ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection (Supplementary)

Yuxin Wang, Hongtao Xie*, Zhengjun Zha, Mengting Xing, Zilong Fu and Yongdong Zhang
University of Science and Technology of China
{wangyx58,metingx,JeromeF}@mail.ustc.edu.cn,{htxie,zhazj,zhyd73}@ustc.edu.cn

1. Discussions about Adaptive-RPN

1.1. Comparisons between Adaptive-RPN and conventional RPN

Due to the large scale variance problem, there are some difficult conditions in scene text detection: 1) the regression distance is large (see in left of Fig.1); 2) the target box has quite different ratio to default box (see in right of Fig.1). Under these difficult conditions, the conventional RPN usually obtains a coarse localization of text region. Benefiting from the awareness of shape information and the scale-invariant training object, the proposed Adaptive-RPN performs better in these cases and achieves finer localization of text regions.

To further demonstrate the effectiveness of proposal representation in Adaptive-RPN, we use the same *IoU* loss to optimize conventional RPN. As shown in Tab.1, our Adaptive-RPN obtains better performance in recall, precision and F-measure respectively compared with conventional RPN. We attribute the improvement to the adaptive local refinement on several pre-defined points, which can automatically account for shape and semantically important local areas from ground-truth bounding box.

1.2. Comparisons between Adaptive-RPN and centernet based regression method

Centernet based regression method [2] models object as a point by keypoint estimation, and regresses a vector to represent its size. We embed this approach in our network and optimize it with the same loss as our Adaptive-RPN for fair comparison (*e.g. IoU* loss). Due to the NMS operation adopted on the center point map and only one size prediction in each position, centernet based regression method [2] is suboptimal to handle the dense predictions with large scale variance, which usually exist in scene text detection task (see in Fig.2). Compared with centernet based regression method [2], our method is proved to obtain a significant improvement in Tab.2.

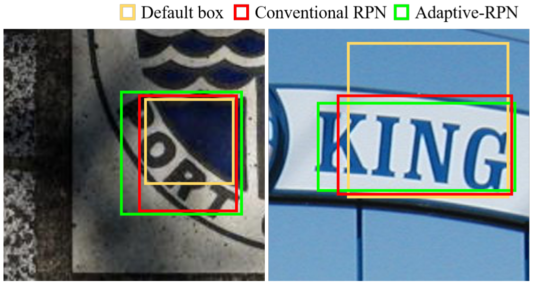

Figure 1. The qualitative examples of text localization of conventional RPN and our Adaptive-RPN.

Method	Recall	Precision	F-measure
RPN + IoU	83.3	86.1	84.7
Adaptive-RPN	**83.9**	**86.9**	**85.4**

Table 1. Performance gain of proposal representation in Adaptive-RPN on Total-Text.

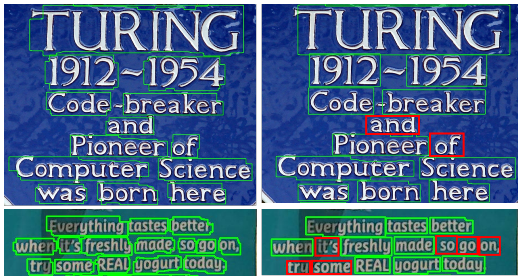

Figure 2. The detection results on Total-Text. Left: our method. Right: centernet based regression method. Green and red bounding boxes mean detected and miss-detected texts respectively.

2. Discussions about LOTM

2.1. Effectiveness of false-positive suppression.

We explore the relationship between the value of θ in *Point Re-scoring Algorithm* and the ratio of suppressed FPs to caused false negetives (FNs). As shown in Fig.3, the value of ratio is considerable when θ goes from 0.1 to 0.9. Thus, our method is much more effective in suppressing F-

*Corresponding author

Method	Recall	Precision	F-measure
Centernet [2]	74.6	86.5	80.1
Adaptive-RPN	**83.9**	**86.9**	**85.4**

Table 2. The comparison between centernet based regression method and Adaptive-RPN on Total-text.

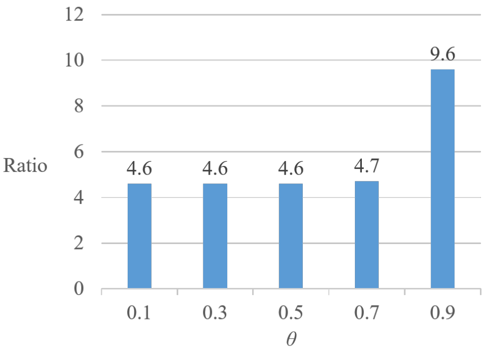

Figure 3. The relationship between the value of θ and the ratio of suppressed FPs to caused FNs on ICDAR2015.

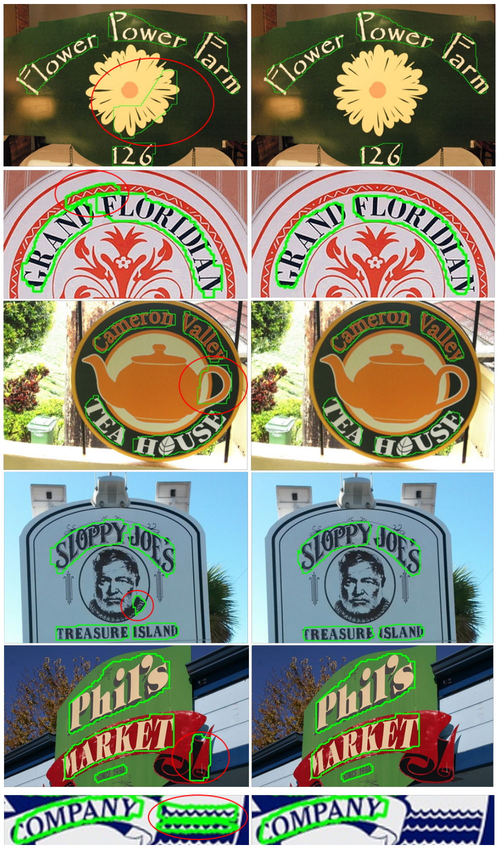

Figure 4. Effectiveness for the suppression of some complex false-positive patterns.

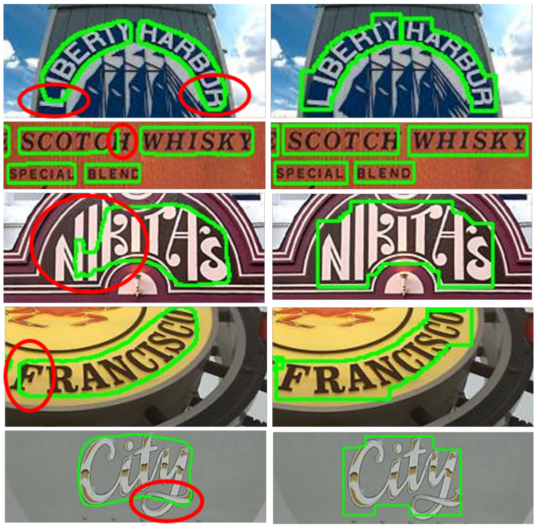

Figure 5. The qualitative examples obtained by Mask-RCNN (left) and our method (right). Our method can capture more completed text outlines in some cases.

Method	Recall	Precision	F-measure
Mask-RCNN [1]	83.7	85.1	84.4
ContourNet	**83.9**	**86.9**	**85.4**

Table 3. The comparison between Mask-RCNN and our method on Total-Text dataset.

Ps than in causing FNs. Though few FNs are caused, it is worth mentioning that the retained positive points with strong texture information in both orthogonal directions are able to accurately represent text region. Furthermore, our method is able to handle some complex false-positive patterns by implementing a high threshold in *Point Re-scoring Algorithm* (see in Fig.4), which gives a novel perspective for FP suppression.

2.2. Comparisons between segmentation-based polygon bounding box generation methods (SPBBGMs) and our method

As described in our paper, the proposed ContourNet is a contour-based polygon bounding box generation technique. Compared with the SPBBGMs (*e.g.* Mask-RCNN [1]), the differences are mainly in three aspects: **1) Perspectives**. Mask-RCNN and other SPBBGMs focus on the whole area of objects, however, our method only concerns about the contour areas which represent the shape of objects. Thus, benefiting from LOTM, our method can capture more accurate and completed text outlines in some cases (see in Fig.5) compared with Mask-RCNN [1], which is important for further text recognition. **2) Technique implementation.** Mask-RCNN and other SPBBGMs use $k \times k$ convolutional kernels in mask head for prediction, which is proved harmful for our contour-based polygon generation approach. In our method, we model the texture characteristics in two orthogonal directions with $1 \times k$ and $k \times 1$ convolutional kernels respectively and further suppress the false-positive predictions through *Point Re-scoring Algorithm*. **3) Performance.** We embed the proposed Adaptive-RPN in Mask-RCNN and train it using the same training setting as our method. As shown in Tab.3, our method achieves better performance compared with Mask-RCNN[1].

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[2] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.