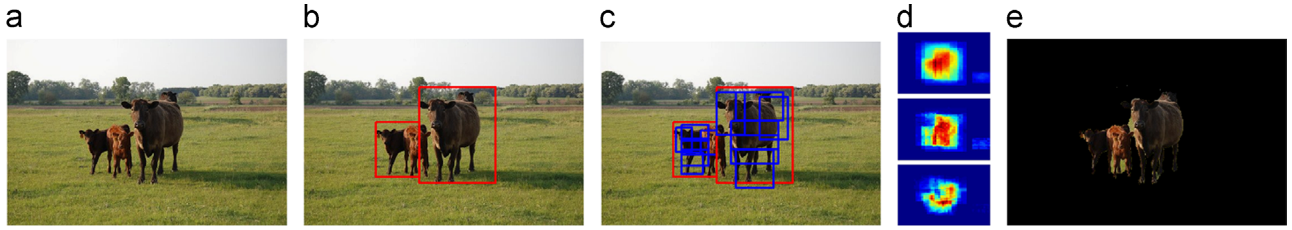

Fig. 1. Overview of our method. (a) Sample input image with multiple foreground regions to be cut. (b) Sample detection bounding windows obtained by using the method described in Section 3.2. (c) The detailed detection results by using lower level deformable part models, each blue bounding window corresponds to a body part of the foreground object. (d) The foreground mask estimated by using the method described in Section 3.4. (e) Final segmentation result estimated by using an energy minimization process. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)