Pattern Recognition 43 (2010) 676 – 684

Contents lists available at ScienceDirect

Pattern Recognition

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

POSIT: Part-based object segmentation without intensive training

Jue Wu*, Wenchao Cai, Albert C.S. Chung

Department of Computer Science and Engineering, and Bioengineering Program, The Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 6 August 2008
Received in revised form 5 May 2009
Accepted 22 July 2009

Keywords:
Object segmentation
Training
Horse and cow segmentation
Part-based model

ABSTRACT

Object segmentation is a well-known difficult problem in pattern recognition. Until now, most of the existing object segmentation methods need to go through a time-consuming training phase prior to segmentation. Both robustness and efficiency of the existing methods have room for improvement. In this work, we propose a new methodology, called POSIT, for object segmentation without intensive training process. We construct a part-based shape model to substitute the training process. In the part-based framework, we sequentially register object parts in the prior model to an image so that the searching space is largely reduced. Another advantage of the sequential matching is that, instead of predefining the weighting parameters for the terms in the matching evaluation function, we can estimate the parameters in our model on the fly. Finally, we fine-tune the previous coarse segmentation by localized graph cuts. In the experiments, POSIT has been tested on numerous natural horse and cow images and the obtained results show the accuracy, robustness and efficiency of the proposed object segmentation method.

1. Object segmentation status quo

Object segmentation is a fundamental and challenging problem in the field of pattern recognition. Its goal is to segment a whole meaningful object in a natural scene. The big challenges for object segmentation methods on real world images are the following: (1) The target object itself is complicated and may have large variance in terms of pose, intensity, color, boundary sharpness, texture, etc. See the examples shown in the 1st and 3rd columns of Fig. 1. (2) The background may be chaotic and can be confused with the foreground (object). See the examples shown in the 1st and 2nd columns of Fig. 1. (3) The images can be noise-corrupted or the object may be occluded by irrelevant objects. See an example shown in the 4th column of Fig. 1.

The conventional low-level segmentation methods usually fail to tackle these notorious obstacles. In order to meet these challenges, object segmentation methods with both top-down and bottom-up styles have received extensive interests [1–6] in the past few years. These methods consider both low-level and high-level information in images and attempt to overcome the shortcomings of the conventional approaches.

The strategy of top-down and bottom-up combo methods is to introduce a prior shape information as a high-level guidance for segmentation. The cost of introducing prior information is the extra complexity added to the methods. Until now, all the top-down and bottom-up combo methods are developed on the basis of intensive training.[1] The information about shape and appearance of the target object is needed to extract in the training phase. This phase can be time-consuming and labor-intensive because plenty of segmented images with the same object are necessary to characterize the target object. With very few exceptions, all the training data sets are segmented by manual work. This produces the most accurate segmentation ground truth but the process is quite inefficient. Among numerous training-based methods, we will review three representative top-down and bottom-up combo methods in this section.

Borenstein and his colleagues have published several papers in an effort to integrate bottom-up with top-down criteria [1,2]. This methodology, in fact, relies on low-level segments more than high-level shapes. Hence, when the low-level segments cannot separate the foreground from the background, the final segmentation will be inaccurate. Moreover, the training procedure either includes non-class training images [2] or needs to extract a large number of informative segments as templates [1]. In order to mitigate the problem of huge training burden, the authors proposed a new learning process to automatically label the unsegmented training images in

[1] Training is defined as the prior procedure of studying the examples of known input/output functionality. Intensive training indicates the involvement of a large number of training examples.

* Corresponding author.
E-mail address: woojohn911@gmail.com (J. Wu).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.07.013

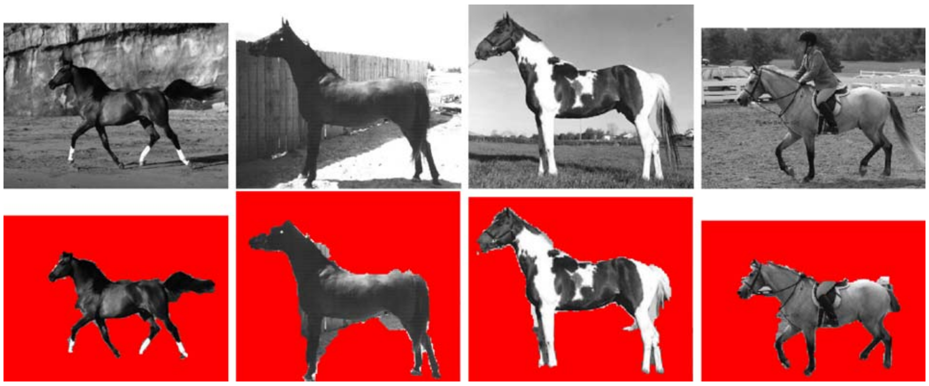

Fig. 1. Examples of horse images (1st row) and the corresponding segmentation results using the proposed method (2nd row). A brown horse may have white hoofs (1st column). The background can be easily confused with the foreground (2nd column). The skin of a horse has both slight and deep colored patches (3rd column). Part of a horse is occluded by a rider (4th column). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

another work [7]. The learning process, which combined top-down and bottom-up cues, could avoid the labor-intensive work of manual segmentation of training images. However, this method does not seem accurate enough because, when the target object has highly variable appearance, it may be in trouble due to the difficulty in matching of fragments[2] in different images.

The work of Levin and Weiss [4] formulated the bottom-up and top-down segmentation into a conditional random fields (CRF) framework. The segmentation component of the algorithm was similar to other shape-based methods but it extracted relatively fewer numbers of fragments from the training data. It was claimed in the paper that the training procedure not only considered the high-level cues but also the low-level features. This may be problematic when the target object has different low-level features (e.g. various colors or textures) in training and testing images. For example, this method may have trouble in segmenting images like the ones shown in the 1st and 3rd columns of Fig. 1. Moreover, it is unknown about how to decide the proper number of fragments. If insufficient number of fragments is chosen, the segmentation can be inaccurate.

It was shown in [3] that Obj Cut was an accurate object-category-specific segmentation method based on top-down and bottom-up cues. It combined the low-level Markov random field (MRF) model and high-level layered pictorial structure (LPS) [8] model. The accuracy of Obj Cut depends on the goodness of LPS samples because the final segmentation is the averaging result over all the samples. The requirement for the training data is demanding in Obj Cut because a number of video frames of the moving object are needed. The features of objects to be trained include both object outline and texture, which are not general enough if the object in images has various features or lacks texture patterns. Besides, although the accuracy of Obj Cut was shown to be good, the size of testing data set was not large in the work [3].

The above three works represent the state-of-the-art object segmentation methods but have a common limitation due to the training procedure. The authors of Obj Cut [3] mentioned an interesting application of nearly automatic object segmentation, namely "magic wand". For example, if the user knows that the image contains a horse, the wand can segment it without the need of manually specifying the near boundary (like intelligent scissor) or casting a set of seeds to differentiate foreground from background. However, not only is Obj Cut unable to implement that wand, but all the other state-of-the-art object segmentation methods are still far from the competent level to accomplish that goal. One of the reasons is that all the current methods rely on the intensive training procedure to acquire the prior information of the target object (shape, intensity/color, texture, etc.). This leads to a problem that, if the magic wand is required to segment multiple objects (e.g. animals, cars, human beings), the computation and storage burden of training-based methods will be tremendous.

The motivation of our work is to propose an accurate and efficient object segmentation method that does not need intensive prior training and also makes more feasible the attempt to achieve the magnificent goal of magic wand. We discard the prevailing procedure of training and pursue a new direction, which exploits part-based model for representing the basic information of the object. The purpose of training before segmentation is to learn the prior information about a specific object such that the segmentation methods can deal with the large variance of morphological and photometric features of the specific object in different images. To achieve this goal, the part-based model represents basic knowledge about the object, e.g. the number of components, the shape of each part, the relative location and orientation of each part, etc. By exploiting the part-based model, our method sequentially registers and matches[3] all parts to an image. Different from other methods that make use of intensity or texture of object, the sequential matching is mainly based on edge/gradient, which shares the same spirit of some basic psychophysical findings [10]. Finally the proposed method fine-tunes the boundary of the object according to the intensity statistics inside the region of each part, which is achieved by the localized graph cuts (LGC).

2. A part-based methodology

To circumvent the problem of intensive training, we design a new methodology for object segmentation. The framework of our method is straightforward, simple, yet effective, as will be shown experimentally. We construct a novel part-based model of the target object. From this model, we know the composition of the object and how the components (parts) are connected to each other. We then match the model to the input image in order to get a coarse segmentation. First, the salient parts are registered and matched to the image and some good candidates are kept. Then the other parts are registered and anchored with reference to the salient parts. At last, we choose the best result from these candidates according to a matching evaluation function. Based on the coarse segmentation, the boundary of each part is slightly deformed by optimizing an energy function

[2] A fragment means a rectangular patch of the image. It is a different concept from an object part.

[3] In this work, the process of aligning one part to the image is called "registration". The interaction of multiple parts is termed "matching" or "anchoring".