Computer Vision and Image Understanding 120 (2014) 81–90

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Efficient semantic image segmentation with multi-class ranking prior ☆

Deli Pei [a,b,c,d], Zhenguo Li [e], Rongrong Ji [f,*], Fuchun Sun [b,c,d,*]

[a] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[c] State Key Laboratory of Intelligent Technology and Systems, Beijing 100084, China
[d] Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China
[e] Huawei Noah's Ark Lab, Hong Kong, China
[f] Department of Cognitive Science, Xiamen University, Xiamen 361005, China

ARTICLE INFO

Article history:
Received 3 October 2012
Accepted 6 October 2013
Available online 25 October 2013

Keywords:
Computer vision
Machine learning
Semantic segmentation
Structural SVMs

ABSTRACT

Semantic image segmentation is of fundamental importance in a wide variety of computer vision tasks, such as scene understanding, robot navigation and image retrieval, which aims to simultaneously decompose an image into semantically consistent regions. Most of existing works addressed it as structured prediction problem by combining contextual information with low-level cues based on conditional random fields (CRFs), which are often learned by heuristic search based on maximum likelihood estimation. In this paper, we use maximum margin based structural support vector machine (S-SVM) model to combine multiple levels of cues to attenuate the ambiguity of appearance similarity and propose a novel multi-class ranking based global constraint to confine the object classes to be considered when labeling regions within an image. Compared with existing global cues, our method is more balanced between expressive power for heterogeneous regions and the efficiency of searching exponential space of possible label combinations. We then introduce inter-class co-occurrence statistics as pairwise constraints and combine them with the prediction from local and global cues based on S-SVMs framework. This enables the joint inference of labeling within an image for better consistency. We evaluate our algorithm on two challenging datasets which are widely used for semantic segmentation evaluation: MSRC-21 dataset and Stanford Background dataset and experimental results show that we obtain high competitive performance compared with state-of-the-art methods, despite that our model is much simpler and efficient.

1. Introduction

Semantic segmentation is a fundamental but challenging problem in computer vision, which aims to assign each pixel in an image a pre-defined semantic label. It can be seen as an extension of the traditional object detection which aims at detecting prominent objects in the foreground of an image, with closed relation to some other fundamental computer vision tasks such as image segmentation and image classification. Semantic segmentation has many applications in practice, including scene understanding, robot navigation, and image retrieval.

Semantic image segmentation algorithms in early stage typically solve this problem from a pixel-wise labeling perspective [1,2]. Although using pixels as labeling units is simple and straightforward, pixel itself contains limited and ambiguous information that cannot always be discriminative enough to determine its correct label. On the other hand, the proliferation of unsupervised image segmentation algorithms, such as mean shift [3], graph based segmentation [4,38], quick shift [5], TurboPixel [6] and SLIC [7], enables higher order features representation of regions. Therefore, more recently semantic segmentation approaches based on region-wise labeling [8–13] are also well investigated, which make use of region-level features that are not only more informative but also robust to noise, clutter, illuminate variance et al. In such a setting, an initial unsupervised segmentation is commonly adopted for pre-processing. However, image segmentation is still far away from being perfect without regard to the extensive attempts in the last several decades. From this point of view, how to make best use of these imperfect unsupervised image segmentation algorithms on the semantic segmentation problem is of fundamental importance yet is still unclear.

Although higher order features extracted from regions are more expressive and informative than those from pixels, sematic ambiguity still exists because of the appearance similarity. A general consent is that contextual information within an image is a very useful cue to attenuate this ambiguity, which can be used to

☆ This paper has been recommended for acceptance by Nicu Sebe.

* Corresponding authors. Addresses: Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (F. Sun). Department of Cognitive Science, School of Information Science and Technology, Xiamen University, Xiamen 361005, China (R. Ji).

E-mail addresses: derrypei@gmail.com (D. Pei), li.zhenguo@huawei.com (Z. Li), rrji@xmu.edu.cn (R. Ji), fcsun@mail.tsinghua.edu.cn (F. Sun).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.10.005

suppress/encourage the presence of object classes during labeling. Context refers to any information that is not extracted directly from local appearance and can be summarized into two categories: pairwise constraints and global cues. Pairwise constraints, such as smoothness based on contrast [14,9], relative location [10,11] and co-occurrence [8,11,15] are used to model the pairwise relationship between regions within an image. Global constraints are usually used to enforce higher level consistency of region sets or image level. Some approaches are proposed to model these cues, such as using image classification results [13], Potts potential [12], p^N Potts potential [16] and its improved versions robust P^N potential [14], p^N-based hierarchical CRFs [17], and Harmony potential [9]. These models will be further discussed in Section 2.

In terms of the methodology, most of the existing methods [16,12,14,10,11,15,9] use conditional random fields (CRFs) to combine these constraints from different levels and make joint inference of labeling within an image, which is also known as structured prediction. In contrast to many sophisticated algorithms for inference, these models [10,11,9,14,15] are usually learned by gradient descent or heuristic search on validation set based on maximum likelihood estimation. On the other hand, Zhu et al. [18] showed that the max-margin based learning algorithm is more robust for structured prediction compared with the maximum likelihood estimation based learning algorithm in many machine learning applications.

In this paper, we use maximum margin based structural support vector machine (S-SVMs) model to combine multiple levels of cues to attenuate the ambiguity of appearance similarity and we propose multi-class ranking based global constraints to confine the object classes to be considered when labeling regions within an image.

For global cues, we first rank all the object classes for an image (class with higher probability present in the image gets larger score) using multi-class ranking algorithm [20] and transform the ranking scores into image-level soft constraint to confine the possible classes present in the image. The advantages of this global cues can be seen from two aspects: on the one hand, compared with robust p^N potential [14] which limits their parent node to take only one single label, our method ranks all the classes for an image and thus is more representative to heterogeneous regions. On the other hand, since we compute the ranking scores for all the classes and transform them to soft constraint, we do not need to make hard decision for every class and thus avoid searching exponential space of possible label combination as harmony potential [9]. The global cues are integrated with the prediction obtained from region feature and logistic regression to encouraging more likely classes while suppressing the others.

We then introduce inter-class co-occurrence statistics as pairwise constraints and combine them with the prediction from previous stage under S-SVMs framework. This enables the joint inference of labeling within an image for better consistency. Moreover, our model can be can be efficiently learned with cutting plane algorithm [19] instead of using heuristic search approach as in CRFs learning. Experimental results show that we obtain high competitive performance with state-of-the-art methods with a much simpler and efficient model on two challenging datasets: MSRS-21 and Stanford Background Dataset.

Probably the most related work is [21], which discussed the application of structural SVM in image semantic segmentation and compared with alternative maximum likelihood method. However, our model is different from their model in designing pairwise and global constraints as well as loss function in parameter learning. The standard contrast-dependent Potts model was used as pairwise constraint in contrast to our co-occurrence property. With regard to global constraints, they used very simple and straightforward K image-level classification results and the advantage of multiple classes ranking over 1-VS-All classifiers is discussed in [20].

The remainder of the paper is organized as follows: In the next section we review the related work. Our model is presented in Section 3, including the problem formulation and model details. Sections 4 and 5 describe the inference and learning methods. Implementation details and performance evaluation are shown in Section 6 while conclusions are drawn in Section 7.

2. Related work

Despite of the success in inferring pixel labels [1,2], more recent methods tend to infer labels over regions or superpixels for the sake of lower computational complexity and incorporating higher level semantic cues. For these approaches, traditional image segmentation algorithms such as Normalized Cut [8], meanshift [14,17,13], graph-based image segmentation [10], quick shift [9,22] are adopted to get initial segments. More recently, several over-segmentation algorithms [6,7] are developed to bypass the problem of tradition segmentation algorithms, such as the semantic ambiguity (regions span multiple object classes) and the difficulty to determine the optimal number of segment regions. These algorithms try to seek the trade-off between reducing image complexity through pixel-grouping and avoiding under-segmentation [6]. Images are decomposed into much smaller regions than object size, e.g. 100–300 regions. Many traditional segmentation algorithms can also be adopted to generate superpixels by setting a finer level region segments. Qualitative results of different segmentation algorithms are given in Fig. 1, where each image is decomposed into approximate 150 superpixels. It can be seen that over-segmentation algorithms tent to segment an image into regions with regions with approximate size while the region size of traditional segmentation may vary a lot with the complexity of the content.

Although various powerful features have been proposed recently (e.g. color histogram, texture and SIFT, these feature are still not informative enough to achieve high classification performance because of the appearance similarity. To attenuate this ambiguity of feature representation, some pairwise constraints, such as smoothness [14,9,23], relative location [10,11] and co-occurrence [8,11,15], are further introduced to attenuate the ambiguity of feature representation: (i) The assumption for pairwise smoothing term is that adjacent regions tend to have same label, and subsequently spatially adjacent regions with different labels will be punished. To keep the boundary, appearance contrast is considered in smoothing term, by which regions with larger appearance contrast will be punished less for their inconsistent labels. However, the dilemma of this smoothing term is that regions with similar appearance will naturally tend to have same label. This is contradicted with the objective of smoothing term that expecting spatially adjacent regions with variant appearance to have same label. (ii) The co-occurrence statistics exploit the property that some classes (e.g. boat, water) are more likely to present within an image than others (e.g. car, water). Thus the existence of one class can be used as the evidence of expecting the presence of some highly related classes and suppress the presence of other unlikely classes. For instance, Rabinovich et al. [8,11] construct context matrices by counting the co-occurrence frequency among object labels in the training set to incorporate semantic contextual information. Ladicky et al. [15] claimed that the co-occurrence cost should depend only on the labels present in an image, it should be invariant to the number and location of pixels that object occupies. (iii) Gould et al. [10] encoded the inter-class spatial relationship as a local feature in a two-stage classification process. However, because of the 2D projection, relative location in images is usually uninformative and hence degenerates to co-occurrence constraint.