Expert Systems With Applications 63 (2016) 231–240

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Modeling spatial layout for scene image understanding via a novel multiscale sum-product network

Zehuan Yuan[a], Hao Wang[a], Limin Wang[b], Tong Lu[a,*], Shivakumara Palaiahnakote[c], Chew Lim Tan[d]

[a] National Key Lab for Novel Software Technology, Nanjing University, China
[b] Computer Vision Laboratory, ETH Zurich, Switzerland
[c] Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
[d] School of Computing, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 19 January 2016
Revised 8 July 2016
Accepted 9 July 2016
Available online 12 July 2016

Keywords:
Scene image understanding
Spatial layout
Multiscale unary potentials
Multiscale sum-product network
MSPN

ABSTRACT

Semantic image segmentation is challenging due to the large intra-class variations and the complex spatial layouts inside natural scenes. This paper investigates this problem by designing a new deep architecture, called *multiscale sum-product network* (MSPN), which utilizes *multiscale unary potentials* as the inputs and models the *spatial layouts* of image content in a hierarchical manner. That is, the proposed MSPN models the joint distribution of multiscale unary potentials and object classes instead of single unary potentials in popular settings. Besides, MSPN characterizes scene spatial layouts in a fine-to-coarse manner to enforce the consistency in labeling. Multiscale unary potentials at different scales can thus help overcome semantic ambiguities caused by only evaluating single local regions, while long-range spatial correlations can further refine image labeling. In addition, higher orders are able to pose the constraints among labels. By this way, multi-scale unary potentials, long-range spatial correlations, higher-order priors are well modeled under the uniform framework in MSPN. We conduct experiments on two challenging benchmarks consisting of the MSRC-21 dataset and the SIFT FLOW dataset. The results demonstrate the superior performance of our method comparing with the previous graphical models for understanding scene images.

1. Introduction

Natural scene image understanding, which aims at labeling each pixel of a scene image to a predefined object class and simultaneously performing segmentation or recognition of multiple objects that occur in the scene, has been extensively studied in the past years (Farabet, Couprie, Najman, & LeCun, 2013; Gritti, Damkat, & Monaci, 2013; Liu, Xu, & Feng, 2011; Rincón, Bachiller, & Mira, 2005; Shotton, Winn, Rother, & Criminisi, 2009; Tighe & Lazebnik, 2013; Tu, Chen, Yuille, & Zhu, 2005; Yin, Jiao, Chai, & Fang, 2015). However, since even the objects of the same class tend to exhibit large intra-class variations in natural scenes, automatically providing satisfying high-level semantics from complex images is still a very challenging task. With the recent development of vision-based hardware and network techniques, it becomes an active research topic and has attracted more and more researchers in computer vision and pattern recognition community.

Fortunately, in addition to low-level cues, most natural scene images contain intrinsic spatial structures, namely, scene contextual information. Thus the state-of-the-art approaches consider the spatial layout of each scene image as a kind of prior information to improve image understanding results. These methods jointly model the low-level appearances of every single patch and the spatial structures between adjacent patch pairs through a unified framework. A common choice for spatial layout modeling is to resort to graphical models typically like Conditional Random Field (CRF) (Krähenbühl & Koltun, 2011; Ladicky, Russell, Kohli, & Torr, 2009), where nodes are built on image pixels or superpixels, while edges incorporate the second order priors like the smoothness between adjacent nodes. Then the problem of image understanding is treated as Maximum-a-Posterior (MAP) inference in the graphical model. These methods are particularly effective for modeling the relationship between adjacent objects. However, they may not perform well for complex scene images due to

[*] Corresponding author.
E-mail addresses: zhyuan001@gmail.com (Z. Yuan), wanghao0711@gmail.com (H. Wang), lmwang.nju@gmail.com (L. Wang), lutong@nju.edu.cn (T. Lu), hudempsk@yahoo.com (S. Palaiahnakote), tancl@comp.nus.edu.sg (C. Lim Tan).
URL: http://cs.nju.edu.cn/lutong/ (T. Lu)

http://dx.doi.org/10.1016/j.eswa.2016.07.015
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

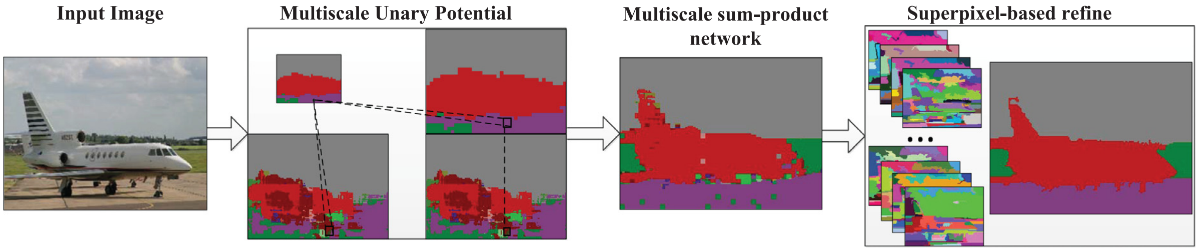

Fig. 1. The pipeline of scene image understanding using the learned MSPN. Left: the original scene image. Middle-left: computing the multiscale unary potentials from the input image. Middle-right: inferring the label for each pixel inside the original scene image by maximizing the posterior with the learned MSPN, which can be conducted efficiently by a two-pass algorithm. Right: refining the results using over-segmented regions.

the following limitations. First, CRF with a grid structure has the ability to utilize adjacent spatial relations, but cannot characterize a wider range of spatial context constraints among non-adjacent scene objects, which sometimes play a more important role than adjacent relations in automatically understanding scene image contents. For example, it is always hard to decide scene content directly from local visual features due to their instabilities brought by intensity, color, texture, illumination, occlusion and viewpoint variations. Instead, by combining a wider range of spatial relations on different scales, scene content understanding results can be greatly improved. Unfortunately, how to combine both the adjacent (short-range) and the nonadjacent (long-range) spatial layouts of image content to enforce the consistency of scene image parsing with such graphical models is still admittedly a hard problem. Second, the inference and training of some complex graphical models may be inefficient, which make the use of graphical models in real-life computer vision related applications inflexible and difficult. For example, CRF with a full connected structure is often hard to infer and train (Farabet et al., 2013). Finally, the state-of-the-art graphical models often face the difficulty on how to integrate high-order object shape priors, which are essential in parsing or understanding image semantics accurately. However, object shape priors sometimes are not well integrated into such models.

In our previous research, we explored scene image co-segmentation by a topic-level random walk framework (Yuan, Lu, & Shivakumara, 2014) and object category discovery in natural scenes using a context-aware graphical model (Yuan & Lu, 2014). To further address the discussed issues in visual scene image understanding, this paper proposes a novel deep architecture named *Multiscale Sum-Product Network* (MSPN), which can be viewed as a stacked sum-product network (SPN) (Poon & Domingos, 2011) to jointly model the distribution over image-level labels and unary potentials from different scene scales. Due to the deep structure of MSPN, the proposed model is able to characterize both the local (short-range) and the global (long-range) spatial relations on different scales from pixel-level, patch-level to image level through a hierarchical manner for better parsing the semantics from complex scene images. Ideally, the combination of both the two types of spatial relations can help understand scene images more accurately since long-range interactions among image patches can be well characterized by MSPN. Additionally, by stacked MSPN, we have the ability to characterize high-order relations among pixels in a flexible and implicit way. To the best of our knowledge, this is the first work by introducing the conceptual deep sum-product network into scene image understanding or content parsing research to reduce the instabilities brought by the unpredictable variations of low-level local visual appearances.

In our architecture, the product operation models various correlations between every two adjacent patches, on which the sum operation further integrates these correlations into the “feature” of a larger patch. On the bottom layer of MSPN, an SPN is designed for each patch to model the joint distribution over the unary potentials and image labels from the previous scale, aiming at modeling the local context information within the image patch. On the up layer of MSPN, a global SPN is proposed for the whole image, aggregating the information from all the SPNs of the patches on the bottom layer, the unary potentials and image labels. Thus, the up layer of MSPN is able to capture long-range interactions among image patches and thereby successfully models the global context information of image content for parsing complex scene semantics more accurately.

In addition to the deep modeling of spatial layouts in every scene, MSPN also allows for efficient inference during the testing phase, which benefits from the fact that the proposed MSPN is a deep tractable model and only contains relatively simple *product* and *max* operations. This is particularly useful for designing real-life systems with a much lower computational load. We show the overall pipeline for understanding the semantics of an unknown scene image in Fig. 1, where we first compute the multiscale unary potentials (middle-left in Fig. 1), and then infer the label for each pixel inside it by maximizing the posterior with the learned MSPN (middle-right), which can be conducted efficiently by a two-pass algorithm. Finally, scene image understanding results will be further refined by using the over-segmented region information from the original scene image (right).

Our main contributions are two-folds: (1) a novel deep network framework named MSPN is proposed to perform semantic image segmentation by a more effective and efficient way comparing with the popular graphic models; and (2) the architecture of MSPN is elaborately designed to model multi-scale features, either local or global spatial layout of any scene image, and higher-order priors under an unified framework. The results on two popular benchmarks show that the proposed MSPN addresses semantic image segmentation effectively.

The rest of the paper is organized as follows. Section 2 discusses the related work. In Section 3, we introduce the structure of MSPN. Section 4 gives the training and inference methods for scene image understanding. Experimental results and discussions are given in Section 5, and finally Section 6 concludes the proposed model.

2. Related work

The problem of image understanding or parsing has been extensively studied in the previous research, and the existing approaches can be roughly classified into three categories, namely, bottom-up scoring, top-down refinement, and region label reasoning.

In *bottom-up scoring methods*, a fairly large number of object hypotheses is first generated and then low-level color, texture and shape features are used on these segments for classifying object regions. For example, Gu, Lim, Arbelaez, and Malik (2009) present a unified max-margin framework for object detection, segmentation, and classification using region-based features, from which the