Neurocomputing 000 (2018) 1–9

Contents lists available at ScienceDirect

Neurocomputing

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

Object detection via deeply exploiting depth information

Saihui Hou, Zilei Wang*, Feng Wu

CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China

ARTICLE INFO

Article history:
Received 2 June 2016
Revised 24 January 2018
Accepted 27 January 2018
Available online xxx

Communicated by Steven Hoi

Keywords:
Property derivation
Property fusion
RGB-D perception
Object detection

ABSTRACT

This paper addresses the issue on how to more effectively coordinate the depth with RGB aiming at boosting the performance of RGB-D object detection. Particularly, we investigate two primary ideas under the CNN model: property derivation and property fusion. Firstly, we propose that the depth can be utilized not only as a type of extra information besides RGB but also to derive more visual properties for comprehensively describing the objects of interest. Then a two-stage learning framework consisting of property derivation and fusion is constructed. Here the properties can be derived either from the provided color/depth or their pairs (e.g. the geometry contour). Secondly, we explore the fusion methods of different properties in feature learning, which is boiled down to, under the CNN model, from which layer the properties should be fused together. The analysis shows that different semantic properties should be learned separately and combined before passing into the final classifier. Actually, such a detection way is in accordance with the mechanism of the primary visual cortex (V_1) in brain. We experimentally evaluate the proposed method on the challenging datasets *NYUD2* and *SUN RGB-D*, and both achieve remarkable performances that outperform the baselines.

1. Introduction

Thanks to the availability of affordable RGB-D sensors, e.g. the Microsoft Kinect, the RGB-D images have been widely provided in the read-world visual analysis systems. Compared with the primitive RGB, the RGB-D can bring remarkable performance improvement for various visual tasks due to the access to the depth information complementary to RGB [1–3]. Actually, the depth has some profitable attributes for visual analysis, e.g. being invariant to illumination or color variations, and providing geometrical cues for image structures [4]. For object detection, which is one of typical complex visual tasks, the acquisition of RGB-D images is applicable and beneficial. However, how to effectively utilize the provided depth information of RGB-D images is still an open question.

Recent years have witnessed the great success of Convolutional Neural Network (CNN) in computer vision, which has boosted the performance of various visual tasks to a new level [5–7]. The CNN is generally considered as an end-to-end extractor to automatically learn discriminative features from millions of input images [8]. In this paper, we also adopt CNN to extract rich features from the RGB-D images, i.e. we are under the CNN model to investigate the exploitation of the depth information.

For the RGB-D object detection with CNN, the key is how to elegantly coordinate the RGB with depth information in feature learning. In the previous literatures, some intuitive methods have been proposed [9,10]. Roughly, we can divide them into two broad categories according to the strategy the depth is treated. The first one is to straightforwardly add the depth map to CNN as the fourth channel along with the RGB [9]. That is, the depth is processed in the same way as the RGB, and they are together convolved for granted. However, it makes no semantic sense to directly merge the depth and color maps, since they contain disparate information. The second is to process the color and depth separately, and they are combined before being fed into the final classifier, where the extracted features are joint. Specifically, two independent CNN networks are learned: one for RGB and one for depth [10]. As for the depth network, the input can be the original depth data or encoded data from the depth, e.g. height above ground, and angle with gravity [10]. It has been empirically shown that the second way usually outperforms the first one. In this paper, we further investigate how to deeply exploit the depth information with the aims of boosting the detection performance.

Before introducing the proposed method, we review the primary mechanism of human visual systems. First, multiple visual properties are always used together to describe one object when people try to recognize it, e.g. geometry contour, color, and contrast [11]. And it is usually thought that exploiting more properties is much helpful. Second, the primary visual cortex (V_1), which con-

* Corresponding author.
E-mail addresses: saihui@mail.ustc.edu.cn (S. Hou), zlwang@ustc.edu.cn (Z. Wang), fengwu@ustc.edu.cn (F. Wu).

https://doi.org/10.1016/j.neucom.2018.01.055
0925-2312/

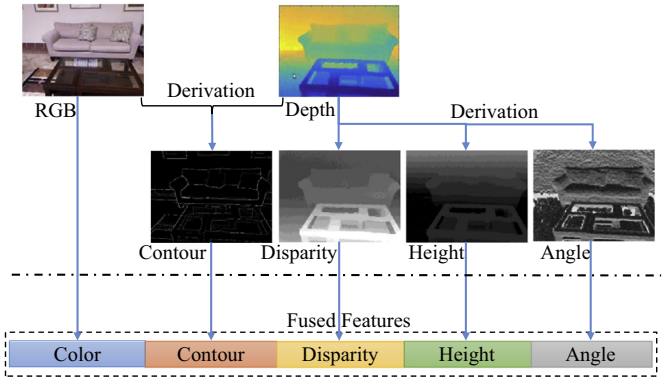

Fig. 1. Illustration of learning rich features for RGB-D object detection. Various property maps are derived to describe the objects from different perspectives. The features for these maps are learned independently and then fused for the final classification. Specifically, the derived maps include geometry contour from the color/depth pairs, and horizontal disparity, height above ground, angle with gravity from the depth data. These maps, as well as the RGB image, are sent into different CNNs for feature learning. And the features are joint before being fed into the classifier.

sists of six functionally distinct layers and is highly specialized in pattern recognition [12], abstracts different visual properties independently in the low layers and integrates in the relatively high layers.

Inspired by the working mechanism of V_1 area, we propose a novel method to deeply exploit the depth information for object detection. Fig. 1 illustrates the main idea of our method. Firstly, various visual property maps are derived through analyzing the provided color and depth pairs. It is believed that more properties can contribute to the accurate description of the objects and thus help boost the detection performance. Specifically, the derived properties include the contour, height, and angle maps. Secondly, we systematically investigate the methods to fuse different visual properties under the CNN model, i.e. how to represent a property, and from which layer the properties need to be fused together. The result of our analysis shows that the multiple properties should have complete and independent semantics in accordance with the human cognition, e.g. RGB channels should be treated as a whole to represent the color property rather than separate them from each other, and it is better to fuse the different properties after they are explicitly transformed into the high-level features.

We evaluate the proposed method on the challenging *NYUD2* and newly published *SUN RGB-D*, and the experimental results demonstrate that our method works reasonably well on both datasets and achieves remarkable performances that outperform the baselines. This is an extended version of the work that appeared in [13]. It differs from [13] in that:

- The proposed model is further evaluated on the *SUN RGB-D* with larger training data, i.e., the generalization ability of the method is validated.
- We study the effect of each involved property map and their combinations on the performance of object detection.
- A proper order is put forward to add the extra property maps in turn complementary to the RGB for the sake of boosting the detection performance, if not all of them are needed.
- More theoretical and experimental details are provided in this paper.

The remainder of the paper is organized as follows. In Section 2, we review the related works on RGB and RGB-D object detection. Section 3 provides the details of our approach, and Section 4 experimentally evaluates the proposed method. Finally, we conclude the work in Section 5.

2. Related work

Object detection [14] is to mark out and label the bounding boxes around the objects in an image. Particularly, the adopted features are critical in determining the detection performance [15]. Traditional methods, including the MRF [16] and DPM [17], are all based on the hand-crafted features such as SIFT [18] and HOG [19]. However, these features are difficult to adapt to the specific characteristics in a given task. And more recent works [20–22] have turned to the Convolutional Neural Network (CNN), which can learn discriminative features automatically from millions of RGB images. A typical CNN consists of a number of convolution and pooling layers optionally followed by the fully connected layers [8], and is able to learn multi-level features ranging from edges to entire object [23].

For object detection with CNN, a classical method is to build a sliding-window detector and then take the CNN for classification, which is usually applied on constrained object categories, e.g. faces [24] and pedestrians [15]. And the object detection is formulated as a regression problem in [20,21] then the CNN is involved in predicting the localization and labels of the bounding boxes. The most remarkable work lies in [22]. The system called R-CNN first generates around 2000 category-independent region proposals for an input image and then computes features of each region with the CNN. A category-specific SVM is appended to predict the label and score for each proposal.

When it comes to the RGB-D object detection, Gupta et al. [10] proves that CNN can also be trained to learn depth features from the depth map. In practice, the extra depth exactly makes it easy to recognize human pose [1], align 3D models [25], and detect objects [4,9,10]. Under the CNN model, two typical methods for RGB-D object detection have been proposed about how to utilize the depth information [9,10]. One is to directly add the fourth channel for depth, and then equally convolve all channels in one network [9]. The other is to separately process the depth and color (RGB) using two independent networks [10]. Obviously, these works mainly focus on the extraction of depth features, rather than considering thoroughly how to better coordinate the color and depth pairs for accurately describing the objects.

A more related work is the one by Gupta et al.[10], in which the depth data is encoded to horizontal disparity (D), height above ground (H), angle with gravity (A), and then form the three-channel DHA image into CNN to learn depth features, besides the RGB network. In our work, differently, the D, H, A are relighted and derived as new maps describing the objects from different perspectives, and used to separately learn particular types of features encoding the multiple visual properties. More than that, we propose to use the depth combining with the RGB to derive new maps to provide extra information, e.g. the geometry contour. Indeed, other properties can be also adopted, which may be obtained by specific sensors or more advanced derivation methods, e.g., colorized depth [26], distance from wall [27]. Considering the simplicity, only several directly computable properties are employed here. Furthermore, we systematically investigate the fusion ways of different properties under the CNN model. We believe that our proposed detection framework and the investigation of feature fusion would inspire more advanced works to significantly improve the performance of object detection.

3. Our approach

Intuitively, acquiring more information contents about the objects could yield more accurate recognition. Meanwhile, for human being, the visual cortex of the brain is exactly to abstract various types of visual information from the input scenes in the inception phase [12]. Inspired by such a principle, more informa-

Please cite this article as: S. Hou et al., Object detection via deeply exploiting depth information, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.01.055