Omnidirectional Scene Text Detection with Sequential-free Box Discretization

Yuliang Liu[1] , **Sheng Zhang**[1] , **Lianwen Jin**[1*] , **Lele Xie**[1] , **Yaqiang Wu**[2] and **Zhepeng Wang**[2]
[1]School of Electronic and Information Engineering, South China University of Technology, China
[2]Lenovo Inc, China
liu.yuliang@mail.scut.edu.cn; lianwen.jin@gmail.com

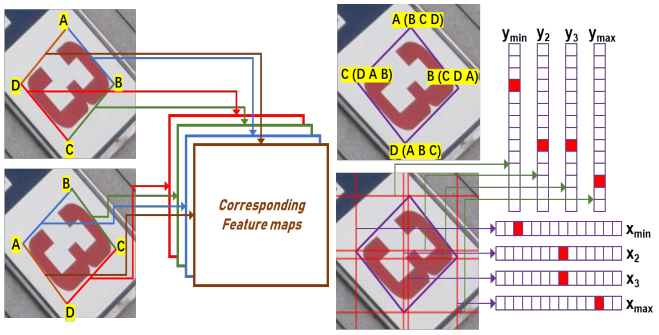
Abstract

Scene text in the wild is commonly presented with high variant characteristics. Using quadrilateral bounding box to localize the text instance is nearly indispensable for detection methods. However, recent researches reveal that introducing quadrilateral bounding box for scene text detection will bring a label confusion issue which is easily overlooked, and this issue may significantly undermine the detection performance. To address this issue, in this paper, we propose a novel method called Sequential-free Box Discretization (SBD) by discretizing the bounding box into key edges (KE) which can further derive more effective methods to improve detection performance. Experiments showed that the proposed method can outperform state-of-the-art methods in many popular scene text benchmarks, including ICDAR 2015, MLT, and MSRA-TD500. Ablation study also showed that simply integrating the SBD into Mask R-CNN framework, the detection performance can be substantially improved. Furthermore, an experiment on the general object dataset HRSC2016 (multi-oriented ships) showed that our method can outperform recent state-of-the-art methods by a large margin, demonstrating its powerful generalization ability. Source code: https://github.com/Yuliang-Liu/Box_Discretization_Network.

1 Introduction

Scene text presented in real images are often found with multi-oriented, low quality, perspective distortions, and various sizes or scales. To recognize the text content, it is an important prerequisite for detecting methods to localize the scene text tightly.

Recently, scene text detection methods have achieved significant progress [Zhou *et al.*, 2017; Liu and Jin, 2017; Deng *et al.*, 2018; Liao *et al.*, 2018a]. One reason for the improvement is that these methods introduce rotated rectangles or quadrangles instead of axis-aligned rectangles to localize the oriented instances, which remarkably improves the detection performance. However, performance of current methods still have a large gap to bridge a commercial application. Recent studies [Liu and Jin, 2017; Zhu and Du, 2018] have found that an underlying problem of introducing quadrilateral bounding box may significantly undermine the detection performance.



(a) Sensitive to label sequence. (b) Irrelevant to label sequence.

Figure 1: (a) Previous detecting methods that are sensitive to the label sequence. (b) The proposed SBD.

Taking East [Zhou *et al.*, 2017] as an example: For each pixel of the high-dimensional representation, the method utilizes four feature maps corresponding to the distances from this pixel to the ground truth (GT). It requires preprocessing steps to sort the label sequence of each quadrilateral GT box so that each predicted feature map can well focus on the targets, otherwise the detecting performance may be significantly worse. Such method is called "Sensitive to Label Sequence" (SLS), as shown in Figure 1 (a). The question is that it is not trivial to find a proper sorting rule that can avoid Learning Confusion (LC) caused by sequence of the points. The rules proposed by [Liu and Jin, 2017; Liao *et al.*, 2018a; He *et al.*, 2018] can alleviate the problem; however, they cannot avoid that a single pixel deviation of a man-made annotation may totally change the corresponding relationships between each feature map and each target of the GT.

Motivated by this issue, this paper proposes a simple but effective method called Sequential-free Box Discretization (SBD) that can parameterize the bounding boxes into key edges. Basically, to avoid LC issue, the basic idea is to find at

*Corresponding author: Lianwen Jin.

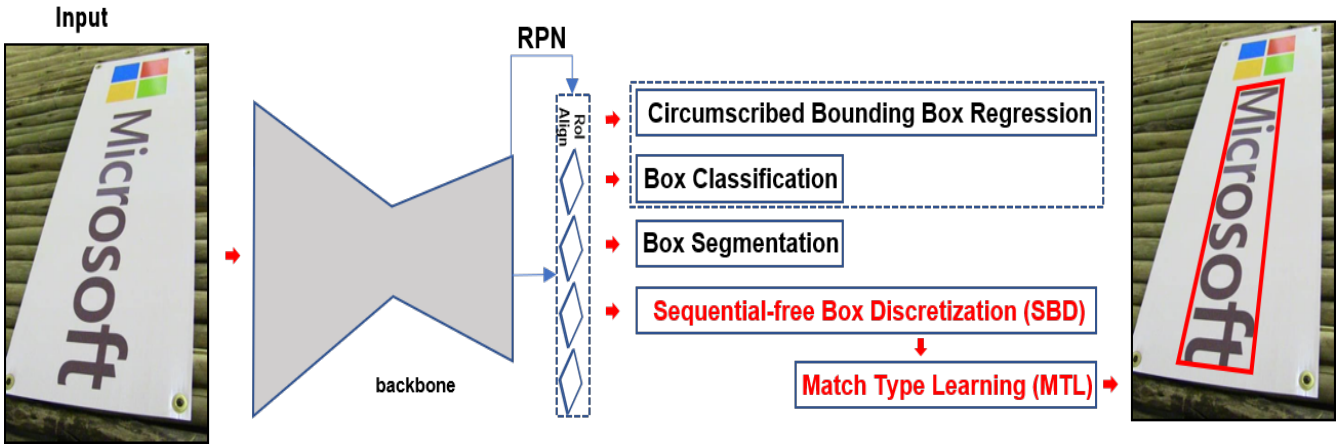


Figure 2: Overall framework. SBD is connected to the Mask R-CNN as an additional branch. The backbone is ResNet-50-FPN in this paper.

least four invariant points (e.g., mean center point, and intersecting point of the diagonals) that are irrelevant to the label sequence and we can use these invariant points to inversely deduce the bounding box coordinates. To simplify parameterization, a novel module called key edge (KE) is proposed to learn the bounding box.

Experiments on many public scene text benchmarks, including MLT [Nayef *et al.*, 2017], MSRA-TD500 [Yao *et al.*, 2012], and ICDAR 2015 Robust Reading Competition Challenge 4 "Incidental scene text localization" [Karatzas and Gomez-Bigorda, 2015], all demonstrated that our method can outperform previous state-of-the-art methods in terms of Hmean. Moreover, ablation studies showed that by seamlessly integrating SBD in Mask R-CNN framework, the detection result can be substantially improved. On multi-oriented ship detection dataset HRSC2016 [Liu *et al.*, 2017], our method can still perform the best, further showing its promising generalization ability.

The main contributions of this paper are manifold: 1) We propose an effective SBD method which can not only solve LC issue but also improve the omnidirectional text detection performance; 2) SBD and its derived post-processing methods can further guarantee tighter and more accurate detections; 3) our method can substantially improve Mask R-CNN and achieve the state-of-the-art performance on various benchmarks.

2 Related Work

The mainstream multi-oriented scene text detection methods can be roughly divided into segmentation-based methods and non-segmentation-based methods.

2.1 Segmentation-based Method

Most of segmentation-based text detection methods are mainly built and improved from the FCN [Long *et al.*, 2015] or Mask R-CNN [He *et al.*, 2017a]. Segmentation-based methods are not SLS methods because the key of segmentation-based method is to conduct pixel-level classification. However, how to accurately separate the adjacent text instances is always a tough issue for segmentation-based methods. Recently, many methods are proposed to solve this issue. For examples, PixelLink [Deng *et al.*, 2018] additionally learns 8-direction information for each pixel to highlight the text margin; [Lyu *et al.*, 2018] proposes a corner detection method to produce position-sensitive score map; and [Wu and Natarajan, 2017] defines text border map for effectively distinguishing the instances.

2.2 Non-segmentation-based Method

Segmentation-based methods require or post-processing steps to group the positive pixels into final detection results, which may easily be affected by the false positive pixels. Non-segmentation methods can directly learn the exact bounding box to localize the text instances. For examples, [Liao *et al.*, 2018b] predicts text location by using different scaled feature; [Liu and Jin, 2017] and [Ma *et al.*, 2018] utilize quadrilateral and rotated anchors to detect the multi-oriented text; [Liao *et al.*, 2018a] utilizes carefully-designed anchors to localize text instances; [Zhou *et al.*, 2017] and [He *et al.*, 2017b] directly regress the text sides or vertexes of the text instances. Although non-segmentation methods can also achieve superior performance, most of the non-segmentation methods are SLS methods, and thus they might easily be affected by the label sequence.

3 Methodology

In this section, we describe the details of the SBD. SBD is theoretically suitable for any general object detection framework, but in this paper we only build and validate SBD on Mask R-CNN. The overall framework is illustrated in Figure 2.

3.1 Sequential-free Box Discretization

The main goal of omnidirectional scene text detection is to accurately predict the compact bounding box which can be rectangular or quadrilateral. As introduced in Section 1, introducing quadrilateral bounding box can also bring the LC issue. Therefore, instead of predicting label-sensitive distances or coordinates, SBD discretizes the quadrilateral GT

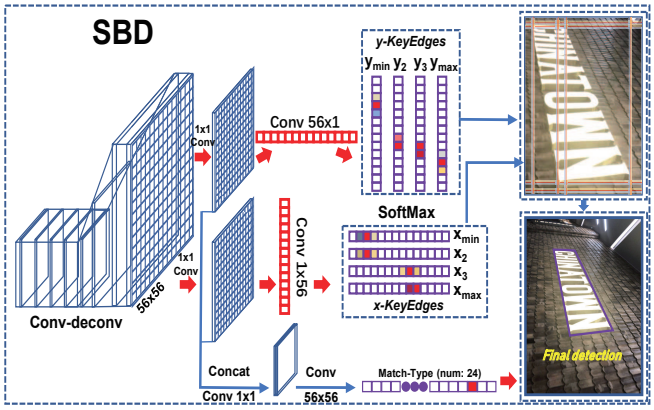


Figure 3: Illustration of SBD. The resolution M in this paper is simply set to 56.

box into 8 lines that only contain invariant points, which are called key edges (KE). As shown in Figure 3, eight KEs in this paper are discretized from the original coordinates: minimum x (x_{min}) and y (y_{min}); the second smallest x (x_2) and y (y_2); the second largest x (x_3) and y (y_3); maximum x (x_{max}) and y (y_{max}).

As shown in the Figure 2 and 3, the inputs of SBD are the proposals processed by RoIAlign [He *et al.*, 2017a]; the feature map is then connected to stacked convolution layers and then upsampled by $2\times$ bilinear upscaling layers, and the resolution of output feature maps F_{out} from deconvolution is restricted to $M \times M$. For each of the x-KEs and y-KEs, we use $1 \times M$ and $M \times 1$ convolution kernels with four output channels to shrink the transverse and longitudinal features, respectively; the number of the output channels are set to the same as the number of x-KEs or y-KEs, respectively. After that, we assign corresponding positions of the GT KEs to each output channel and update the network by minimizing the cross-entropy loss L_{KE} over a M-way softmax output. We found detection in such classification manner instead of regression would be much more accurate.

Taking t_i (t can be x or y, and i can be min, 2, 3, max) as an example, we do not directly learn the t_i-th KE; instead, the GT KE is the vertical line t_{ihalf}, and $t_{ihalf} = (t_i + t_{mean})/2$, where t_{mean} represents the t value of the mean central point of the GT box. Learning t_{ihalf} has two important advantages:

- Breaking RoI restriction. The original Mask R-CNN only learns to predict inside the RoI, and if parts of the target instances are outside the RoI, it would be impossible to recall these missing pixels. However, as shown in Figure 4, learning t_{ihalf} can output the real border even if the border is outside the RoI.
- Even if the border of the text instance is outside the RoI, in most cases, the t_{ihalf} remains inside the RoI. Therefore, the integration of the text instance can be guaranteed and loss can be well propagated (because if a learning target is outside the RoI, the loss is zero).

Formally, a multi-task loss on each foreground RoI is defined as $L = L_{cls} + L_{box} + L_{mask} + L_{ke}$. The first three terms

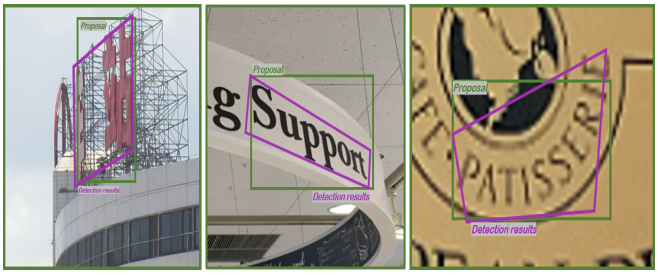


Figure 4: Detection examples that the results of SBD can break the restriction of proposal (RoI).

L_{cls}, L_{box}, and L_{mask} are the same as [He *et al.*, 2017a]. It is worth mentioning that [He *et al.*, 2017a] pointed out that the additional keypoint branch reduces the performance of box detection in Table 5; however, from our experiments, the proposed SBD is the key component for boosting detection performance, which we think is mainly because: 1) For keypoint learning, there are M^2 classes against each other, while for SBD, the number of competitive pixels is only M; 2) the keypoint might not be very explicit for a specific point (it could be a small region), while the KEs produced by SBD represent the borders of GT instances, which are absolute and exclusive, and thus the supervision information would not be confused.

Match-Type Learning

Based on the box discretization, we can learn the values of all x and y, but we do not know which y-KEs should be matched to which x-KEs. Intuitively, as shown in the top right of the Figure 3, designing a proper matching procedure is a very important issue, otherwise the detection results could be significantly worse.

To solve this problem, we propose a simple but effective match-type learning (MTL) method. As shown in Figure 3, we concatenate the x-KE and y-KE feature maps followed by 1×1, $M \times M$ convolutions, and softmax loss is used to learn a total of 24 (A_4^4) match-types (because we have 4 x-KEs and y-KEs), including {1234,1243,1324, ..., 4312, 4321}. For example, in the case of the Figure 2, the predicted match-type is "2413" which represents the matching results are (x_{min}, y_2), (x_2, y_{max}), (x_3, y_{min}), (x_{max}, y_3).

During training, we find the MTL can be very easy to learn and the loss can quickly converge within ten thousand iterations with 1 image per batch. Moreover, in some cases, the segmentation branch would somehow produce non-positive pixel while both the proposals and SBD predictions are accurate, as shown in Figure 5. Through MTL, SBD can output the final bounding box and improve the detection performance by offsetting the weakness of segmentation branch.

Rescoring and Post Processing

Based on our observations, some unconsolidated detections could also have virtual high confidence. This is mainly because the confidence outputted from the softmax in Fast R-CNN [Girshick, 2015] is a classification loss but not for localization. Therefore, the compactness of the bounding box cannot be directly supervised by the score.

Figure 5: Examples that SBD can recall many instances that segmentation branch fails to recall. Green bounding boxes and scores represent RoIs. Rotated cyan bounding box and transparent pixels represent the result from segmentation branch. Transparent pixels are predicted by mask branch. Purple quadrangles are final detection results from SBD. KEs are simplify by colorful points.

We thus compute a refined confidence that takes the advantages of SBD prediction which learns the specific position of the final detections. Formally, we refine final instance-level detection score as follow:

$$score(\Re) = \frac{(2-\gamma)S_{box} + \gamma S_{SBD}}{2}, \tag{1}$$

where, γ is the weighting coefficient, and it satisfies $0 \leq \gamma$, and $\gamma \leq 2$. S_{box} is the original softmax confidence for the bounding box, and S_{SBD} represents the mean score of all the KEs, which is defined below:

$$S_{SBD} = \frac{1}{K}\sum_{k=1}^{K} \max_{x_i} f_k(x_i), \tag{2}$$

where, K is the number of the KEs (which is 8, including 4 x-KEs and 4 y-KEs); f is the function to calculate the sum of adjacent 5 scores. We have found that using Equation (1) can not only suppress some false positives but also make the results more reliable. Examples are shown in Figure 6.

4 Experiments

4.1 Implemented Details

We used synthetic data [Gupta *et al.*, 2016] to pretrain the model and finetuned on the provided training data from MLT [Nayef *et al.*, 2017], and ICDAR 2015 [Karatzas and Gomez-Bigorda, 2015]. For MSRA-TD500 [Yao *et al.*, 2012], because the limited number of the Chinese samples, we pretrained the model from 4k well annotated samples from [Shi *et al.*, 2017a] and finetuned by official training samples.

The number of maximum iterations is 40 epochs for each dataset on four NVIDIA 1080ti GPUs. The initial learning rate is 10^{-2} and reduces to 10^{-3} and 10^{-4} on the 25th and 32th epoch, respectively. In order to balance the learning weights of all branches, the weights of KEs and match-type learning are empirically restricted to 0.2 and 0.01, respectively.

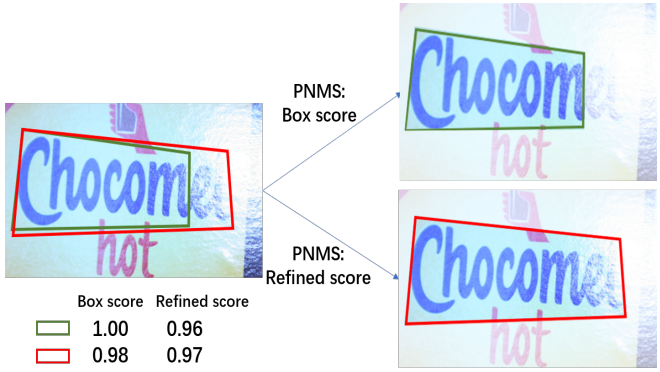


Figure 6: Example of the effect of rescoring. Original confidence is mainly for classification, while our refined score further considers the localization possibility.

The resolutions of training images were randomly selected from 600 to 920 with the interval of 40, and the maximum size was restricted to 1480. For testing, we only used single scale for all datasets (public methods have numerous settings for multi-scale testing, which is hard to conduct a fair comparison), and the scale and maximum size is (1200, 1600). Polygon non-maximum suppression (PNMS) [Yuliang *et al.*, 2017] with threshold 0.2 is used to suppress the redundant detections.

4.2 Experiments on the Scene Text Benchmarks

ICDAR 2017 MLT. [Nayef *et al.*, 2017] is the largest multi-lingual (9 languages) oriented scene text dataset, including 7.2k training samples, 1.8k validation samples and 9k testing samples. The challenges of this dataset are manifold: 1) Different languages have different annotating styles, e.g., most of the Chinese annotations are long (there is not specific word interval for a Chinese sentence) while most of the English annotations are short, the annotations of Bangla or Arabic may be frequently entwined with each other; 2) more multi-oriented, perspective distortion text on various complexed backgrounds; 3) many images have more than 50 text instances. All instances are well annotated with compact quadrangles. The results of MLT are given in Table 1. Our method outperforms previous state-of-the-art methods by a large margin, especially in terms of recall rate. Some of the detection results are visualized in Figure 7. Instead of merely using segmentation predictions to group the rotated rectangular bounding boxes, SBD can directly predict the compact quadrilateral bounding boxes which should be more reasonable. Although there are some text instances missed, most of the text can be robustly recalled.

MSRA-TD500. [Yao *et al.*, 2012] is a text-line based oriented dataset with 300 training images and 200 testing images captured from indoor and outdoor scenes. Although this dataset contains less text per image and most of the text is clean, the major challenge of this dataset is that most of the text in this dataset has the large variance in orientations. The results of MSRA-TD500 are given in Table 2. Although our

Algorithms	$R(\%)$	$P(\%)$	$H(\%)$
[Nayef *et al.*, 2017]	25.59	44.48	32.49
[Nayef *et al.*, 2017]	34.78	67.75	45.97
[Ma *et al.*, 2018]	67.0	55.0	61.0
[Ma *et al.*, 2018]	55.5	71.17	62.37
[Nayef *et al.*, 2017]	69.0	67.75	45.97
[Nayef *et al.*, 2017]	62.3	80.28	64.96
[Zhong *et al.*, 2018]	66.0	75.0	70.0
[Lyu *et al.*, 2018] (SS)	55.6	**83.8**	66.8
[Liu *et al.*, 2018] (SS)	62.3	81.86	70.75
Proposed method	**70.1**	83.6	**76.3**

Table 1: Experimental results on MLT dataset. SS represents single scale. R: Recall rate. P: Precision. H: Harmonic mean of R and P. Note that we only use single scale for all experiments.

Algorithms	$R(\%)$	$P(\%)$	$H(\%)$	FPS
[Kang *et al.*, 2014]	62.0	71.0	66.0	-
[Zhang *et al.*, 2016]	67.0	83.0	74.0	0.48
[Yao *et al.*, 2016]	75.3	76.5	75.9	1.61
[Zhou *et al.*, 2017]	67.4	87.3	76.1	**13.2**
[Shi *et al.*, 2017b]	70.0	86.0	77.0	8.9
[He *et al.*, 2017b]	70.0	77.0	74.0	1.1
[Wu and Natarajan, 2017]	78.0	77.0	77.0	-
[Deng *et al.*, 2018]	73.2	83.0	77.8	-
[Lyu *et al.*, 2018]	76.5	87.6	81.5	5.7
[Liao *et al.*, 2018b]	73.0	87.0	79.0	10
Proposed method	**80.5**	**89.6**	**84.8**	3.2

Table 2: Experimental results on MSRA-TD500 benchmark.

method is slower than some of the previous methods, it has a significant improvement in terms of the $Hmean$, which demonstrates its robustness in detecting long and strong tilted instances.

ICDAR 2015 Incidental Scene Text. [Karatzas and Gomez-Bigorda, 2015] is one of the most popular benchmarks for oriented scene text detection. The images are incidentally captured mainly from streets and shopping malls, and thus the challenges of this dataset rely on the oriented, small, and low resolution text. This dataset contains 1k training samples and 500 testing samples, with about 2k content-recognizable quadrilateral word-level bounding boxes. The results of ICDAR 2015 are given in Table 3. From this table, we can observe that our method can still perform the best.

4.3 Ablation Studies

In this section, we further conducted ablation studies to validate the effectiveness of SBD, and the results are shown in Table 5 and Figure 9. Table 5 showed that adding SBD can lead to 2.4% improvement in terms of Hmean. One reason is that the SBD can recall more instances, as discussed in Section 3 and shown in Figure 5; the other reason maybe the SBD branch can bring the effect of mutual promotion just like how segmentation branch improves the performance of the Mask R-CNN. In addition, Figure 9 showed our method can substantially outperform the baseline Mask R-CNN under different confidence thresholds of the detections, which further demonstrated its effectiveness.

We also conducted experiments to compare and validate

Algorithms	$R(\%)$	$P(\%)$	$H(\%)$
[Zhang *et al.*, 2016]	43.0	71.0	54.0
[Tian *et al.*, 2016]	52.0	74.0	61.0
[Shi *et al.*, 2017b]	76.8	73.1	75.0
[Liu and Jin, 2017]	68.2	73.2	70.6
[Zhou *et al.*, 2017]	73.5	83.6	78.2
[Hu *et al.*, 2017]	77.0	79.3	78.2
[Liao *et al.*, 2018b]	79.0	85.6	82.2
[Deng *et al.*, 2018]	82.0	85.5	83.7
[Ma *et al.*, 2018]	82.2	73.2	77.4
[Lyu *et al.*, 2018]	79.7	**89.5**	84.3
[He *et al.*, 2017b]	80.0	82.0	81.0
Proposed method	**83.8**	89.4	**86.5**

Table 3: Experimental results on ICDAR 2015 dataset. For fair comparison, this table only listed the single scale results without recognition supervision.

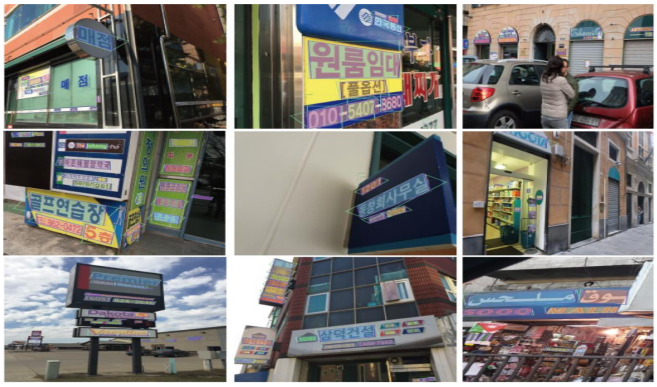


Figure 7: Examples of detection results. Purple detections are the final detection results of SBD. The transparent regions are the segmentation results from the segmentation branch, and rotated rectangles are the minimum area bounding boxes grouped by the transparent regions. Horizontal thin green bounding boxes are the rois. Zoom in for better visualization.

	Textboxes++	East	CTD	Ours
Δ Hmean	$\downarrow$ **9.7%**	$\downarrow$ **13.7%**	$\downarrow$ **24.6%**	$\uparrow$ **0.3%**

Table 4: Comparison on ICDAR 2015 dataset showing different methods' ability of resistant to the LC issue (by adding rotated pseudo samples). East and CTD are both SLS methods.

different methods' ability of resistant to the LC issue. Specifically, we first trained the East [Zhou *et al.*, 2017], CTD [Yuliang *et al.*, 2017], and proposed method with original 1k training images of ICDAR 2015 dataset. Then, we randomly rotated the training images among $[0°, 15°, 30°, ..., 360°]$ and randomly picked up additional 2k images from the rotated dataset to finetune on the these three methods. The results are given in Table 4, which demonstrated the powerful sequential-free ability of the proposed SBD.

4.4 Experiments on the Ship Detection Benchmark

To demonstrate generalization ability of SBD, we further evaluated and compared SBD on Level 1 task of the HRSC2016 dataset [Liu *et al.*, 2017] to show our method's performance on multi-directional object detection. The ship

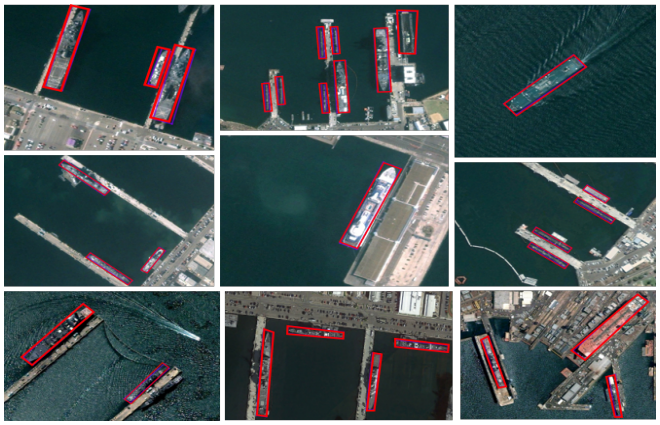


Figure 8: Experimental results on HSRC 2016. The detections are highlighted with red bounding boxes.

Datasets	Algorithms	Hmean
ICDAR2015	Mask R-CNN baseline	83.5%
	Baseline + SBD	85.9% (↑ **2.4%**)
	Baseline + SBD + Rescoring	86.5% (↑ **0.6%**)

Table 5: Ablation studies to show the effectiveness of the proposed method. The γ of rescoring is set to 1.4 (best practice).

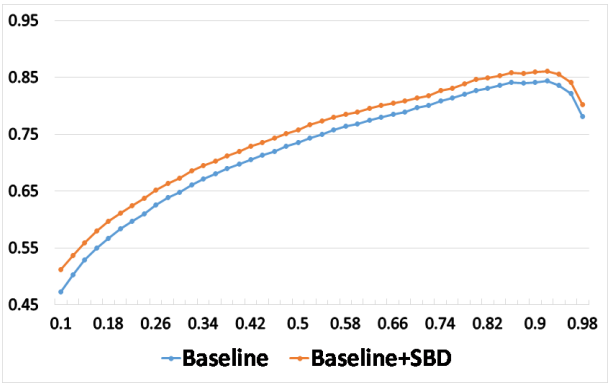


Figure 9: Ablation study on ICDAR 2015 benchmark. X-axis represents confidence threshold and Y-axis represents Hmean result. Baseline represents Mask R-CNN. By integrating with proposed SBD, the detection results can be substantially better than the results of Mask R-CNN baseline.

instances in this dataset might appear in various orientations, and annotating bounding box is based on rotated rectangles. There are 436, 181, and 444 images for training, validating, and testing set, respectively. The evaluating metric is the same as [Karatzas and Gomez-Bigorda, 2015]. Only the training and validation sets are used for training, and because of the small amount of the training data, the whole training procedure took us only about two hours.

The result showed that our method can easily surpass previous methods by a large margin (as shown in Table 6), 7.7% higher than recent state-of-the-art RRD [Liao *et al.*, 2018b] in mAP score. Some of the detection results are presented in Figure 8. Both the quantitative and qualitative results all show

Algorithms	mAP
[Girshick, 2015; Liao *et al.*, 2018b]	55.7
[Girshick, 2015; Liao *et al.*, 2018b]	69.6
[Girshick, 2015; Liao *et al.*, 2018b]	75.7
[Liao *et al.*, 2018b]	84.3
Proposed method	**93.7**

Table 6: Experimental results on HRS_2016 dataset.

that the proposed method can perform well on common oriented object detections even with very limited training data, further demonstrating its powerful generalization ability.

5 Conclusion

This paper proposed SBD - a novel method that uses discretization methodology for oriented scene text detection.

SBD solves the LC issue by discretizing the point-wise prediction into sequential-free KEs that only contain invariant points, and using a novel match-type learning method to guide the compound mode. Benefiting from SBD, we can improve the reliability of the confidence of the bounding box and adopt more effective post-processing methods to improve performance.

Experiments on various oriented scene text benchmarks (MLT, ICDAR 2015, MSRA-TD500) all demonstrate the outstanding performance of the SBD. To test generalization ability, we further conducted an experiment on oriented general object dataset HRSC2016, and the results showed that our method can outperform recent state-of-the-art methods with a large margin.

Acknowledgements

This research is supported in part by the National Key Research and Development Program of China (No. 2016YFB1001405), GD-NSF (no.2017A030312006), NSFC (Grant No.: 61673182, 61771199), and GDSTP (Grant No.:2017A010101027) , GZSTP(no. 201704020134).

References

[Deng *et al.*, 2018] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. *arXiv preprint arXiv:1801.01315*, 2018.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[Gupta *et al.*, 2016] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[He *et al.*, 2017a] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[He *et al.*, 2017b] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Deep direct regression for multi-oriented scene text detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[He *et al.*, 2018] Zheqi He, Yafeng Zhou, Yongtao Wang, Siwei Wang, Xiaoqing Lu, Zhi Tang, and Ling Cai. An end-to-end quadrilateral regression network for comic panel extraction. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 887–895. ACM, 2018.

[Hu *et al.*, 2017] Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, and Errui Ding. Wordsup: Exploiting word annotations for character based text detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[Kang *et al.*, 2014] Le Kang, Yi Li, and David Doermann. Orientation robust text line detection in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4034–4041, 2014.

[Karatzas and Gomez-Bigorda, 2015] Dimosthenis Karatzas and et al. Gomez-Bigorda, Lluis. Icdar 2015 competition on robust reading. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 1156–1160. IEEE, 2015.

[Liao *et al.*, 2018a] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 27(8):3676–3690, 2018.

[Liao *et al.*, 2018b] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5909–5918, 2018.

[Liu and Jin, 2017] Yuliang Liu and Lianwen Jin. Deep matching prior network: Toward tighter multi-oriented text detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Liu *et al.*, 2017] Zikun Liu, Jingao Hu, Lubin Weng, and Yiping Yang. Rotated region based cnn for ship detection. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 900–904. IEEE, 2017.

[Liu *et al.*, 2018] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5676–5685, 2018.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor. Darrell. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[Lyu *et al.*, 2018] Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. Multi-oriented scene text detection via corner localization and region segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7553–7563, 2018.

[Ma *et al.*, 2018] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 2018.

[Nayef *et al.*, 2017] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, volume 1, pages 1454–1459. IEEE, 2017.

[Shi *et al.*, 2017a] B. Shi, Yao, M. Liao, Yang M., Xu P., L. Cui, Lu S. Serge Belongie, and Bai X. Icdar2017 competition on reading chinese text in the wild (rctw-17). *arXiv preprint arXiv:1708.09585*, 2017.

[Shi *et al.*, 2017b] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Tian *et al.*, 2016] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *European Conference on Computer Vision*, pages 56–72. Springer, 2016.

[Wu and Natarajan, 2017] Yue Wu and Prem Natarajan. Self-organized text detection with minimal post-processing via border learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5000–5009, 2017.

[Yao *et al.*, 2012] C. Yao, X. Bai, W. Liu, and Y. Ma. Detecting texts of arbitrary orientations in natural images. In *Computer Vision and Pattern Recognition*, pages 1083–1090, 2012.

[Yao *et al.*, 2016] Cong Yao, Xiang Bai, Nong Sang, Xinyu Zhou, Shuchang Zhou, and Zhimin Cao. Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002*, 2016.

[Yuliang *et al.*, 2017] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017.

[Zhang *et al.*, 2016] Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai. Multi-oriented text detection with fully convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4159–4167, 2016.

[Zhong *et al.*, 2018] Zhuoyao Zhong, Lei Sun, and Qiang Huo. An anchor-free region proposal network for faster r-cnn based text detection approaches. *arXiv preprint arXiv:1804.09003*, 2018.

[Zhou *et al.*, 2017] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Zhu and Du, 2018] Yixing Zhu and Jun Du. Sliding line point regression for shape robust scene text detection. *arXiv preprint arXiv:1801.09969*, 2018.

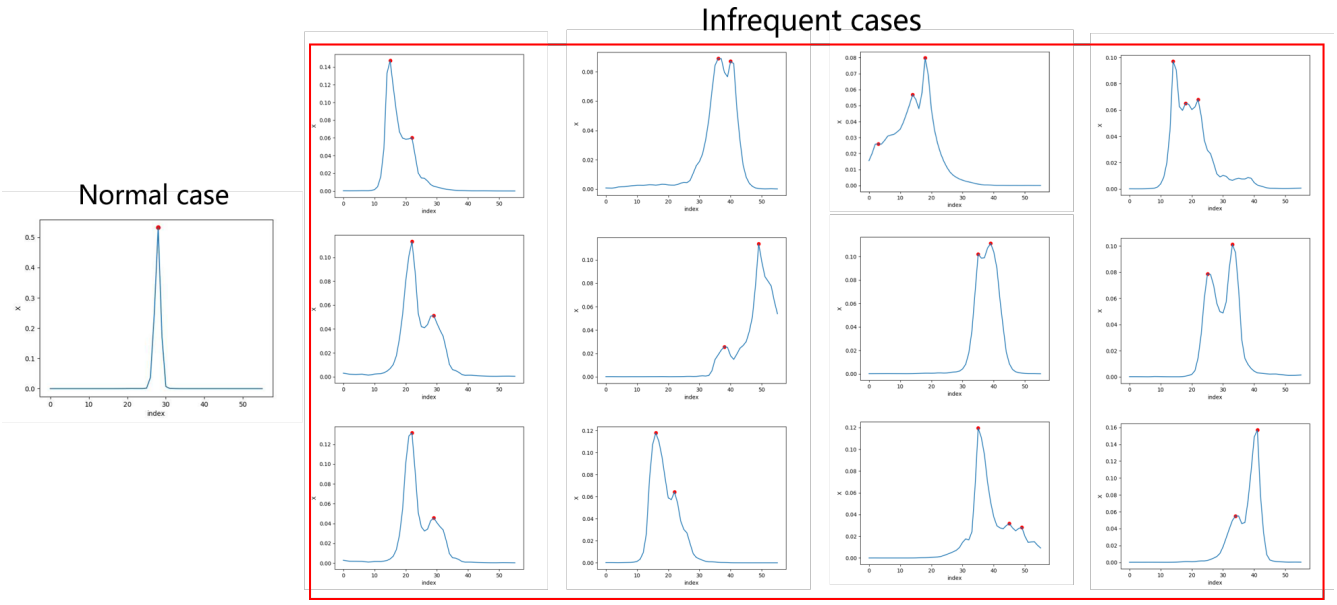


Figure 10: Examples of KE score results.

Appendix

Some additional data and figures are provided here for better understanding our method. Our method is built on MaskRCNN-benchmark[1], which is based on pytorch framework.

KE score. Figure 10 shows some example results of the ke scores. Normally, detection results will produce the similar shapes as the normal case in Figure 10. Note that even if in the normal case, the highest score may still obviously below 1.0, and that explains why we use the sum of adjacent 5 score in the rescoring operation.

False positive suppression. Match type can also be used for false positives. Because for some false positives, there is not clear edge, and in such case the match type learning may predict an abnormal result as shown in Figure 11. These abnormal results can be easily removed by judging if the quadrangle is valid (sides should only have two intersections on the head and tail). By doing so, we can further eliminate some false positives that might cheat mask branch, as shown in Figure 12.

OKS-NMS. [Papandreou *et al.*, 2017] adopted object keypoint similarity non-maximum suppression (NMS-OKS) that can be effective to suppress some unnecessary box-in-box. We can follow similar implement on our KEs detections except removing σ^2, which is because all KEs should weight the same. The formula is given as follows:

$$OKS_p = \frac{\sum_i exp\{-d_{pi}^2/2s_p^2\}\delta(v_{pi}-1)}{\sum_i \delta(v_{pi}=1)}. \qquad (3)$$

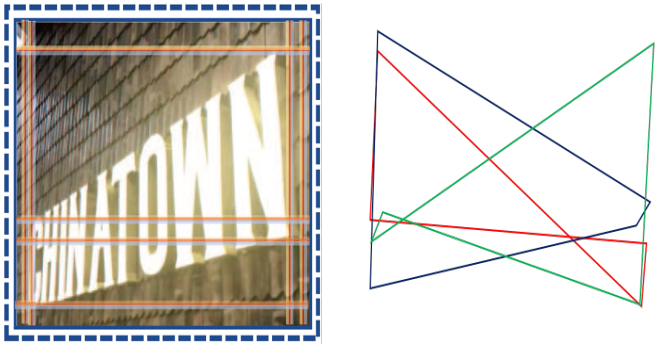


Figure 11: Examples of wrong match type results (different colors).

Figure 12: Examples of false positives suppression.

[1]https://github.com/facebookresearch/maskrcnn-benchmark