Accurate Scene Text Detection through Border Semantics Awareness and Bootstrapping

Chuhui Xue[0000−0002−3562−3094], Shijian Lu[0000−0002−6766−2506], and Fangneng Zhan[0000−0003−1502−6847]

School of Computer Science and Engineering, Nanyang Technological University
xuec0003@e.ntu.edu.sg, {shijian.lu,fnzhan}@ntu.edu.sg

Abstract. This paper presents a scene text detection technique that exploits bootstrapping and text border semantics for accurate localization of texts in scenes. A novel bootstrapping technique is designed which samples multiple 'subsections' of a word or text line and accordingly relieves the constraint of limited training data effectively. At the same time, the repeated sampling of text 'subsections' improves the consistency of the predicted text feature maps which is critical in predicting a single complete instead of multiple broken boxes for long words or text lines. In addition, a semantics-aware text border detection technique is designed which produces four types of text border segments for each scene text. With semantics-aware text borders, scene texts can be localized more accurately by regressing text pixels around the ends of words or text lines instead of all text pixels which often leads to inaccurate localization while dealing with long words or text lines. Extensive experiments demonstrate the effectiveness of the proposed techniques, and superior performance is obtained over several public datasets, e. g. 80.1 f-score for the MSRA-TD500, 67.1 f-score for the ICDAR2017-RCTW, etc.

Keywords: Scene text detection, data augmentation, semantics-aware detection, deep network models

1 Introduction

Scene text detection and recognition has attracted increasing interests in recent years in both computer vision and deep learning research communities due to its wide range of applications in multilingual translation, autonomous driving, etc. As a prerequisite of scene text recognition, detecting text in scenes plays an essential role in the whole chain of scene text understanding processes. Though studied for years, accurate and robust detection of texts in scenes is still a very open research challenge as witnessed by increasing benchmarking competitions in recent years such as ICDAR2015-Incidental [19], ICDAR2017-MLT [30], etc.

With the fast development of convolutional neural networks (CNN) in representation learning and object detection, two CNN-based scene text detection approaches have been investigated in recent years which treat words or text lines as generic objects and adapt generic object detection techniques for the