Neurocomputing 304 (2018) 82–103

Contents lists available at ScienceDirect

Neurocomputing

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

Methods and datasets on semantic segmentation: A review

Hongshan Yu[a,b,*], Zhengeng Yang[a], Lei Tan[a,c], Yaonan Wang[a], Wei Sun[a], Mingui Sun[d], Yandong Tang[e]

[a] National Engineering Laboratory for Robot Visual Perception and Control Technology, College of Electrical and Information Engineering, Hunan University, Changsha, China
[b] Shenzhen Research Institute of Hunan University, Shenzhen, Guangdong 518057, China
[c] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[d] Laboratory for Computational Neuroscience, University of Pittsburgh, Pittsburgh, USA
[e] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China

ARTICLE INFO

Article history:
Received 22 June 2017
Revised 31 January 2018
Accepted 19 March 2018
Available online 1 May 2018

Communicated by XIANG Xiang Bai

Keywords:
Semantic segmentation
Convolutional neural network
Markov random fields
Weakly supervised method
3D point clouds labeling

ABSTRACT

Semantic segmentation, also called scene labeling, refers to the process of assigning a semantic label (e.g. car, people, and road) to each pixel of an image. It is an essential data processing step for robots and other unmanned systems to understand the surrounding scene. Despite decades of efforts, semantic segmentation is still a very challenging task due to large variations in natural scenes. In this paper, we provide a systematic review of recent advances in this field. In particular, three categories of methods are reviewed and compared, including those based on hand-engineered features, learned features and weakly supervised learning. In addition, we describe a number of popular datasets aiming for facilitating the development of new segmentation algorithms. In order to demonstrate the advantages and disadvantages of different semantic segmentation models, we conduct a series of comparisons between them. Deep discussions about the comparisons are also provided. Finally, this review is concluded by discussing future directions and challenges in this important field of research.

1. Introduction

With the ever-increasing range of intelligent applications (e.g. mobile robots), there is an urgent need for accurate scene understanding. As an essential step towards this goal, semantic segmentation thus has received significant attention in recent years. It refers to a process of assigning a semantic label (e.g. car, people) to each pixel of an image. One main challenge of this task is that there are a large amount of classes in natural scenes and some of them show high degree of similarity in visual appearance.

The emergence of terminology “semantic segmentation” can be dated back to 1970s [1]. At that time, this terminology was equivalent to image segmentation but emphasized that the segmented regions must be “semantically meaningful”. In 1990s, “object segmentation and recognition” [2] further distinguished semantic objects of all classes from background and can be viewed as a two-class image segmentation problem. As the complete partition of foreground objects from the background is very challenging, a relaxed two-class image segmentation problem: the sliding window object detection [3], was proposed to partition objects with bounding boxes. It is useful to find where the objects in the scenes with excellent two-class image segmentation algorithms such as constrained parametric min-cuts(CPMC) [4]. However, two-class image segmentation cannot tell what these objects segmented are. As a result, the generic sense of object recognition(or detection) was gradually extended to multi-class image labeling [5], i.e., semantic segmentation in present sense, to tell both where and what the objects in the scene.

In order to achieve high-quality semantic segmentation, there are two commonly concerned questions: how to design efficient feature representations to differentiate objects of various classes, and how to exploit contextual information to ensure the consistency between the labels of pixels. For the first question, most early methods [6–8] benefit from using the hand-engineered features, such as Scale Invariant Feature Transform (SIFT) [9] and Histograms of Oriented Gradient(HOG) [10]. With the development of deep learning [11,12], the using of learned features in computer vision tasks, such as image classification [13,14], has achieved great success in past few years. As a result, the semantic segmentation community recently paid lots of attention to the learned features [15–26], which are usually refer to Convolutional Neural

* Corresponding author at: National Engineering Laboratory for Robot Visual Perception and Control Technology, College of Electrical and Information Engineering, Hunan University, Changsha, China.
E-mail address: yuhongshancn@hotmail.com (H. Yu).

https://doi.org/10.1016/j.neucom.2018.03.037
0925-2312/

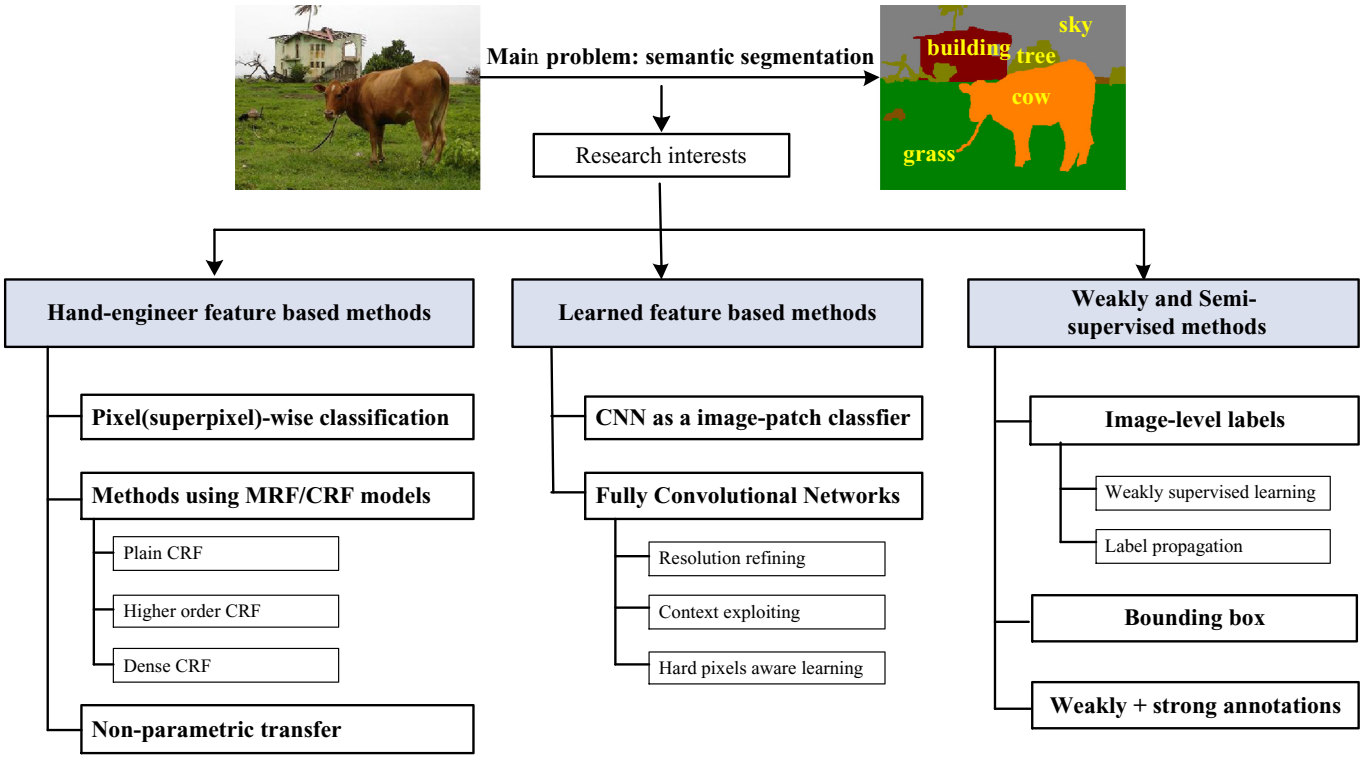

Fig. 1. Classification of existing semantic segmentation methods. According to the current research focus, existing methods can be roughly divided into three main categories. Each of them can be further classified into several sub-categories based on some key characteristics. Here we only provide a very simple description for the name of each sub-category. We refer readers to the corresponding section for more details. The image and its truth are taken from Stanford Background Dataset [32].

Networks(CNN or ConvNets) [27]. For the second issue, the most common strategy, no matter the feature used, is to use contextual models such as Markov Random Field(MRF) [28] and Conditional Random Field(CRF) [6,8,15,16,20,29–34]. These graphical models make it very easy to leverage a variety of relationships between classes via setting links between adjacent pixels. More recently, the use of Recurrent Neural Networks(RNN) [35,36] are more commonly seen in retrieving contextual information. Under the weakly supervised framework [28,37–41], another challenging issue is how to learn class models from weakly annotated images, whose labels are provided at image-level rather than pixel-level. To address this challenge, many methods resort to multiple instance learning(MIL) techniques [42].

Although there are many strategies available for addressing the problems mentioned above, these strategies are not yet mature. For example, there are still no universally accepted hand-engineered features while research on learned features has become a focus again only in recent few years. The inference of MRF or CRF is a very challenging issue in itself and often resort to approximation algorithms. Thus, new and creative semantic segmentation methods are being developed and reported continuously.

The main motivation of this paper is to provide a comprehensive survey of semantic segmentation methods, focus on analyzing the commonly concerned problems as well as the corresponding strategies adopted. Semantic segmentation is now a vast field and is closely related to other computer vision tasks. This review cannot fully cover the entire field. Since excellent reviews on research achievements on traditional image segmentation, object segmentation and object detection already exist [43,44], we will not cover these subjects. We will instead focuses on generic semantic segmentation, i.e., multi-class segmentation. Based on the observation that most works published after 2012 are CNN based, we will divide existing semantic segmentation methods into those based on hand-engineered features and learned features (see Fig. 1). We will discuss weakly supervised methods separately because this challenging line of methods is being investigated actively. It should be emphasized that there are no clear boundaries between these three categories. For each category, we further divide it into several sub-categories and then analyze their motivations and principles.

The rest of this paper is organized as follows. Before introducing recent progresses on semantic segmentation, preliminaries of commonly used theories are given in the next section. Methods using hand-engineered features and learned features are systematically reviewed in Sections 3 and 4, respectively. The efforts devoted to weakly supervised semantic segmentation are described in Section 5. In Section 6, we describe several popular datasets for semantic segmentation tasks. Section 7 compares some representative methods using several common evaluation criteria. Finally, we conclude the paper in Section 8 with our views on future perspectives. Note that semantic segmentation also called as scene labeling in literature, we will not differentiate between them in the rest of this paper.

2. Preliminaries

We start by describing the commonly used theories and technologies in the semantic segmentation community, including superpixels and contextual models.

2.1. Superpixels

As argued by Ren and Malik [45], the superpixel that consists of a set of similar and connected pixels is more appropriate for representing entities compared to the pixel. The benefits of using superpixel can be summarized into two aspects. First, the computational complexity is greatly reduced by treating a set of pixels as a single pixel, i.e., the superpixel. Second, a region is able to