2019 IEEE Winter Conference on Applications of Computer Vision

TextContourNet: a Flexible and Effective Framework for Improving Scene Text Detection Architecture with a Multi-task Cascade

Dafang He[1], Xiao Yang[2], Daniel Kifer[2], C. Lee Giles[1]

duh188,xuy111@psu.edu, giles@ist.psu.edu, dkifer@cse.psu.edu
School of Information Science and Technology, the Penn State University[1]
Department of Computer Science, the Penn State University[2]

Abstract

We study the problem of extracting text instance contour information from images and use it to assist scene text detection. We propose a novel and effective framework for this and experimentally demonstrate that: (1) A CNN that can be effectively used to extract instance-level text contour from natural images. (2) The extracted contour information can be used for better scene text detection. We propose two ways for learning the contour task together with the scene text detection: (1) as an auxiliary task and (2) as multi-task cascade. Extensive experiments with different benchmark datasets demonstrate that both designs improve the performance of a state-of-the-art scene text detector and that a multi-task cascade design achieves the best performance.

1. Introduction

Scene Text Detection extracts text information from natural images and has received an increasing amount of attention from both academia and industry. Usually as an end-to-end scene text reader first detects the location of each word in the image and then a trained recognizer reads the text for each word. For such a system, scene text detection is usually the bottleneck and much research has been devoted to improving its performance.

Following recent work [6], we categorize the scene text detection methods into three classes: (1) Proposal based scene text detection [16, 20, 25] which uses state-of-the-art object detection methods [24, 22] to classify and regress from each default proposal locations. (2) Regression based scene text detection [13, 32] which usually does direct regression on each output pixel location as opposed to relative regression from default bounding boxes. A text-nontext classification map is also needed to remove background nontext areas. (3) Segmentation based scene text detection [10, 31] which extracts text block at the first stage and uses low-level methods to obtain each individual text instances.

Due to the complex post-processing as well as the errors that incur from segmentation based methods, most recent methods follow direct regression based designs or proposal based designs. Both of them achieved breakthroughs in multi-oriented scene text detection, and state-of-the-art end-to-end performance has been achieved by combining them with a scene text recognizer [4, 19].

However, both of these categories have their drawbacks. (1) Proposal based methods are usually less accurate in terms of recovering multi-oriented bounding boxes [13]. (2) Regression based methods, though being able to recover accurately oriented bounding boxes, exhibit difficulties when the text has large variances of scales since each output bounding box is generated from a single output pixel. Usually then, a multi-scale testing is needed.

Recently, He [12] proposed to jointly learn a text attention map that suppresses background interference for a better word-level text proposal. The text attention map is essentially a text-nontext segmentation, and could be used as a mask to remove background noise. This design could be seen as a semantic segmentation based method that is used to assist text proposal generation. Such a combination also achieved better performance.

In other research, learned contour detection [3, 26, 23] has become popular. Unlike regular edge detection, learned contour information provides more instance-level semantic information.

Inspired by these two works, we also adopt the idea of using an easier task to assist scene text detection task. However, instead of using a semantic text-nontext segmentation mask to suppress background interference, we propose to learn a contour map which directly encodes the text instance information to better assist the text detection task.