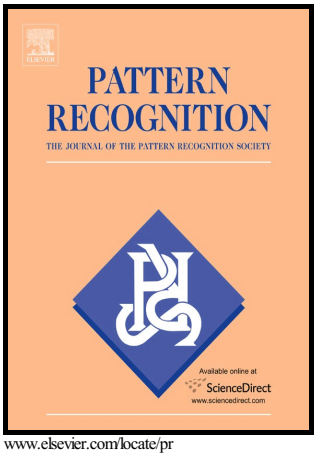
Scene Parsing using Inference Embedded Deep Networks

Shuhui Bu, Pengcheng Han, Zhenbao Liu, Junwei Han



www.elsevier.com/locate/pr

PII: S0031-3203(16)00048-0
DOI: http://dx.doi.org/10.1016/j.patcog.2016.01.027
Reference: PR5629

To appear in: *Pattern Recognition*

Received date: 5 August 2015
Revised date: 10 December 2015
Accepted date: 22 January 2016

Cite this article as: Shuhui Bu, Pengcheng Han, Zhenbao Liu and Junwei Han, Scene Parsing using Inference Embedded Deep Networks, *Pattern Recognition*, http://dx.doi.org/10.1016/j.patcog.2016.01.027

Scene Parsing using Inference Embedded Deep Networks

Shuhui Bu[a], Pengcheng Han[a], Zhenbao Liu[a,*], Junwei Han[a,**]

[a]*Northwestern Polytechnical University, China*

Abstract

Effective features and graphical model are two key points for realizing high performance scene parsing. Recently, Convolutional Neural Networks (CNNs) have shown great ability of learning features and attained remarkable performance. However, most researches use CNNs and graphical model separately, and do not exploit full advantages of both methods. In order to achieve better performance, this work aims to design a novel neural network architecture called Inference Embedded Deep Networks (IEDNs), which incorporates a novel designed inference layer based on graphical model. Through the IEDNs, the network can learn hybrid features, the advantages of which are that they not only provide a powerful representation capturing hierarchical information, but also encapsulate spatial relationship information among adjacent objects. We apply the proposed networks to scene labeling, and several experiments are conducted on SIFT Flow and PASCAL VOC Dataset. The results demonstrate that the proposed IEDNs can achieve better performance.

Keywords: Convolutional Neural Networks (CNNs), Conditional Random Fields (CRFs), Inference Embedded Deep Networks (IEDNs), Hybrid Features

1. Introduction

As the intelligent time is coming, computer vision as an important technical field of artificial intelligence, has achieved rapid development in recent years. Scene parsing, a complex high level vision task not only detecting and segmenting the different objects but also recognizing what classes the objects belong to, is primarily important for a wide scope of applications. The core technique for realizing the scene parsing is to label every pixel in images with accuracy as high as possible [1, 2].

Compared to scene parsing, image classification task [3] recently has made a significant breakthrough. The image classification usually assumes that the object of interest is centered and at a fixed scale, as a consequence the object localization problem is not vital. However, real scenes are complex and volatile, rarely just containing a single object or object class, hence scene parsing belongs to a multi-label classification task caring the object location and recognizing every isolated object in a scene, which leads to some challenging problems. The latest researches show that there are two keys affecting the performance of scene labeling: one is how to extract good representations of images [4, 5, 6], and the other is how to infer and improve the object class based on their spatial relationship between different objects in images [7, 8].

In the last decade, lots of researches have been conducted to find good features that represent the intrinsic properties of objects and many effective features are proposed, such as Gist [9], HoG [10], SIFT [11], SURF [12], and so on. Although these features have achieved great performance in some vision applications, they just depict one part information of object in images, which causes inconsistent performance in different tasks, thereby they have limited performance and application range. Moreover, these features are extracted with systems relying on carefully engineered design, which increases the difficulty for further improvement. Some researches [13, 14] are explored through simulating human vision mechanism, and some achievements have been made. In order to overcome the shortcomings of these features, deep learning [15, 16, 17, 18, 19, 20, 21, 22] based feature learning methods have been investigated and become the mainstream of vision researches. Over the several past years, Convolutional Neural Networks (CNNs) [23, 24], one type of deep leaning methods, have pushed the performance of computer vision systems to soar heights on a broad array of high-level techniques, including image classification [23, 24], object detection [24, 25], fine-grained categorization [26], and so on. The success is partially attributed to three aspects: First, CNNs, a powerful machine learning model, automatically learn feature from image databases and are independent of the target dataset, which removes the demand to compute over lots of hand-crafted features and the need for feature selection. Second, they are deep architectures having the capacity to learn more complex non-linear model than the traditionally shallow ones [4], such as Support Vector Machine (SVM) [27] or Neural

*Corresponding author
**Corresponding author
Email addresses: bushuhui@nwpu.edu.cn (Shuhui Bu), hanpc@mail.nwpu.edu.cn (Pengcheng Han), liuzhenbao@nwpu.edu.cn (Zhenbao Liu), jhan@nwpu.edu.cn (Junwei Han)