CRAFT: Character Region Awareness for Text Detection

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee *

Clova AI Research, NAVER Corp.

Baek Y, Lee B, Han D, et al. Character Region Awareness for Text Detection[J]. arXiv preprint arXiv:1904.01941, 2019.

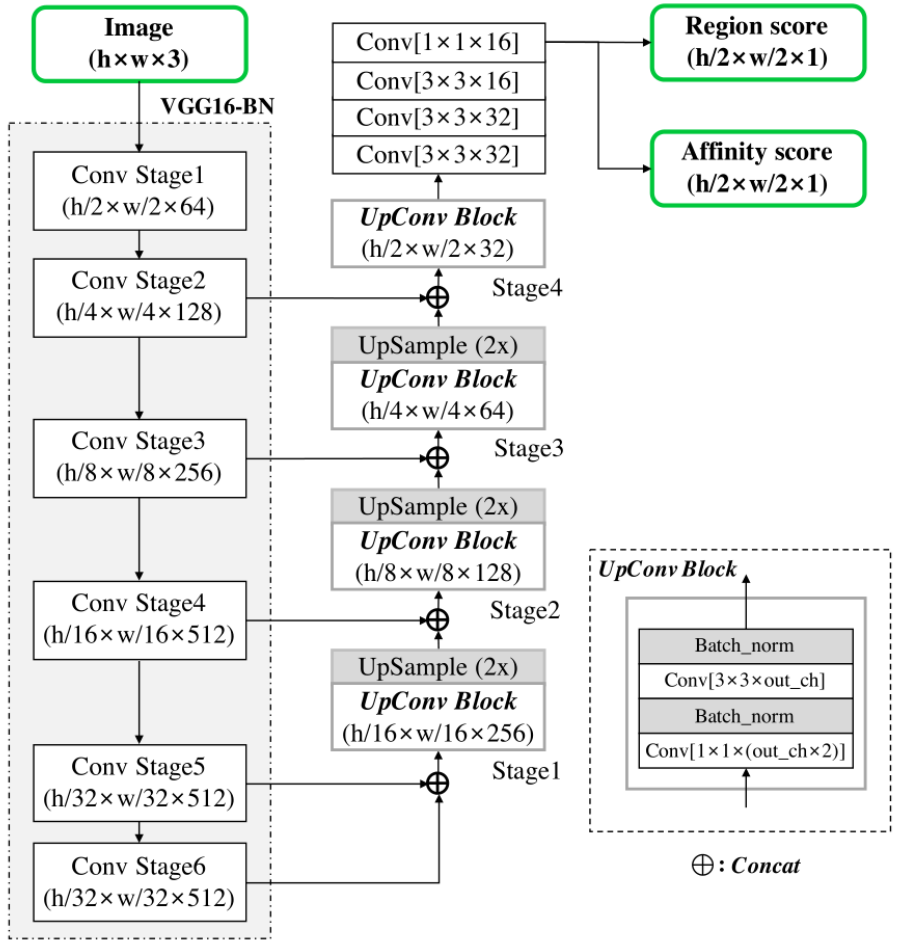
Introduction

- 提出了一种新的场景文本检测方法：
  - 以自下而上的方式，定位单个字符区域，并将检测到的字符连接成文本实例。
  - 为了克服真实图像中character-level annotation缺少的情况，利用：
    ① 合成图像中给定的character-level annotations；
    ② 真实图像中估计的character-level GT值（弱监督学习框架）。
  - 在检测复杂的场景文本图像中，有很高的灵活性。
    比如：任意形状的、弯曲的、变形的文本等。

Figure 1. Visualization of character-level detection using CRAFT. (a) Heatmaps predicted by our proposed framework. (b) Detection results for texts of various shape.

Architecture

- Objective：定位单个字符（预测字符区域和字符间关系）。
- Backbone：VGG-16 with batch normalization.
- Output：the region score and the affinity score.



Training：Ground Truth Label Generation

- 对于训练图像，需要生成GT label：
  - region score：表示像素是字符中心点的概率，用于定位单个字符；
  - affinity score：表示相邻字符间的空间的中心概率，用于对字符分组；
- 对于合成图像label生成过程：
  ① 准备二维各向同性Gaussian heatmap；
  ② 计算高斯图区域和每个字符框之间的透视变换；
  ③ 将高斯图扭曲到框区域。
- 这种GT定义方式能够使模型检测大的或者极长的文本实例。

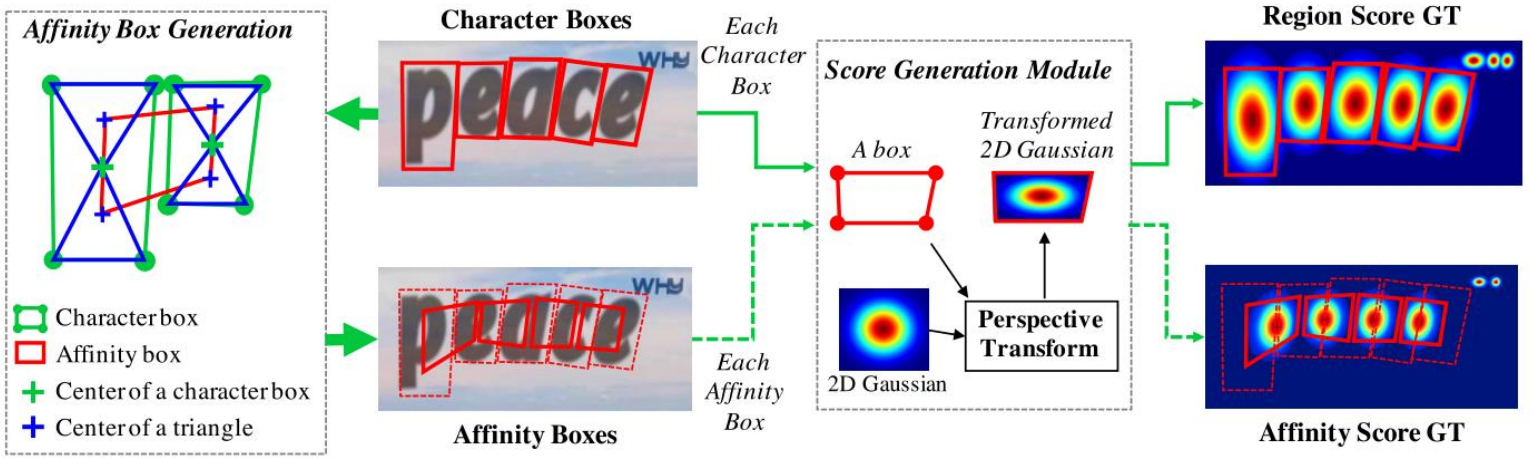


Figure 3. Illustration of ground truth generation procedure in our framework. We generate ground truth labels from a synthetic image that has character level annotations.

Training：Weakly-Supervised Learning

- 对于自然场景中的真实图像，从word-level annotation到character-level annotation的字符分割过程：
  ① 从原始图像裁剪下word-level的图像；
  ② 使用训练得到的中间模型预测region score；
  ③ 使用分水岭算法分割字符区域；
  ④ 获得字符边界框；
  ⑤ 字符框坐标变换。

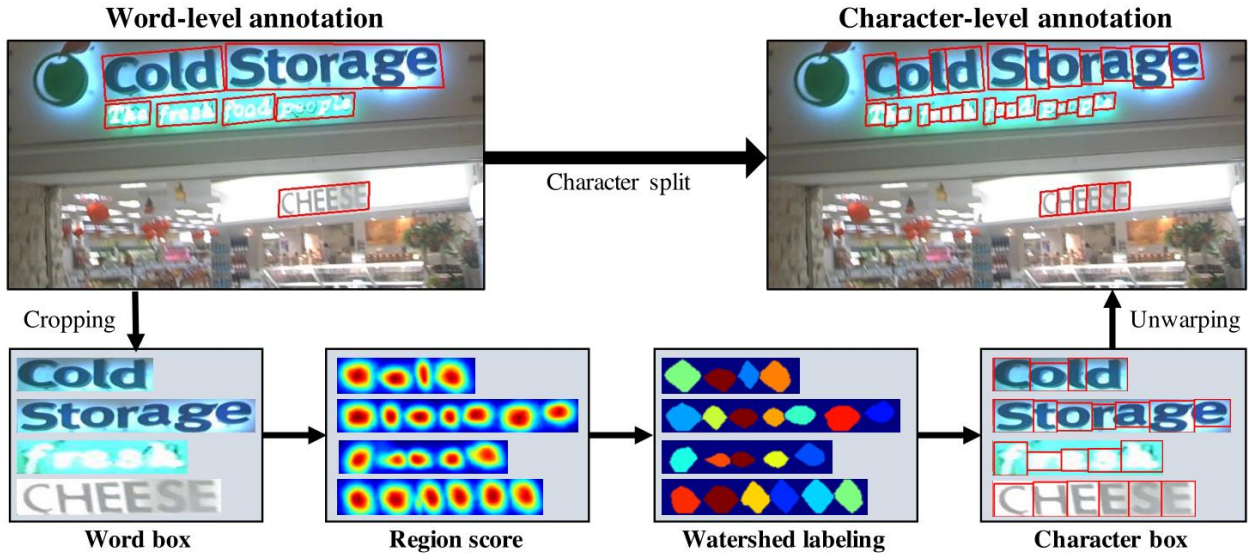


Figure 6. Character split procedure for achieving character-level annotation from word-level annotation: 1) crop the word-level image; 2) predict the region score; 3) apply the watershed algorithm; 4) get the character bounding boxes; 5) unwarp the character bounding boxes.

Training：Weakly-Supervised Learning

- 对于单词标注框w，R(w)和L(w)表示w的边界框区域和单词长度：
  - Confidence score： $s_{conf}(w) = \frac{l(w) - \min(l(w), |l(w) - l^c(w)|)}{l(w)}$,
  - 像素级的confidence map： $S_c(p) = \begin{cases} s_{conf}(w) & p \in R(w), \\ 1 & \text{otherwise}, \end{cases}$
  - 损失函数： $L = \sum_p S_c(p) \cdot \left(||S_r(p) - S_r^*(p)||_2^2 + ||S_a(p) - S_a^*(p)||_2^2 \right)$,

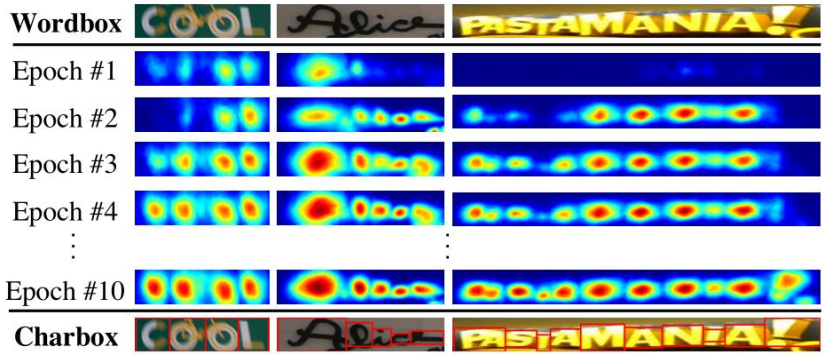


Figure 5. Character region score maps during training.

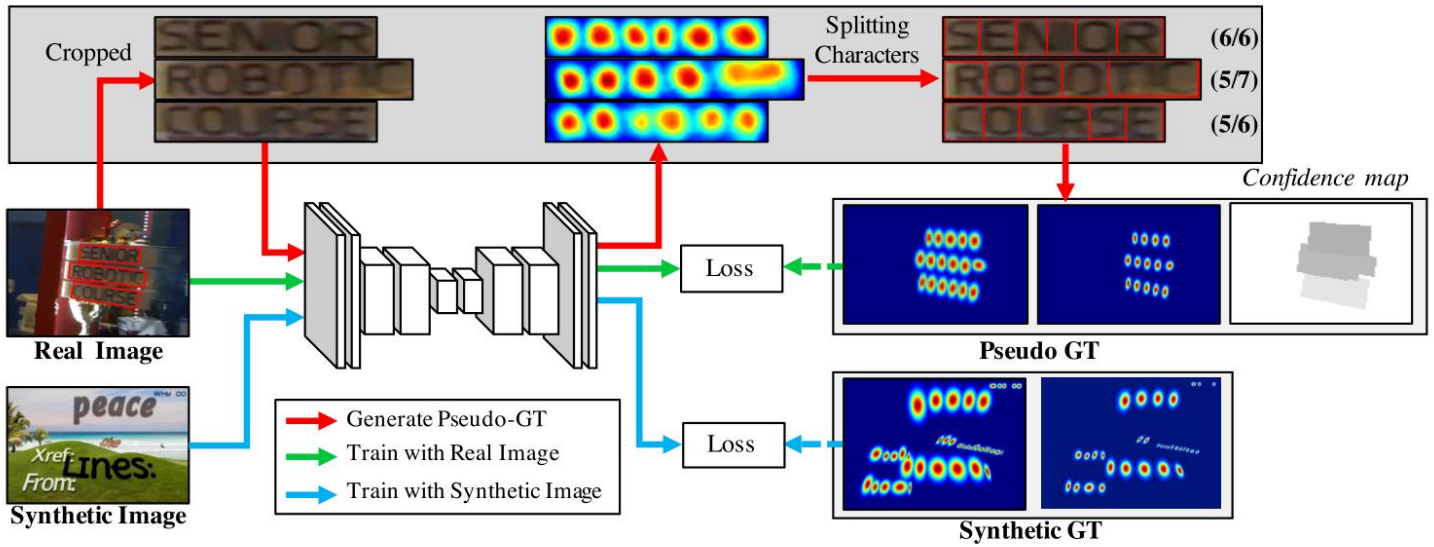


Figure 4. Illustration of the overall training stream for the proposed method. Training is carried out using both real and synthetic images in a weakly-supervised fashion.

Inference

- 在测试阶段，输出的结果可以转换成多种形状，字符框、单词框、多边形框。
- 对于矩形单词框，可以直接通过OpenCV中提供的connectedComponents()和minAreaRect()函数获得。
- 对于多边形单词框：

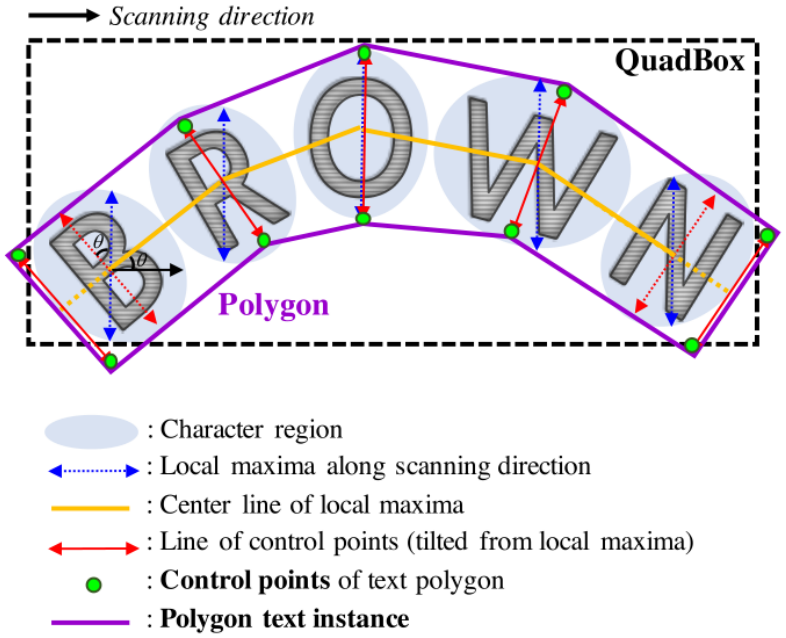


Figure 7. Polygon generation for arbitrarily-shaped texts.

Experiment and Results

- 实验数据集：ICDAR2013，ICDAR2015，ICDAR2017，MSRA-TD500，TotalText，CTW-1500
- 弱监督的训练需要两种类型的数据：
  - ① 四边形坐标标注：用于裁剪单词图
  - ② 文本内容：用于计算单词长度
  - 满足上述条件的数据集包括：ICDARs
- 只在ICDARs数据集上训练CRAFT，在其它数据集上测试（without fine-tuning）。

（一）对于四边形的数据集（ICDARs和MSRA-TD500）

Method	IC13(DetEval)			IC15			IC17			MSRA-TD500			FPS
	R	P	H	R	P	H	R	P	H	R	P	H	
Zhang et al. [39]	78	88	83	43	71	54	-	-	-	67	83	74	0.48
Yao et al. [37]	80.2	88.8	84.3	58.7	72.3	64.8	-	-	-	75.3	76.5	75.9	1.61
SegLink [32]	83.0	87.7	85.3	76.8	73.1	75.0	-	-	-	70	86	77	20.6
SSTD [8]	86	89	88	73	80	77	-	-	-	-	-	-	7.7
Wordsup [12]	87.5	93.3	90.3	77.0	79.3	78.2	-	-	-	-	-	-	1.9
EAST* [40]	-	-	-	78.3	83.3	80.7	-	-	-	67.4	87.3	76.1	13.2
He et al. [11]	81	92	86	80	82	81	-	-	-	70	77	74	1.1
R2CNN [13]	82.6	93.6	87.7	79.7	85.6	82.5	-	-	-	-	-	-	0.4
TextSnake [24]	-	-	-	80.4	84.9	82.6	-	-	-	73.9	83.2	78.3	1.1
TextBoxes++* [17]	86	92	89	78.5	87.8	82.9	-	-	-	-	-	-	2.3
EAA [10]	*87*	*88*	*88*	*83*	*84*	*83*	-	-	-	-	-	-	-
Mask TextSpotter [25]	*88.1*	*94.1*	*91.0*	*81.2*	*85.8*	*83.4*	-	-	-	-	-	-	4.8
PixelLink* [4]	87.5	88.6	88.1	82.0	85.5	83.7	-	-	-	73.2	83.0	77.8	3.0
RRD* [19]	86	92	89	80.0	88.0	83.8	-	-	-	73	87	79	10
Lyu et al.* [26]	84.4	92.0	88.0	79.7	89.5	84.3	**70.6**	74.3	72.4	76.2	87.6	81.5	5.7
FOTS [21]	-	-	*87.3*	*82.0*	*88.8*	*85.3*	*57.5*	*79.5*	*66.7*	-	-	-	23.9
CRAFT(ours)	**93.1**	**97.4**	**95.2**	**84.3**	**89.8**	**86.9**	68.2	**80.6**	**73.9**	**78.2**	**88.2**	**82.9**	8.6

（二）对于多边形的数据集（TotalText和CTW-1500）

Method	TotalText			CTW-1500		
	R	P	H	R	P	H
CTD+TLOC [38]	-	-	-	69.8	77.4	73.4
MaskSpotter [25]	55.0	69.0	61.3	-	-	-
TextSnake [24]	74.5	82.7	78.4	**85.3**	67.9	75.6
CRAFT(ours)	**79.9**	**87.6**	**83.6**	81.1	**86.0**	**83.5**

Experimental Results

Figure 8. Results on the TotalText dataset. First row: each column shows the input image (top) with its respective region score map (bottom left) and affinity map (bottom right). Second row: each column only shows the input image (left) and its region score map (right).