An Efficient and Accurate Scene Text Detector

Hongzhi Liu

Jun 14, 2018

Abstract

Recently, extracting and understanding textual information embodied in natural scenes have become increasingly important and popular, which is evidenced by the unprecedented large numbers of participants of the ICDAR series contests and the launch of the TRAIT 2016 evaluation by NIST. Today, I read a thesis written by Xinyu Zhou, who is from Megvii Technology Inc.. His team propose a simple yet powerful pipeline that yields fast and accurate text detection in natural scenes. Experiments on standard datasets including MSRA-TD500 demonstrate that the proposed algorithm significantly outperforms state-of-the-art methods in terms of both accuracy and efficiency. On the ICDAR 2015 dataset, the proposed algorithm achieves an F-score of 0.7820 at 13.2fps at 720p resolution.

1. Overview of EAST

Scene text detection and recognition have been active research topics in computer vision for a long period of time. The core of text detection is the design of features to distinguish text from backgrounds. Conventional approaches rely on manually designed features. Stroke Width Transform (SWT) [1] and Maximally Stable Extremal Regions (MSER) [4] based methods generally seek character candidates via edge detection or extremal region extraction. In recent years, the area of scene text detection has entered a new era that deep neural network based algorithms [5] have gradually become the mainstream. However, existing methods, either conventional or deep neural network based, mostly consist of several stages and components, which are probably sub-optimal and time-consuming. Therefore, the accuracy and efficiency of such methods are still far from satisfactory.

In this paper, the team propose a efficient and accurate scene text detection pipeline that has only two stages [10]. The pipeline utilizes a fully convolutional network (FCN) model that directly produces word or text-line level predictions, excluding redundant and slow intermediate steps. The produced text predictions, which can be either rotated rectangles or quadrangles, are sent to Non-Maximum Suppression to yield final results. Compared with existing methods, the proposed algorithm achieves significantly enhanced performance, while running much faster, according to the qualitative and quantitative experiments on standard benchmarks.

2. Method of EAST

Zhou devises a deep FCN-based pipeline that directly targets the final goal of text detection: word or textline level detection. As depicted in Fig. 1(e), the model abandons unnecessary intermediate components and steps and allows for end-to-end training and optimization.

The key component of the proposed algorithm is a neural network model, which is trained to directly predict the existence of text instances and their geometries from full images. The model is a fully-convolutional neural network adapted for text detection that outputs dense per-pixel predictions of words or text lines. This eliminates intermediate steps such as candidate proposal, text region formation and word partition. The post-processing steps only include thresholding and NMS on predicted geometric shapes. The detector is named as EAST, since it is an Efficient and Accuracy Scene Text detection pipeline.

2.1. Network Design

Several factors must be taken into account when designing neural networks for text detection. Since the sizes of word regions, vary tremendously, determining the existence of large words would require features from late-stage of a neural network. A schematic view of our model is depicted in Fig. 2. The model can be decomposed in to three parts: feature extractor stem, feature-merging branch and output layer.

In Fig. 2, PVANet [3] is depicted. In their experiments, the team also adopted the well-known VGG16 [6] model, where feature maps after pooling-2 to pooling-5 are extracted. In the feature-merging branch, they gradually

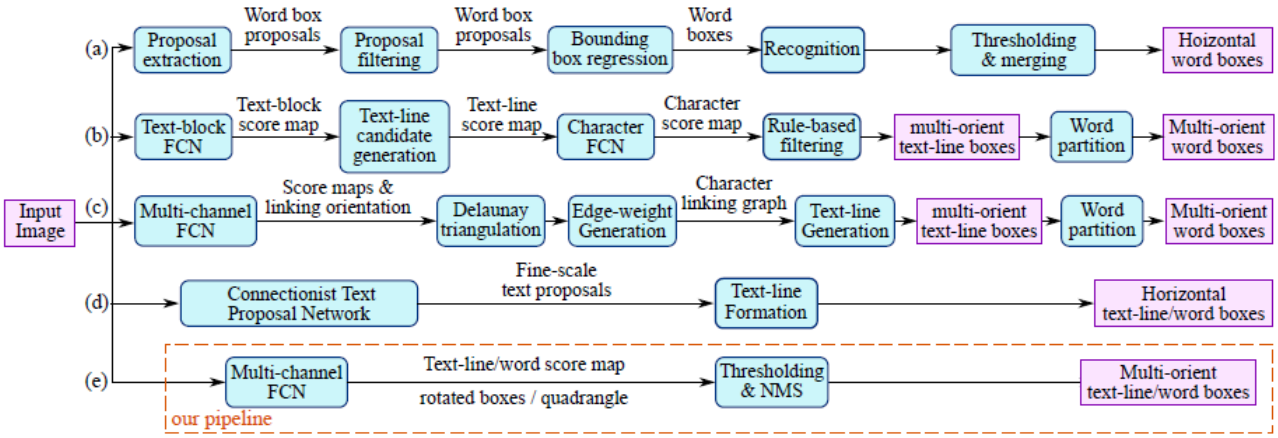

Figure 1. Comparison of pipelines of several recent works on scene text detection: (a) Horizontal word detection and recognition pipeline proposed by Jaderberg *et al.* [2]; (b) Multi-orient text detection pipeline proposed by Zhang *et al.* [9]; (c) Multi-orient text detection pipeline proposed by Yao *et al.* [8]; (d) Horizontal text detection using CTPN, proposed by Tian *et al.* [7]; (e) Their pipeline, which eliminates most intermediate steps, consists of only two stages and is much simpler than previous solutions.

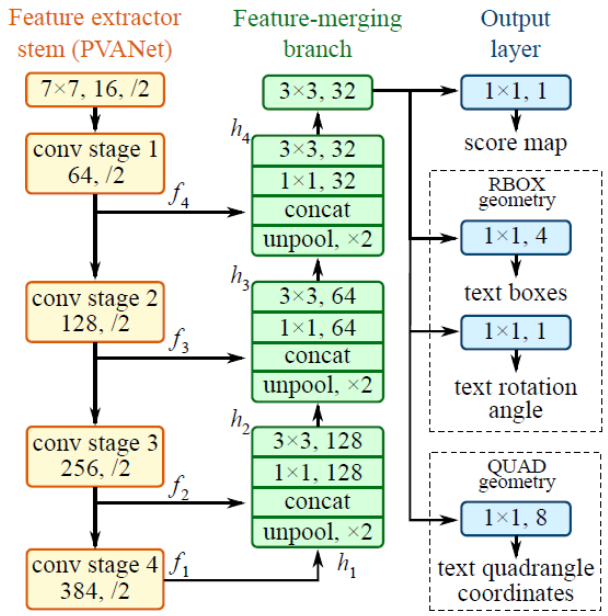

Figure 2. Structure of the text detection FCN.

merge them as Equation 1:

$$g_i = \begin{cases} unpool(h_i), & \text{if } i \leq 3, \\ conv_{3\times3}(h_i), & \text{if } i = 4. \end{cases} \tag{1a}$$

$$h_i = \begin{cases} f_i, & \text{if } i = 1, \\ conv_{3\times3}(conv_{1\times1}([g_{i-1}; f_i])), & \text{otherwise.} \end{cases} \tag{1b}$$

where g_i is the merge base, and h_i is the merged feature map. Next, a $conv_{1\times1}$ bottleneck cuts down the number of channels and reduces computation, followed by a $conv_{3\times3}$ that fuses the information to finally produce the output of this merging stage. Following the last merging stage, a $conv_{3\times3}$ layer produces the final feature map of the merging branch and feed it to the output layer.

2.2. Label Generation

Without loss of generality, Zhou only considers the case where the geometry is a quadrangle. The positive area of the quadrangle on the score map is designed to be roughly a shrunk version of the original one.

For those datasets whose text regions are annotated in QUAD style (*e.g.*, ICDAR 2015), the team first generate a rotated rectangle that covers the region with minimal area. Then for each pixel which has positive score, they calculate its distances to the 4 boundaries of the text box, and put them to the 4 channels of RBOX ground truth. For the QUAD ground truth, the value of each pixel with positive score in the 8-channel geometry map is its coordinate shift from the 4 vertices of the quadrangle.

2.3. Loss Functions

In most state-of-the-art detection pipelines, training images are carefully processed by balanced sampling and hard negative mining to tackle with the imbalanced distribution of target objects. Doing so would potentially improve the network performance. However, using such techniques inevitably introduces a non-differentiable stage and more parameters to tune and a more complicated pipeline, which contradicts our design principle.

The loss can be formulated as Equation 2:

$$L = L_s + \lambda_g L_g. \tag{2}$$

Table 1. Results on MSRA-TD500

Algorithm	Recall	Precision	F-score
Theirs + PVANET2x	0.6743	0.8728	0.7608
Theirs + PVANET	0.6713	0.8356	0.7445
Theirs + VGG16	0.6160	0.8167	0.7023
Yao et al. [8]	0.7531	0.7651	0.7591
Zhang et al. [9]	0.67	0.83	0.74

where L_s and L_g represents the losses for the score map and the geometry, respectively, and λ_g weighs the importance between two losses.

3. Evaluation of the Model

As shown in Tab. 1, on MSRA-TD500 all of the three settings of the method achieve excellent results. The F-score of the best performer (Theirs+PVANET2x) is slightly higher than that of [8]. Compared with the method of Zhang *et al.* [9], the previous published state-of-the-art system, the best performer (Theirs+PVANET2x) obtains an improvement of 0.0208 in F-score and 0.0428 in precision.

References

[1] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *CVPR*, 2010. 1

[2] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *IJCV*, 2016. 2

[3] K. H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park. PVANet: Deep but lightweight neural networks for real-time object detection. *arXiv preprint arXiv:1608.08021*, 2016. 1

[4] L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In *ACCV*, 2010. 1

[5] L. Neumann and J. Matas. Robust scene text detection with convolution neural network induced MSER trees. In *ECCV*, 2014. 1

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014. 1

[7] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network. In *ECCV*, 2016. 2

[8] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao. Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002*, 2016. 2, 3

[9] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multi-oriented text detection with fully convolutional networks. In *CVPR*, 2016. 2, 3

[10] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. EAST: An efficient and accurate scene text detector. In *CVPR*, 2017. 1