

Arbitrary-Oriented Scene Text Detection via Rotation Proposals

Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, Xiangyang Xue

Abstract—This paper introduces a novel rotation-based framework for arbitrary-oriented text detection in natural scene images. We present the Rotation Region Proposal Networks (RRPN), which is designed to generate inclined proposals with text orientation angle information. The angle information is then adapted for bounding box regression to make the proposals more accurately fit into the text region in orientation. The Rotation Region-of-Interest (RRoI) pooling layer is proposed to project arbitrary-oriented proposals to the feature map for a text region classifier. The whole framework is built upon the Faster-RCNN architecture, which ensures the computational efficiency of the arbitrary-oriented text detection comparing with previous text detection systems. We conduct experiments using the rotation-based framework on three real-world scene text detection datasets, and demonstrate its superiority in terms of effectiveness and efficiency over previous approaches.

Index Terms—Scene text detection, rotation proposals.

I. INTRODUCTION

Text detection is an important prerequisite for many computer vision and document analysis problems, which aims to identify text regions of the given images. Although there are a few commercial OCR systems for documental texts or internet contents, detect text in natural scene image is challenge due to hard situations such as uneven lighting, blurring, perspective distortion, orientation, and etc. .

In the past years there have been many attentions in the text detection task [1], [2], [3], [4], [5], [6], [7], [8], [9]. Although these approaches have shown promising results, most of them rely on horizontal or nearly-horizontal annotations and return the detection of horizontal regions. However, in real-world applications, a great number of the text regions are not horizontal, and even applying the non-horizontal aligned text lines as the axis-aligned proposals may not be accurate, so the horizontal-specific methods cannot be widely applied in practice.

Recently, a few works have been proposed to address the arbitrary-oriented text detection [10], [11], [12]. In general, there are mainly two steps included in the task of these methods, i.e. , segmentation networks such as Fully Convolutional Network (FCN) to generate the text prediction maps and geometric approaches for inclined proposals. However, prerequisite segmentation is usually time-consuming. Besides, some systems need several post-processing steps to generate

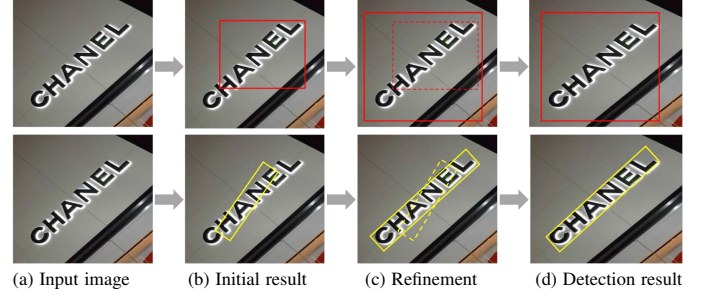


Fig. 1: Overview for text detection. First row: text detection based on horizontal bounding box. Second row: detection using rotation region proposal.

final text region proposal with orientation, which are not as efficient as those based on detection network directly.

In this paper, we come up with a rotation-based approach and an end-to-end text detection system for arbitrary-oriented text detection. Particularly, orientations are incorporated so that the detection system can generate proposals for arbitrary orientation. Comparison between the previous horizontal-based approach and ours are illustrated in Figure 1. We present the Rotation Region Proposal Networks (RRPN), which is designed to generate inclined proposals with text orientation angle information. The angle information are then adapted for bounding box regression to make the proposals more accurately fit the text region. The Rotation Region-of-Interest (RRoI) pooling layer is proposed to project arbitrary-oriented proposals to the feature map. Finally, a two-layer network is deployed to classify the regions as text or background. The main contributions of this paper include:

- Different from previous framework, ours has ability to predict the orientation of a text line, so that the proposals can better fit the text region and the ranged text region can be easily rectified and more convenient for text reading.
- The framework incorporates new components into the Faster-RCNN architecture [13], which ensures the computational efficiency for text detection comparing with segmentation-based text detection systems.
- We also proposed novel strategies for select region proposals with arbitrary orientation to improve the performance on arbitrary-oriented text detection.
- We perform the rotation-based framework on three real-world text detection datasets, i.e. , MSRA-TD500 [14], ICDAR2013 [15] and ICDAR2015 [16], and find that it is accurate and significantly more efficient than previous approaches.

J. Ma, L. Wang, and X. Xue are with School of Computer Science, Fudan University, Shanghai, China.

W. Shao, H. Ye, H. Wang, and Y. Zheng are with Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China (email: zhengyb@sari.ac.cn).

II. RELATED WORK

Reading text in the wild has been studied during the last few decades and the comprehensive surveys can be found at [17], [18], [19], [20]. Methods based on the sliding window, connected component and bottom-up strategy are designed to handle the horizontal-based text detection. Sliding window-based methods [21], [22], [4], [23], [7] tend to use a sliding window of a fixed size to slide the text area and find the region most likely to be the presence of text. In order to take more presence styles of text into account, [7], [24] applies multi-scale and multi-ratio to the sliding window methods. However, the process of sliding window leads to a large computational cost and leads to inefficiency. Representative approaches of connected component-based like Stroke Width Transform (SWT) [3] and Maximally Stable Extremal Regions (MSER) [2] were with leading performance in ICDAR 2011 [25] and ICDAR 2013 [15] robust text detection competitions. They mainly focus on the edge and pixel point of the image by detecting the character by edge detection or extreme region extraction and then combining the sub MSER components into a word or text-line region. The efforts of these methods are limited in some hard situations such as multiple connected characters, segmented stroke characters and non-uniform illumination [26].

Scene text in the wild is usually aligned from any orientation in real-world applications and approaches for arbitrary orientation are needed. For example, [27] use Mutual Magnitude Symmetry and Gradient Vector Symmetry to identify text pixel candidates regardless of any orientation including curves from natural scene image, and [28] design the Canny Text Detector by taking the similarity between image edge and text to detect text edge pixels and text localization. Recently, convolution network based approaches are proposed to perform text detection, e.g., Text-CNN [29] by first using optimized MSER detector to find the approximate region of the text and then sending region features into a character-based horizontal text CNN classifier to further recognize the character region. And the orientation factor is adopted in segmentation models by Yao et al. [11]. Their model tries to predict more accurate orientation with an explicit manner of text segmentation, and performs outstanding results on ICDAR2013 [15], ICDAR2015 [16] and MSRA-TD500 [14] benchmarks.

A technique similar to text detection is generic object detection. The detection process can be faster if the number of proposals is largely reduced. There is a wide variety of region proposal methods such as Edge Boxes [30], Selective Search [31], and Region Proposal Network (RPN) [32]. For example, Jaderberg *et al.* [33] extend the region proposal method and apply the Edge Boxes method [30] to perform text detection. Their text spotting system achieves outstanding result on several text detection benchmarks. The Connectionist Text Proposal Network (CTPN) [34] is also a detection based framework for scene text detection by employing the image feature from CNN network in LSTM to predict the text region and generate robust proposals.

This work is inspired by the Fast-RCNN [13] and RPN

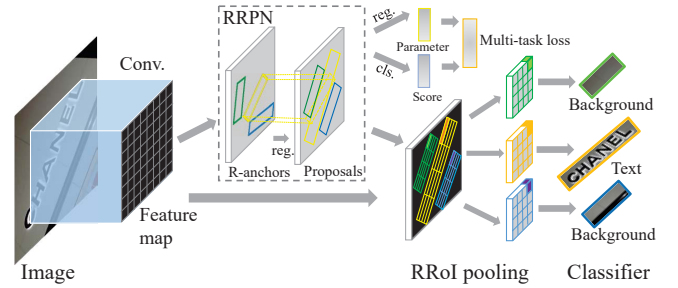


Fig. 2: Rotation based text detection pipeline.

detection pipeline. The RPN can be combined with Fast-RCNN framework to further accelerate the proposal process and achieve the state-of-the-art detection performance on competitions such as ImageNet [35]. The idea is also inspired by Spatial Transformer Networks (STN) [36], which suggests a neural network model can learn to transform an image with an angle degree; here we try to add angle information into the model for multi-oriented text detection. Perhaps the work most related to ours is [37], where the authors proposed Inception-RPN and made further text detection-specific optimization to adapt the text detection. Comparing with previous works, we incorporate the rotation factor into the region proposal network so that it is able to generate arbitrary-oriented proposals. We also extend the RoI pooling layer into the Rotation RoI pooling layer and apply angle regression in our framework to perform the rectification process and finally achieve a very good result.

III. APPROACH

We now elaborate the construction of the rotation-based framework. The architecture is illustrated in Figure 2; details of each phase will be described in the following parts.

A. Horizontal Region Proposal

As mentioned in previous section, RPN [32] is able to further accelerate the process of proposal generation. Part of VGG-16 [38] is employed as sharable layers, and the horizontal region proposals are generated by sliding over the feature map of the last convolutional layer. The feature extracted from each sliding window are fed into two sibling layers (a box-regression (reg) layer and a box-classification (cls) layer), where there are $4k$ (4 coordinates per proposal) outputs from reg layer representing coordinates and $2k$ (2 scores per proposal) scores from cls layer for k anchors of each sliding position.

To fit the objects in different sizes, RPN uses two parameters to control the size and shape of anchors, i.e., scale and aspect ratio. Scale parameter decides the size of the anchor, and aspect ratio controls the ratio of width and height for anchor box. In [32], the authors set the scale as 8, 16 and 32 and the ratio as 1:1, 1:2 and 2:1 for generic object detection task. This anchor selection strategy can cover almost shapes of all natural objects and keep the total proposals at a small number. However, in the text detection task, especially with the scene images, texts are usually presented in an unnatural shape

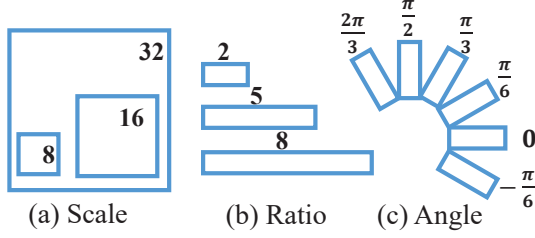


Fig. 3: Anchor strategy in our framework.

with different orientations; axis-aligned proposals generated by RPN are not robust for scene text detection. To make the network more robust for text detection and keep its efficiency, we think that it is necessary to build a detection framework, which encodes the rotation information with the region proposals.

B. Proposed Architecture

As shown in Figure 2, we employ the convolutional layers of VGG16 [38] in the front of the framework, which are shared by two sibling branches, i.e., the RRPN and a clone of feature map of last convolutional layer. RRPN generates arbitrary-oriented proposals for text instances and further performs bounding box regression for proposals to better fit the text instances. The sibling layers branched from the RRPN are the classification layer (cls) and regression layer (reg) of RRPN, respectively. Outputs from these two layers are the scores from cls and proposal information from reg and their losses are computed and summed to form multi-task loss. Then RRoi pooling layer acts as a max pooling layer by projecting arbitrary-oriented text proposals from RRPN onto the feature map. Finally, a classifier formed by two fully-connected layers is used and the region with the RRoi features is classified as text or background.

C. Data Preprocessing

In the training stage, the ground truth of a text region is represented as a rotated bounding box with 5 tuples (x, y, h, w, θ) . The coordinate (x, y) represents geometric center of the bounding box. The height h is set as short side of bounding box and the width w as the long side. The orientation θ is the orientation of long side. For convenience, we fix the range to $(-\frac{\pi}{4}, \frac{3\pi}{4}]$ by shifting to the opposite direction if the angle is out of this range.

D. Rotation Anchors

Traditional anchors using scale and aspect ratio parameters are not enough for in-the-wild text detection, therefore we design the rotation anchors (R-anchors) with several adjustments. First, orientation parameter is added to control the orientation of a proposal. Six different orientations, i.e., $-\frac{\pi}{6}$, 0 , $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$ and $\frac{2\pi}{3}$, are used, which are trade-off between orientation coverage and computational efficiency. Second, as text regions are usually with special shapes, the aspect ratio is changed to 1:2, 1:5 and 1:8 to cover a wide range of text line. Besides,

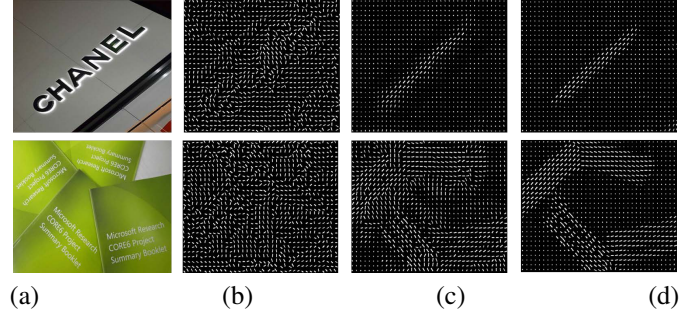


Fig. 4: Visualization of different multi-task loss values. The orientation is the direction of the white line on each point, and longer lines indicate higher response score for text. (a) Input images; (b) 0 iteration; (c) 15,000 iterations; (d) 150,000 iterations.

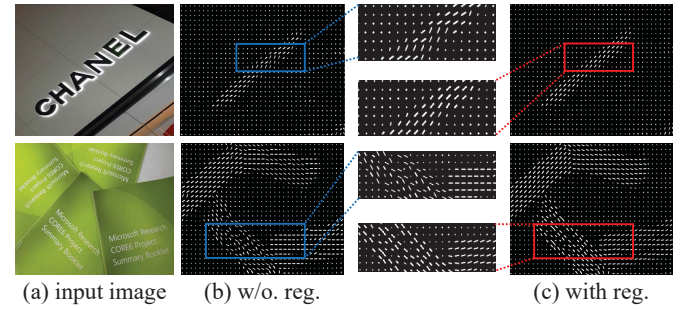


Fig. 5: Visualization of the impact for regression. (a) Input images; (b)(c) orientation and response of the anchors without regression (b) and with regression (c).

the scale of 8, 16 and 32 are kept. Summary of the anchor strategy is in Figure 3. Following our data preprocess step, a proposal is generated from the R-anchors with 5 variables (x, y, h, w, θ) . For each point on feature map, 54 R-anchors ($6 \times 3 \times 3$) are generated, and respectively 270 outputs (5×54) for reg layer and 108 scores output (2×54) for cls layer on each sliding position. Then, we slide the feature map with RRPN and generates $H \times W \times 54$ anchors in total for the feature map with width W and height H .

E. Learning of Rotated Proposal

As the R-anchors are generated, the sampling strategy for R-anchors is needed for network learning. We first define the Intersection-over-Union (IoU) overlap as the overlap between the skew rectangles of ground truth and R-anchor. Then Positive R-anchors are featured with: (i) the highest IoU overlap or IoU larger than 0.7 with ground truth, and (ii) intersection angle with the ground truth less than $\frac{\pi}{12}$. Negative R-anchors are defined as: (i) IoU lower than 0.3, or (ii) IoU larger than 0.7 but the intersection angle with ground truth larger than $\frac{\pi}{12}$. Regions which are not positive or negative are not used during training.

Our loss function for the proposal uses the form of multi-task loss [13], which is defined as:

$$L(p, l, v^*, v) = L_{\text{cls}}(p, l) + \lambda L_{\text{reg}}(v^*, v) \quad (1)$$

where l is the indicator of the class label ($l = 1$ for text and $l = 0$ for background, no regression for background), the parameter $p = (p_0, p_1)$ is the probability over classes computed by the softmax function, $v = (v_x, v_y, v_h, v_w, v_\theta)$ is the tuple of ground truth bounding box, and $v^* = (v_x^*, v_y^*, v_h^*, v_w^*, v_\theta^*)$ is the predicted tuple for the text label. The trade-off between two terms are controlled by the balancing parameter λ . We define the classification loss for class l as:

$$L_{cls}(p, l) = -\log p_l \quad (2)$$

For the bounding box regression, the background RoIs are ignored and we adopt smooth- L_1 loss for the text RoIs:

$$L_{reg}(v^*, v) = \sum_{i \in \{x, y, h, w, \theta\}} \text{smooth}_{L_1}(v_i^* - v_i) \quad (3)$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \quad (4)$$

The scale-invariant parameterizations of a tuple $t = (t_x, t_y, t_h, t_w, t_\theta)$ are adopted as follows (t can be v or v^*):

$$\begin{aligned} v_x &= \frac{x - x_a}{w_a}, v_y = \frac{y - y_a}{h_a} \\ v_h &= \log \frac{h}{h_a}, v_w = \log \frac{w}{w_a}, v_\theta = \theta - \theta_a \end{aligned} \quad (5)$$

$$\begin{aligned} v_x^* &= \frac{x^* - x_a}{w_a}, v_y^* = \frac{y^* - y_a}{h_a} \\ v_h^* &= \log \frac{h^*}{h_a}, v_w^* = \log \frac{w^*}{w_a}, v_\theta^* = \theta^* - \theta_a \end{aligned} \quad (6)$$

where x, x_a and x^* are for the predicted box, anchor and ground-truth box; and it is the same for y, h, w and θ .

As we describe in the prior section, we give R-anchors fixed orientations within the range $[-\frac{\pi}{4}, \frac{3\pi}{4}]$ and each of the 6 orientations can fit the ground truth that has an intersection angle less than $\frac{\pi}{12}$, so that every R-anchor has its fitting range, which we call it fit domain. When orientation of a ground truth box is in the fit domain of an R-anchor, this R-anchor is most likely to be a positive sample of the ground truth box. As a result, the fit domains of 6 orientations divide the angle range $[-\frac{\pi}{4}, \frac{3\pi}{4}]$ into 6 equal parts. Thus ground truth in any orientation can be fitted with an R-anchor of appropriate fit domain. Figure 5 shows a comparison of the utility of regression term. We can observe that the orientations of the regions are similar in a neighborhood region.

In order to verify the ability of the network for learning text region orientation, we visualize the intermediate results in Figure 4. For an input image, the feature maps of RPN training after different iterations are visualized. The short white line on feature map represents the R-anchor with the highest response to the text instance. The orientation of short line is orientation of this R-anchor, while short line length indicates the level of response. We can observe that brighter field of feature map focuses on the text region and other region becomes darker after 150,000 iterations. Moreover, the orientations of the regions become closer to the orientation of text instance as the iteration increases.

F. Accurate Proposal Refinement

Skew IoU Computation The rotation proposals can be generated in any orientations, so IoU computation on axis-aligned proposals may lead to an inaccurate IoU of skew interactive

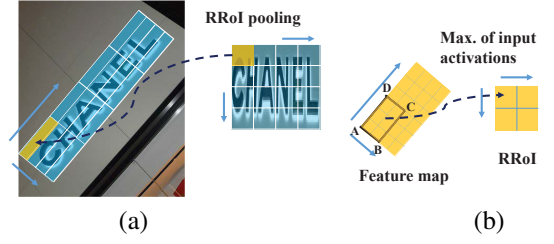


Fig. 6: RRoI pooling layer: (a) divide arbitrary-oriented proposal into subregions; (b) max pooling a single region from a inclined proposal to a point in RRoI.

Algorithm 1 RRoI pooling

```

1: RRoI_width  $\leftarrow W_r$ 
2: RRoI_height  $\leftarrow H_r$ 
3: RRoI_map[1, RRoI_height][1, RRoI_width]  $\leftarrow 0$ 
4: for  $i \in [1, \text{RRoI\_height}], j \in [1, \text{RRoI\_width}]$  do
5:    $A, B, C, D \leftarrow \text{affineTransformation}(\text{subRegion}(i, j))$ 
6:    $\text{RRoI\_map}[i][j] \leftarrow \max(\text{featureMap}(A, B, C, D))$ 
7: end for
8: return RRoI_map

```

proposals and further ruin the proposal learning. Therefore we design an implementation for Skew IoU computation with the thought of triangulation [39] and Figure 7 shows the geometric principles.

Skew Non-Maximum Suppression (Skew-NMS) Traditional NMS only takes the IoU factor into consideration (e.g. IoU threshold is 0.7), but it is insufficient for arbitrary-oriented proposals. For instance, anchor with ratio 1:8 and angle difference of $\frac{\pi}{12}$ are with the IoU of 0.31, which are less than 0.7; however, it may be annotated as positive samples. Therefore, the Skew-NMS consists of 2 phases: (i) keep the max IoU for proposals with IoU larger than 0.7; (ii) if all proposals with an IoU in range $[0.3, 0.7]$, keep the proposal with minimum angle difference with ground truth (the angle difference shall be less than $\frac{\pi}{12}$).

G. RRoI Pooling Layer

As presented in Fast-RCNN [13], RoI pooling layer extracts a fixed-length feature vector from the feature map for each proposal. Each feature vector is feed into fully-connected layers that finally branch in to the sibling cls and reg layers and the outputs are predicted localization and class of an object in an input image. As the feature map of image only needs to be computed once per image rather than computed every generated proposal, the object detection framework is accelerated. The RoI pooling layer uses max pooling to convert the feature inside any valid RoI into a small feature map with a fixed spatial extent of $h_r \times w_r$, where h_r and w_r are layer hyper-parameters that are independent of any RoI.

For the arbitrary-oriented text detection task, the traditional RoI pooling layer can only handle axis-aligned proposals, thus we present the Rotation RoI (RRoI) pooling layer to adjust arbitrary-oriented proposals generated by RPN. We first set the RRoI layer hyper-parameters to H_r and W_r .

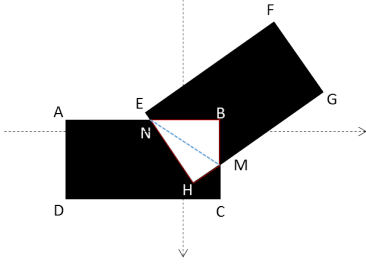


Fig. 7: IoU computation of two rectangles in arbitrary orientation. First find all the vertices of the intersection region, and divide the polygon into triangles (triangulation) and summarize all the area of the triangles. $Area_{intersection} = Area_{\triangle NHB} + Area_{\triangle NMB}$

for the RRoIs, the rotated proposal region can be divided into $H_r \times W_r$ subregions of $\frac{h}{H_r} \times \frac{w}{W_r}$ size for a proposal with height h and width w (as shown in Figure 6(a)); each subregions have the same orientation as the proposal. Figure 6(b) displays an example with 4 vertices (A, B, C, and D) of the subregion on the feature map. The 4 vertices are calculated using affine transformation and grouped to range the border of subregion. Then max pooling is performed in every subregion and max-pooled values are saved in the matrix of each RRoI; the pseudo-code for RRoI pooling is shown in Algorithm 1. Finally, the proposals are transferred into RRoIs and sent to the classifiers to give the result of text or background.

IV. EVALUATION

We perform the rotation-based framework on three popular text detection benchmarks: MSRA-TD500 [14], ICDAR2015 [16] and ICDAR2013 [15]; we follow the evaluation protocol of both benchmarks. MSRA-TD500 dataset contains 300 training images and 200 testing images. Annotations of the images consist of both position and orientation of each text instance, and the benchmark can be used to evaluate text detection performance over the multi-oriented text instance. ICDAR 2015 is released for the text localization of incidental scene text challenge (Task 4.1) in ICDAR 2015 Robust Reading Competition, which has 1500 images in total. The images are acquired using Google Glass in first-person view and some text instances in these images suffer from blurred, skewed or not well-placed issues. Different from previous ICDAR Robust Reading Competition, the text instance annotations are with four vertices, which form an irregular quadrilateral bounding box with orientation information. We roughly generate an inclined rectangle to fit the quadrangle and its orientation. The ICDAR 2013 dataset is from the ICDAR 2013 Robust Reading Competition. There are 229 natural images for training and 233 natural images for testing. All the text instances in this dataset are in English and are horizontally aligned. We conduct experiments on this horizontal-based benchmark to see the adaptability of our approach on specific orientation.

Model training. For the MSRA-TD500 experiment, we use its training set (contains 300 images) as well as the HUST-TR400 (contains 400 images with text instance annotated with the same parameters as MSRA-TD500) [40] for training. For

a.	b.	c.	d.	e.	P	R	F	ΔF
Faster-RCNN [32]					38.7%	30.4%	34.0%	-
Baseline					57.4%	54.5%	55.9%	-
✓					65.6%	58.4%	61.8%	5.9%
	✓				63.3%	58.5%	60.8%	4.9%
		✓			63.1%	55.4%	59.0%	3.1%
✓	✓	✓			68.4%	58.9%	63.3%	7.4%
✓	✓	✓	✓		71.8%	67.0%	69.3%	13.4%
✓	✓	✓	✓	✓	82.1%	67.7%	74.2%	18.3%

TABLE I: Evaluation on MSRA-TD500 with different strategies and settings. Experiments of Faster-RCNN are based on the original source code. P, R and F are abbreviation of Precision, Recall, and F-measure respectively. ΔF is the improvement of F-measure over baseline. The strategies include: a. context of text region; b. training dataset enlargement; c. border padding; d. scale jittering; e. post processing.

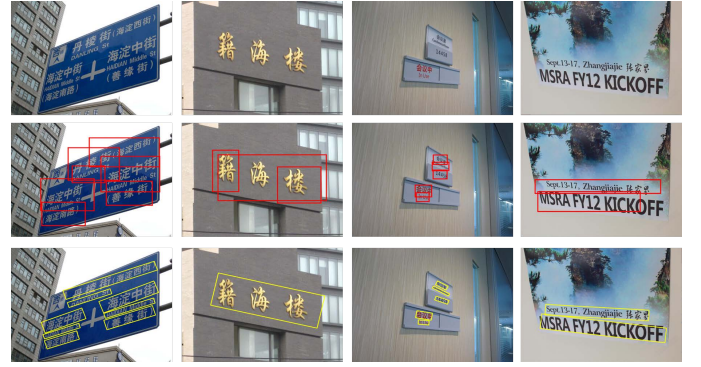


Fig. 8: Comparison of rotation and horizontal region proposal. First row: original images; second row: text detection based on horizontal region proposal; third row: text detection based on rotation region proposal.

ICDAR2015 benchmark, both the training set of ICDAR2015 and previous ICDAR competition datasets [41], [15] are also used. As the training set may not be large enough for the network to fully learn the characteristic of arbitrary orientation, we further make the data argumentation by rotating the training image patches to ensure that text instance in any orientation could be learnt. All of the experiments follow the standard evaluation protocols with the measurements of Precision, Recall, and F-measure. The training set of ICDAR2013 follows training set of ICDAR2015 experiments.

Implementation details. Our network is initialized by pre-training of a model for ImageNet classification [38]. The weights of the network are updated by using a learning rate of 10^{-3} for the first 200,000 iterations and 10^{-4} for next 100,000 iterations, weight decay of 5×10^{-4} and a momentum of 0.9.

Due to our different R-anchor strategy, the total number of proposals for each image is nearly 6 times of that in previous approaches such as Faster-RCNN. In order to make the efficient detection, we filter R-anchors of breaking the border of image. Therefore, the speed of our system is similar with previous works in both training and testing stage.

Factor	Precision	Recall	F-measure
1.0	57.4%	54.5%	55.9%
1.2	59.3%	57.0%	58.1%
1.3	66.0%	56.8%	61.1%
1.4	65.6%	58.4%	61.8%
1.5	63.8%	55.0%	59.1%
1.6	63.8%	56.8%	60.1%
1.7	66.3%	56.6%	61.1%

TABLE II: Exploit the text region context by enlargement of the text bounding box with different factor of original size.



Fig. 9: Examples of the failed detection on MSRA-TD500. The red box is the detection result and the green box is groundtruth.

A. MSRA-TD500

We train the baseline system with 300 images from MSRA-TD500 training set; the input image patch is resized with long side of 1000 pixels. The evaluation result is precision of 57.4%, recall of 54.5%, and F-measure of 55.9%, which are much better performance than original Faster-RCNN with P, R and F of 38.7%, 30.4% and 34.0%. We make the comparison between rotation and horizontal region proposal and some detection results are illustrated in Figure 8. The rotation based approach is able to have accurate detection with less background area, which indicates the effectiveness of incorporation the rotation strategy.

Further analysis on the baseline results give us the following insights: (i) the hard situations (e.g., blur and uneven lighting) in the image can hardly be detected; (ii) some text instances in extremely small size cannot properly be detected, and this issue result in a large recall loss of the performance; (iii) the extremely long text line, i.e., height-width ratio of the bounding box larger than 1:10, cannot be correctly detected, often split in several shorter proposals, so that all the proposals become false detection according to the evaluation of MSRA-TD500 (some failed detection are shown in Figure 9). A few alternative strategies and settings from the baseline approach are tested and the summary is listed in Table I.

Context of text region. Incorporating the contextual information is proved to be useful for the general object detection task (e.g., [45]) and we wonder whether it can promote a text detection system. We keep the center of the rotated bounding box and orientation, and enlarge both width and height by a

Approach	P	R	F	Time
This work	82.1%	67.7%	74.2%	0.3s
Yao <i>et al.</i> [11]	76.5%	75.3%	75.9%	0.62s
Zhang <i>et al.</i> [10]	83%	67%	74%	2.1s
Yin <i>et al.</i> [42]	81%	63%	71%	1.4s
Kang <i>et al.</i> [43]	71%	62%	66%	-
Yin <i>et al.</i> [44]	71%	61%	65%	0.8s
Yao <i>et al.</i> [40]	64%	62%	61%	7.2s

TABLE III: Comparison with state-of-the-art on MSRA-TD500.



Fig. 10: Text detection results on MSRA-TD500.

factor of 1.X on the data preprocess step. During testing phase, we divide the enlargement for every proposal. As shown in Table II, all the experiments are with obvious increase in F-measure. The reason may probably be that as the bounding box become larger, more context information of text instance is grasped and the information of orientation can be better captured, so that the orientation of the proposals can be more precisely predicted.

Training dataset enlargement. we adopt HUST-TR400 [40] as an additional dataset and form the training set of 700 images from both datasets. There is a significant improvement of all the measurements and the F-measure is 60.8%, showing that the network is better trained and more robust when handling noisy inputs.

Border padding. With our filtering strategy, most of the breaking bound R-anchors are eliminated. However, as the bounding box is rotated with angles, it may still exceed the image border, especially when we enlarge the text region for the contextual information. Thus we set a border padding of 0.25 times of each side to reserve more positive proposals. Experiment shows that adding border padding to the image improves the detection results. Besides, combining border padding with enlargement of text region and training dataset makes a further improvement with F-measure of 63.3%.

Scale jittering. There are still a number of small text regions in both training dataset, and we add the scale jittering for the robustness of our system. The image patch is re-scaled with long side of a random size less than 1300 pixels and then sent into the network. The F-measure based on scale jittering is 69.3%, which has 6% improvement over the experiment without jittering.

Post processing. The annotation of MSRA-TD500 prefers to label the region of a whole text line. Thus, the length of text line is not with a fixed range, sometimes to be very long. However, the ratios for the R-anchor are fixed and may not be large enough to cover all the length, which leads to several short bounding box results for a single text region. To handle this extremely long text line issue, a post-processing step is incorporated by combine multiple short detection boxes in the neighborhood area to one proposal. With the post-processing, the performance further boost to F-measure of 74.2%.

Comparisons with state-of-the-art. The experimental results of our method compared with related works are given in Table III. The performance of our RRPN reaches the same magnitude of start-of-the-art approaches, such as [11] and [10]. As our system make the end-to-end text detection, it is more efficient than others with only 0.3s processing time per testing image. Some detection results on MSRA-TD500 are illustrated in Figure 10.

B. ICDAR2015

We conduct experiments on ICDAR2015 text localization task. Table IV summarizes the results of our framework comparing with state-of-the-art approaches. Following the strategies and parameters on MSRA-TD500 dataset, we train the baseline system with the training dataset of ICDAR2015 containing 1000 images with 10886 text instances. The evaluation result is precision 45.42%, recall 72.56%, and F-measure 55.87%.

We observe that precision of our approach is lower than other methods that reach the same magnitude of F-measure. We analysis the detected text regions and find three reasons. First, the ICDAR2015 training dataset may not contain enough rotated text region to train a robust system for the challenging ICDAR2015 benchmark. Second, even we resize the images using the scale jittering strategy, some of the incidental text regions are still too small. Finally, there exist some small unreadable text instances (labeled with '####') in ICDAR2015 training set, which may lead to some false detection results of text-like instance; some are shown in Figure 11. To further improve our detection system, we incorporate a few text datasets for training, i.e., SVT [4], ICDAR2003 [41] and ICDAR2013 [15]. The whole training dataset includes 2077 images with 14902 text instances. Besides, we make a larger scale by jittering the image patch with long side in a random size less than 1700 pixels and then sent into the network. By considering all of the instances in ICDAR2015, our method shows a huge performance gain with precision 68.53%, recall 72.65%, and F-measure 70.53%.

We also check the impact of small unreadable text instances by randomly removing these instances from training set. Figure 12 displays the curves of the measurements. The recall rate remains the same level around 72%-73% unless we remove all the unreadable instances, while precision significantly increases with the proportion. We finally get Precision of 82.17%, Recall of 73.23%, and F-measure of 77.44% when removing 80% unreadable text instances; some of the detection results are shown in Figure 13. As listed in Table



Fig. 11: Text detection in ICDAR2015 with the model trained on ICDAR2015 training set (including all unreadable text instances). The green areas are the positive detections with $\text{IoU} > 0.5$, red areas represent false detection.

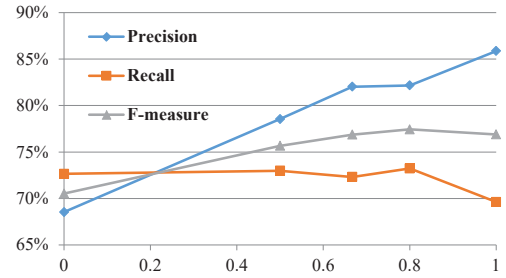


Fig. 12: The effect of unreadable text instance proportion on F-measure. Horizontal axis represents the removing proportion of unreadable text instances and vertical axis represents the percentage of F-measure.

IV, the substantial performance gains over the state-of-the-arts confirm the effectiveness of using rotation region proposal and rotation RoI for text detection task.

C. ICDAR2013

In order to examine the adaptability of our approach, we also conduct experiments on ICDAR2013 dataset, which is a horizontal-based benchmark. The model trained in last subsection for ICDAR2015 is used and the result is Precision 90.22%, Recall 71.89%, and F-measure 80.02%. As shown in Table V, there are 7% improvement comparing with Faster-RCNN. The approach DeepText [37] is also detection based approach, but is based on the Inception structure. We believe that our rotation-based framework is also complementary to Inception since they focus on different levels of information.

V. CONCLUSION

In this paper, we have introduced a rotation based detection framework for arbitrary-oriented text detection. Inclined rectangle proposals are generated with the text region orientation angle information from higher convolutional layer of network, which is able to detect text in multi-orientation. A novel RRoi pooling layer is also designed and adapted to the rotated RoIs. Experimental comparisons with state-of-the-arts on MSRA-TD500, ICDAR2013 and ICDAR2015 show the effectiveness and efficiency of proposed RRPN and RRoi on text detection task.

Approach	Precision	Recall	F-measure
CTPN [34]	93%	83%	88%
DeepText [37]	85%	81%	85%
Yao <i>et al.</i> [11]	89%	80%	84%
RRPN	90%	72%	80%
Faster-RCNN [32]	75%	71%	73%

TABLE V: Comparison with state-of-the-art on ICDAR2013. The results of Faster-RCNN are reported in [37].

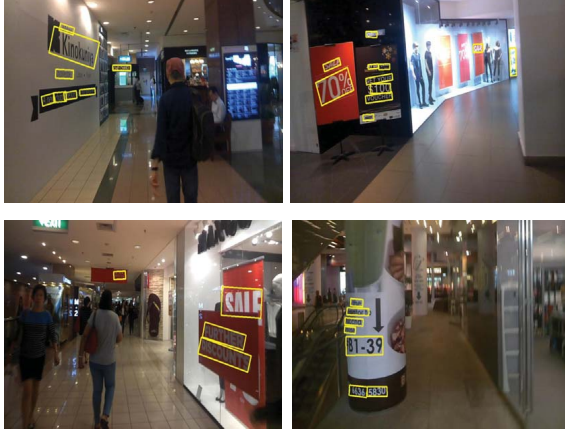


Fig. 13: Text detection results on ICDAR2015.

Approach	Precision	Recall	F-measure
RRPN	82.17%	73.23%	77.44%
Baidu IDL v2	77.41%	72.75%	75.01%
HUST_OrientedText	69.03%	74.15%	71.49%
SCUT_DMPNet	73.23%	68.22%	70.63%
Yao <i>et al.</i> [11]	72.26%	58.69%	64.77%
CTPN [34]	74.22%	51.56%	60.85%
MCLAB_FCN [10]	70.81%	43.09%	53.58%
Stradvision-2 [16]	77.46%	36.74%	49.84%

TABLE IV: Comparison with state-of-the-art on ICDAR2015

REFERENCES

- [1] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. 366–373.
- [2] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image & Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.
- [3] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2963–2970.
- [4] K. Wang and S. Belongie, "Word spotting in the wild," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 591–604.
- [5] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Asian Conference on Computer Vision*, 2010, pp. 770–783.
- [6] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "Photoocr: Reading text in uncontrolled conditions," in *International Conference on Computer Vision (ICCV)*, 2013, pp. 785–792.
- [7] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 512–528.
- [8] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced msr trees," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 497–511.
- [9] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Text flow: A unified text detection system in natural scene images," in *International Conference on Computer Vision (ICCV)*, 2015.
- [10] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [11] 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.
- [12] T. He, W. Huang, Y. Qiao, and J. Yao, "Accurate text localization in natural image with cascaded convolutional text network," *arXiv preprint arXiv:1603.09423*, 2016.
- [13] R. Girshick, "Fast r-cnn," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
- [14] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1083–1090.
- [15] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "Icdar 2013 robust reading competition," in *International Conference on Document Analysis and Recognition*, 2013.
- [16] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *International Conference on Document Analysis and Recognition*, 2015, pp. 1156–1160.
- [17] D. Chen and J. Luetttin, "A survey of text detection and recognition in images and videos," IDIAP, Tech. Rep., 2000.
- [18] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: a survey," *Pattern Recognition*, vol. 37, no. 5, 2004.
- [19] S. Uchida, "Text localization and recognition in images and video," in *Handbook of Document Image Processing and Recognition*, 2014, pp. 843–883.
- [20] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1480–1500, 2015.
- [21] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1631–1639, 2003.
- [22] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2004, pp. II–366.
- [23] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 97–104.
- [24] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *International Conference on Pattern Recognition*, 2012, pp. 3304–3308.
- [25] D. Karatzas, P. P. Roy, and L. D., "Icdar 2011 robust reading competition," 2011.
- [26] S. Zhang, M. Lin, T. Chen, L. Jin, and L. Lin, "Character proposal network for robust text extraction," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2633–2637.
- [27] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8027–8048, 2014.
- [28] H. Cho, M. Sung, and B. Jun, "Canny text detector: Fast and robust scene text localization algorithm," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [29] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-attentional convolutional neural network for scene text detection," *IEEE Transaction on Image Processing*, vol. 25, no. 6, pp. 2529–2541, 2016.
- [30] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 391–405.
- [31] J. R. R. Uijlings, K. E. A. V. D. Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.
- [32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2016.
- [33] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2014.
- [34] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 56–72.
- [35] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

- [36] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” in *Neural Information Processing Systems (NIPS)*, 2015, pp. 2017–2025.
- [37] Z. Zhong, L. Jin, S. Zhang, and Z. Feng, “Deeptext: A unified framework for text proposal generation and text detection in natural images,” *arXiv preprint arXiv:1605.07314*, 2016.
- [38] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [39] D. A. Plaisted and J. Hong, “A heuristic triangulation algorithm,” *Journal of Algorithms*, vol. 8, no. 3, pp. 405–437, 1987.
- [40] C. Yao, X. Bai, and W. Liu, “A unified framework for multioriented text detection and recognition,” *IEEE Transaction on Image Processing*, vol. 23, no. 11, pp. 4737–49, 2014.
- [41] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “ICDAR 2003 robust reading competitions,” in *International Conference on Document Analysis and Recognition*, 2003.
- [42] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao, “Multi-orientation scene text detection with adaptive clustering,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, 2015.
- [43] L. Kang, Y. Li, and D. Doermann, “Orientation robust text line detection in natural images,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 4034–4041.
- [44] X. C. Yin, X. Yin, K. Huang, and H. W. Hao, “Robust text detection in natural scene images,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970–83, 2014.
- [45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.