Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network

Wenhai Wang*, Enze Xie*, Xiaoge Song, Yuhang Zang, Wenjia Wang,
Tong Lu†, Gang Yu, and Chunhua Shen

ICCV 2019 Seoul, Korea

Abstract

In this paper, we propose an efficient and accurate arbitrary-shaped text detector, termed Pixel Aggregation Network (PAN), which is equipped with a low computational-cost segmentation head and a learnable post-processing. The segmentation head is made up of Feature Pyramid Enhancement Module (FPEM) and Feature Fusion Module (FFM). FPEM is a cascadable U-shaped module, which can introduce multi-level information to guide the better segmentation. FFM can gather the features given by the FPEMs of different depths into a final feature for segmentation. The learnable post-processing is implemented by Pixel Aggregation (PA), which can precisely aggregate text pixels by predicted similarity vectors. It is worth noting that our method can achieve a competitive F-measure of 79.9% at 84.2 FPS on CTW1500.

1. Proposed Method

1.1 Pipeline (see Fig. 1)

Two steps:

a) Predicting the text regions, kernels and similarity vectors by segmentation network;

b) Rebuilding complete text instances from the predicted kernels.

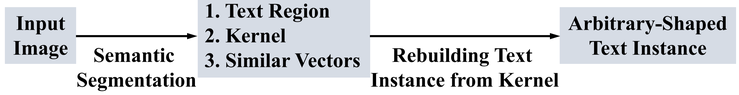

Fig. 1. The overall pipeline of PAN

1.2 Overall Architecture (see Fig. 2)

Two parts:

a) Low computational-cost segmentation head:

  a. Feature Pyramid Enhancement Module (FPEM)

  b. Feature Fusion Module (FFM)

b) Learnable post-processing

  a. Pixel Aggregation (PA)

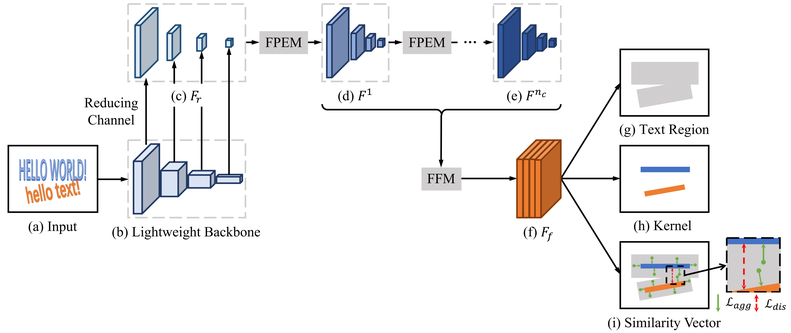

Fig. 2. The overall architecture of PAN. The features from lightweight backbone network are enhanced by a low computational-cost segmentation head which is composed of Feature Pyramid Enhancement Module (FPEM) and Feature Fusion Module (FFM). The network predicts text regions, kernels and similarity vectors to describe the text instances.

1.3 Feature Pyramid Enhancement Module (see Fig. 3)

1.4 Feature Fusion Module (see Fig. 4)

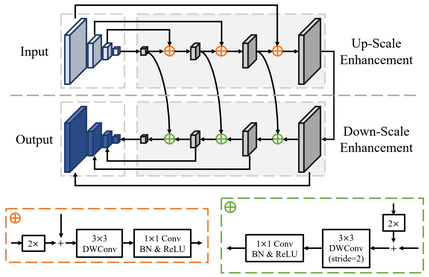

Fig. 3. The details of FPEM. "+", "2×", "DWConv", "Conv" and "BN" represent element-wise addition, 2×linear upsampling, depthwise convolution, regular convolution and batch normalization respectively.

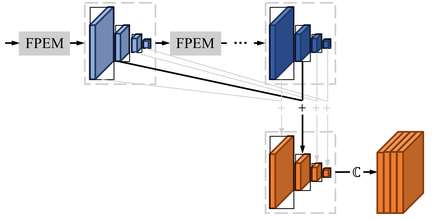

Fig. 4. The detail of FFM. "+" is element-wise addition. "ℂ" is the operation of upsampling and concatenating.

1.5 Pixel Aggregation (PA)

$$\mathcal{L}_{agg} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|T_i|}\sum_{p\in T_i}\ln(\mathcal{D}(p,K_i)+1)$$

$$\mathcal{L}_{dis} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\ln(\mathcal{D}(K_i,K_j)+1)$$

2. Results

2.1 Speed analysis (see Table 1)

Method	F	Time consumption (ms)			FPS
		Backbone	Head	Post	
PAN-320	77.10	**4.4**	**5.4**	**2.1**	**84.2**
PAN-512	80.32	6.4	7.3	3.5	58.1
PAN-640	**81.00**	9.8	10.1	5.2	39.8

Table 1. Time consumption of PAN on CTW-1500. The total time consists of backbone, segmentation head and post-processing. "F" represents the F-measure.

2.2 Performance (see Fig. 5)

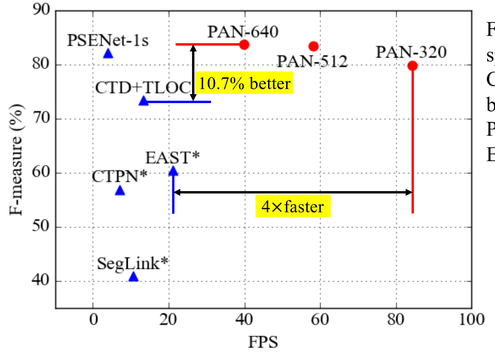

Fig. 5. The performance and speed on curved text dataset CTW1500. PAN-640 is 10.7% better than CTD+TLOC, and PAN-320 is 4 times faster than EAST.

Contact Us:
wangwenhai362@163.com
Johnny_ez@163.com