Proceedings of Machine Learning Research 101:631–645, 2019 ACML 2019

An Anchor-Free Oriented Text Detector with Connectionist Text Proposal Network

Chenhui Huang HCH373933165@GMAIL.COM

Jinhua Xu JHXU@CS.ECNU.EDU.CN

School of Computer Science and Technology
East China Normal University, Shanghai 200062, China

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

Deep learning approaches have made great progress for the scene text detection in recent years. However, there are still some difficulties such as the text orientation and varying aspect ratios. In this paper, we address these issues by treating a text instance as a sequence of fine-scale proposals. The vertical distances from a text pixel to the text borders are directly regressed without the commonly used anchor mechanism, and then the small local proposals are connected during the post-processing. A U-shape convolutional neural network (CNN) architecture is used to incorporate the context information and detect small text instances. In experiments, the proposed approach, referred to as Anchor-Free oriented text detector with Connectionist Text Proposal Network (AFCTPN), achieves better or comparable performance with less time consumption on benchmark datasets.

Keywords: Text Detection, Convolutional Neural Network, Deep learning

1. Introduction

Scene text detection is an important task in computer vision. Driven by its numerous potential applications like product identification and autonomous driving, scene text detection has obtained great development these years. However, it is still challenging to detect the scene text coming with complex scenarios, diverse shapes and various aspect ratios.

Scene text can be detected at three levels, including character, word and line level. Most traditional text detectors work at the character level, for example, Maximally Stable Extremal Regions (MSER) based methods (Neumann and Matas (2011)) assumed chromatic consistency within each character and Stroke Width Transform (SWT) based methods (Epshtein et al. (2010)) assumed consistent stroke width within each character. Deep learning approaches usually work at the word level or the line level. Different from characters which have constant aspect ratios, words or lines may have varying aspect ratios. A sentence in Latin languages consisting of a sequence of words separated by spaces can be detected at the word level. However, a sentence in many non-Latin languages, such as Chinese, Japanese and Korean, consists of a long sequence of characters and no obvious mark can be found to split neighboring characters. Compared to the character-level detection, it seems easier and more efficient to locate the whole sentence for non-Latin languages. Unfortunately, the text lines usually come with extreme aspect ratios, thus common methods

designed for the word-level detection fail on this task due to the anchor mechanism used and the limitation of receptive field.

Some approaches have been proposed to solve the aspect ratio problem. In TextBoxes (Liao et al. (2017)), anchors with large aspect ratios and one-dimensional horizontal convolutional kernels were used. The CTPN (Tian et al. (2016)) detected a text line in a sequence of fine-scale text proposals in convolutional feature maps. A vertical anchor mechanism was proposed that jointly predicted location and text/non-text score of each fixed-width proposal. The sequential proposals were then connected by a recurrent neural network to explore the context information.

Another issue for the scene text detection is the text orientation. Scene texts may have arbitrary orientations and even irregular shapes, i.e. curved texts. Some work has been done for the oriented scene text detection (Liao et al. (2018); Ma et al. (2018); Dai et al. (2018); Deng et al. (2018); Lyu et al. (2018b)). In the paper (Liao et al. (2018)), rotation-sensitive features were extracted by actively rotating the convolutional filters for regression. RRPN (Ma et al. (2018)) followed the standard Faster R-CNN framework (Ren et al. (2015)) and replaced the standard axis-aligned rectangles with rotation region proposals for RPN. Instance-aware semantic segmentation methods were used by Dai et al. (2018); Deng et al. (2018); Lyu et al. (2018b), which leveraged segmentation maps for generating arbitrary-shaped text mask.

In this paper, we propose a fast and easy-training network named Anchor-Free oriented text detector with Connectionist Text Proposal Network (AFCTPN) for the oriented scene text detection. As indicated by its name, AFCTPN follows the idea of CTPN to treat a text instance as a sequence of fine-scale proposals, which makes it able to detect texts with varying aspect ratios. We remove the vertical anchors in CTPN and directly predict the vertical distances from a text pixel to the top and bottom borders. To be more accurate with smaller text instances, we use a U-shape network to explore the context information and improve the resolution of the final feature map. By using the higher resolution map and smaller proposals, our method can detect small texts and long text lines with arbitrary orientation. Meanwhile, the efficiency is improved by removing the anchors and RNN layer in CTPN.

In summary, our contributions in this paper are three-fold. First, we directly regress a sequence of fine-scale proposals, which are not affected by the aspect ratio and can be used for the long oriented text detection. Second, we improve the detection speed by removing the anchor mechanism and simplifying the pipeline. Finally, our AFCTPN achieves better or comparable F-measure with much higher detection speed on two benchmark datasets.

2. Related Works

Over the past few years, the scene text detection has been a hot research area due to the great potential on real-world applications. Numerous inspiring ideas and effective methods based on deep learning have been proposed to distinguish text instances from complex natural scenes. These approaches can be roughly divided into two categories: regression-based methods and segmentation-based methods.

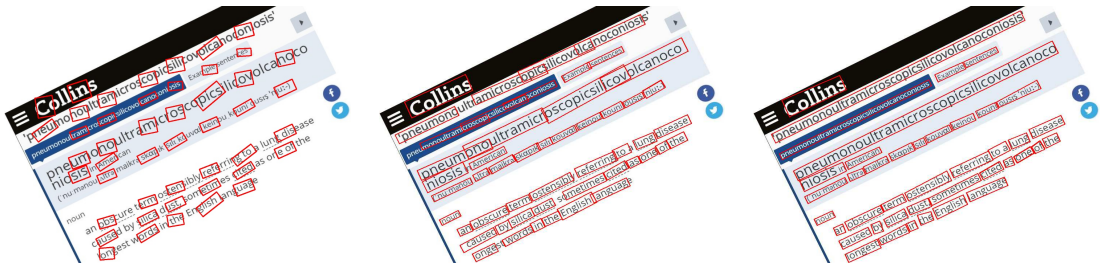
Regression-based methods usually follow the standard object detection frameworks, such as Faster R-CNN and SSD (Liu et al. (2016)). TextBoxes (Liao et al. (2017)) modified the

anchor shape and kernel size of SSD to adapt to various aspect ratios of text regions. RRPN (Ma et al. (2018)) used rotation region proposals for the regional proposal network (RPN) of Faster R-CNN to detect the arbitrary oriented text. However, common regression-based methods usually rely on the custom anchor design and fail to detect the extremely long texts. In the paper (He et al. (2017c)), a deep direct regression method was proposed for the oriented text detection, in which the vertex coordinates of quadrilateral text boundaries were directly predicted without anchors (or proposals).

Inspired by the fully convolutional network (FCN) (Long et al. (2015)), many approaches treat the text detection as a semantic segmentation problem. They usually output a dense pixel-level score map indicting whether a pixel belongs to a text region, thus segmentation-based methods can be used for the oriented text detection with extreme aspect ratios. However, text instances in scene images usually lie very close to each other. In such cases, they are very difficult, and are sometimes even impossible to be separated via semantic segmentation. Instance-aware semantic segmentation method was used by Dai et al. (2018), which leveraged the merits from accurate region proposal based methods and flexible segmentation based methods generating arbitrary-shaped text mask. PixelLink (Deng et al. (2018)) generated a connection map to divide pixels into different text instances. TextSnake (Long et al. (2018)) was proposed to represent the curved text region with ordered disks. PSENet (Wang et al. (2019)) predicted segmentation mask with different scales and introduced a progressive scale expansion algorithm to obtain the final detection result.

In the papers (Zhou et al. (2017); Lyu et al. (2018b)) the regression-based method and the segmentation based method were combined. EAST (Zhou et al. (2017) predicted a dense score map like segmentation-based methods and a geometry map to regress the shapes of text instances. The final text region was retrieved from the post-processing on the two maps. In the paper (Lyu et al. (2018b)), 4 corner points and 4 position sensitive segmentation maps were predicted. In inference stage, candidate boxes were generated by sampling and grouping corner points, which were further scored by segmentation maps and suppressed by non-maximum suppression (NMS).

Our approach is mostly related to EAST (Zhou et al. (2017)) and CTPN (Tian et al. (2016)). Similar to EAST, we use direct regression without anchors. But different from EAST, we predict a sequence of fine-scale proposals and only predict the vertical distances to the top and bottom borders, since it is harder to predict the horizontal distances precisely due to the text orientation and aspect ratio. That is why EAST may fail to detect the long text. Our method is inspired by CTPN. Considering the fact that text instances seldom overlap each other, we eliminate the anchor mechanism in CTPN. And we also remove the RNN layer in CTPN, which is used to encode the text-line context information. We argue that the horizontal RNN enrolls too much background while detecting oriented text and it is time-consuming and poses a challenge to train the whole network. A U-shape network architecture is used in our model and the context information can be obtained from top layer features. Our approach has outperformed the EAST and CTPN by a large margin on ICDAR2015 dataset in both accuracy and speed. Comparison of the detection results of an image is shown in Fig. 1. It can be seen that CTPN fails for the oriented text detection, and EAST performs poorly for the long text lines.



(a) (b) (c)

Figure 1: Detection results of CTPN (a) , EAST (b) and our method (c), respectively.

3. Methodology

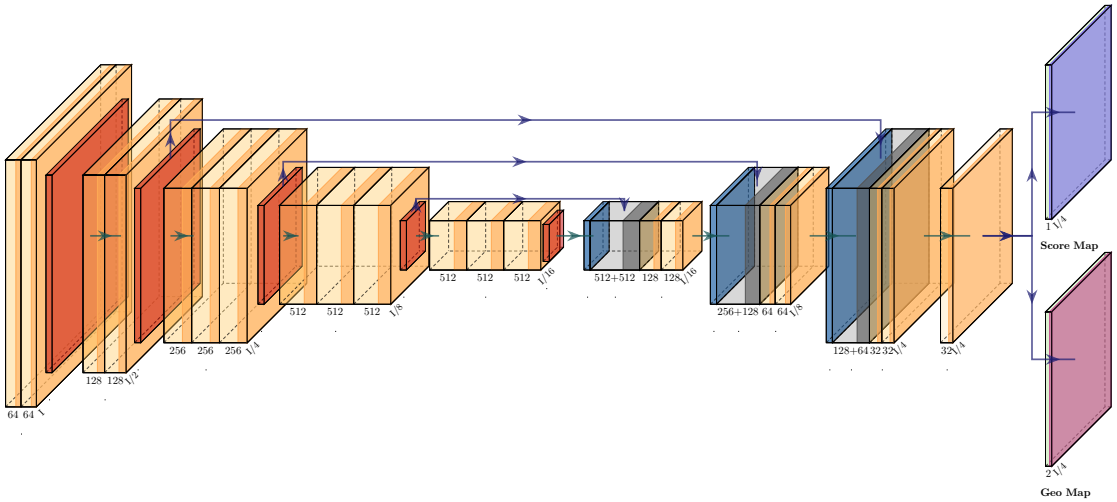
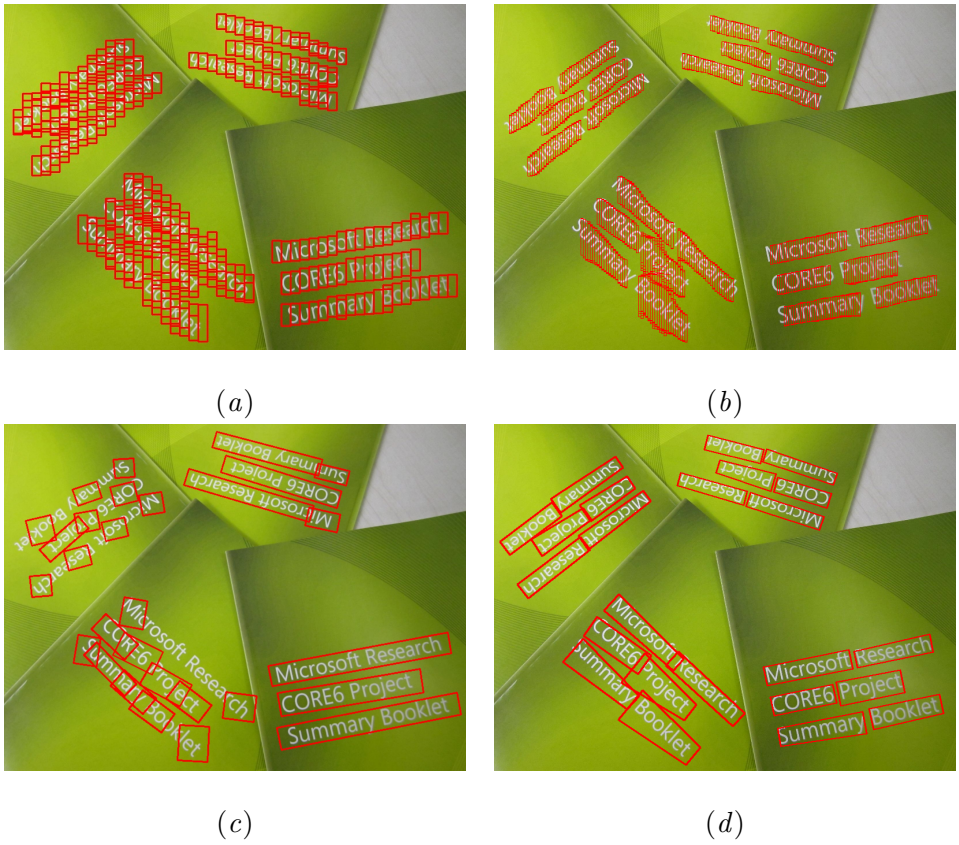


Figure 2: Network Architecture based on VGG16. The maps after convolution, pooling and upsampling are represented by yellow, red and blue, respectively.

3.1. Network Architecture

The network architecture is shown in Fig. 2. There are five blocks in the backbone network. We use VGG16 (Liu and Deng (2015)) and ResNet50 (He et al. (2016)) in our experiments as the backbone network.

Instead of directly conducting the output layer after the backbone as CTPN, we merge feature maps gradually like U-net (Ronneberger et al. (2015)) and finally get a feature map of 1/4 size of the original input image. Since the final feature map contains both texture details from lower layers and semantic information from higher layers, our network is able to predict the dense text score map and geometry map more precisely. As shown in Fig. 3, the predicted proposals of our method are more smooth, which is the key point to detect oriented text.



(a) (b)

(c) (d)

Figure 3: The predicted proposals (top) and detection results (bottom) of CTPN (left) and our method (right).

We also remove the anchor mechanism, considering the fact that text instances seldom overlap each other. As a result, the output just consists of a text score map and 2 geometry maps that indicate distances to top and bottom borders of corresponding quadrilateral bounding box. Compared to the original CTPN with k score maps and $2k$ regression targets (k is the number of anchors, set to 10 in the paper), our approach is much more efficient with only 1 score map and 2 regression channels.

3.2. Label Generation

In order to distinguish words from each other and decrease the noise brought by the less accurate ground truth, we adopt the idea in EAST to shrink the ground truth with a coefficient of 0.3. Only the pixels inside the shrunk region of the ground truth are considered positive. For each positive location, two vertical distances are calculated at the scale of the original input image as regression targets.

3.3. Loss Function

The loss function can be formulated as

$$L = L_{cls} + \lambda_g L_{reg} \tag{1}$$

where L_{cls} and L_{reg} represent the classification loss for the text/non-text score map and the regression loss for vertical coordinates regression, respectively. λ_g is the loss weight to balance different tasks and is set to 1.0 in our experiments.

Loss for Score Map It is common to use balanced sampling and hard negative mining in anchor-based approaches. To facilitate a simpler training procedure, we eliminate anchor mechanism and transfer the text classification into a semantic segmentation problem. Inspired by Milletari et al. (2016), we choose the Dice loss which is based on Dice coefficient. The Dice coefficient between the predicted score map and ground truth can be written as

$$D(P, G) = \frac{2\sum_{x,y}(P_{x,y} \cdot G_{x,y})}{\sum_{x,y} P_{x,y} + \sum_{x,y} G_{x,y}} \tag{2}$$

where $P_{x,y}$ and $G_{x,y}$ refer to the text score of prediction and ground truth at each location (x, y), respectively. Using this formulation, we do not need to add extra processing on training images to establish the right balance between foreground and background voxels. L_{cls} is the loss for text/non-text classification, which can be formulated as

$$L_{cls} = 1 - D(P, G) \tag{3}$$

Loss for Regression Since the scales of scene text usually vary tremendously, it is not proper to simply use L1 or L2 loss for regression. Aiming to detect both large and small text instances, we adopt IoU loss proposed in Yu et al. (2016), which is robust to the variation in scales of object.

$$L_{reg} = \frac{1}{|\Omega|}\sum_{x\in\Omega} -\log IoU(\hat{R}_x, R^*_x) = \frac{1}{|\Omega|}\sum_{x\in\Omega} -\log \frac{|\hat{R}_x \cap R^*_x|}{|\hat{R}_x \cup R^*_x|} \tag{4}$$

Note that Ω is the set of valid points. $\hat{R}_x$ and R^*_x represent the predicted geometry of fine-scale proposal and its corresponding ground truth, respectively. As the width of fine-scale proposals is fixed, the intersection can be formulated as

$$|\hat{R}_x \cap R^*_x| = \min(\hat{d}_1, d^*_1) + \min(\hat{d}_2, d^*_2) \tag{5}$$

where d_1 and d_2 represent the distances of a pixel from top and bottom borders of its corresponding text instance, respectively. The union is given by

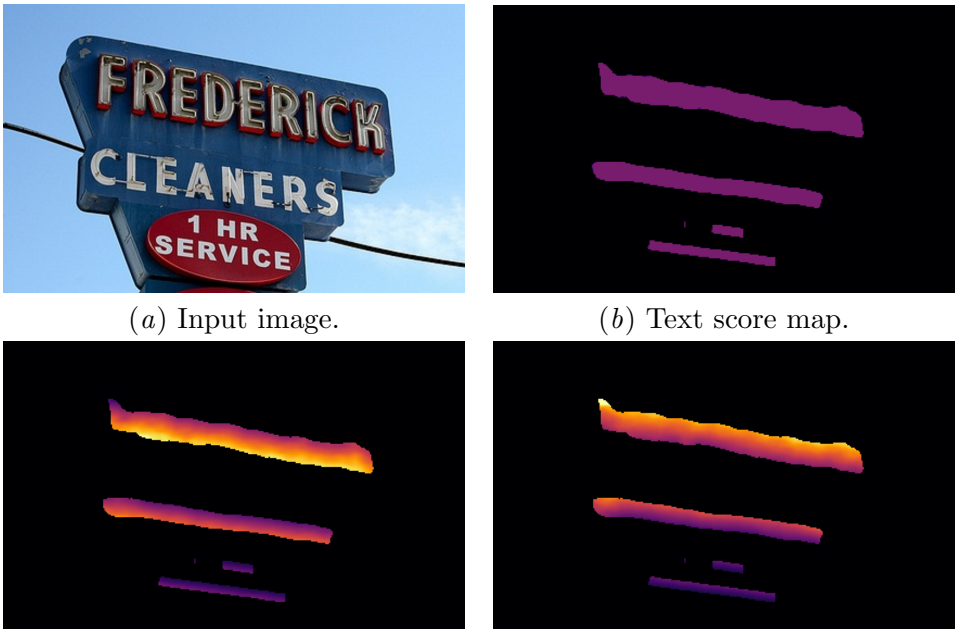
$$|\hat{R}_x \cup R^*_x| = |\hat{R}_x| + |R^*_x| - |\hat{R}_x \cap R^*_x| \tag{6}$$

$$|R_x| = d_1 + d_2 \tag{7}$$

Therefore, L_{reg} can be easily calculated by the two output distances.

3.4. Post-processing

After the feedforward, the network outputs a score map and a geometry map. We use the score map as a mask and only the pixels with a text score no less than the threshold (set to 0.99 in our experiments) are considered valid. Then the fine-scale proposals are restored from valid points on the geometry map, followed by a non-maximum suppression (NMS). Finally, we connect the proposals to text regions. Since we have adopted a shrinked region as ground truth, offsets are added to both ends of a segmentation mask. In our experiments, we set the offset to $0.3H$, where H is the average height of the corresponding text instance. The whole process is shown in Fig. 4.



(a) Input image. (b) Text score map.

(c) Geometry map, indicates the distances to top and bottom borders of corresponding text, respectively.

(d) Fine-scale proposals. (e) Result.

Figure 4: Steps of post-processing.

4. Experiments

4.1. Datasets

We test our model on the following datasets.

ICDAR2015 The ICDAR2015 dataset (Karatzas et al. (2015)) is from challenge 4 of the 2015 Robust Reading Competition. It includes 1000 training images and 500 testing images. Scene text images in this dataset are taken by Google Glasses in an incident way. The text instances from this dataset are labeled as word level quadrangles.

ICDAR2013 The ICDAR2013 dataset (Karatzas et al. (2013)) is from challenge 2 of the 2013 Robust Reading Competition. It contains 229 training images and 233 testing images. This dataset aims at focused scene text detection, which is the typical scenario for text reading and text translating applications.

4.2. Data Augmentation

Data augmentation is important to keep a model robust when the training data is quite limited. We resize the height in range $[0.8, 1.2]$ randomly and then the image is rotated in range $[-10°, 10°]$. Finally, the image is cropped to the size of 512x512 and we adjust its brightness, contrast and saturation randomly.

4.3. Implement Details

Our method is implemented on PyTorch 1.0 (Paszke et al. (2017)). We adopt Adam optimizer (Kingma and Ba (2014)) as our learning rate scheme. For each training, the learning rate decays from 10^{-3} by 1/10 every 300 epochs and finally stops at 10^{-5}. All the experiments are conducted on a server with a NVIDIA Tesla K40 GPU and Google Cloud Platform with a NVIDIA Tesla P100 GPU.

4.4. Quantitative Results

ICDAR2015 To get a direct and fair comparison with the original CTPN, we choose VGG16 (Liu and Deng (2015)) as our backbone and train the model for 900 epochs only with the 1000 training images from ICDAR2015. As shown in Tab. 1, our AFCTPN achieves a F-measure of 0.792, which outperforms CTPN (0.608) by a large margin and surpasses both VGG16-based EAST (0.764) and PVANET2x-based EAST (0.782). Note that we evaluate the model with the original input resolution of 1280x720 and use no magic such as multi-scale testing.

The frames per second (FPS) of the CTPN, VGG16-based EAST and our AFCTPN is retrieved from the same environment with a NVIDIA TESLA P100 GPU and our approach achieves FPS of 9.4, which is faster than CTPN (7.1 FPS) and EAST (8.3 FPS). We also test our method with ResNet50 and it achieves FPS of 12.0 with a F-measure of 0.774.

Compared with other methods, our F-measure of 0.792 is not as good as PixelLink (0.837) and TextSnake (0.826), but our FPS of 9.4 is much higher than PixelLink (3.0) and TextSnake (1.1).

Table 1: Quantitative results on ICDAR2015. FPS of ours, CTPN and VGG16-based EAST is from our test with NVIDIA Tesla P100 for comparison, others are from the original papers.

Method	Precision	Recall	F-measure	FPS
Ours + VGG16	0.827	0.760	0.792	9.4
Ours + ResNet50	0.807	0.743	0.774	12.0
CTPN (Tian et al. (2016))	0.742	0.515	0.608	7.1
EAST + VGG16 (Zhou et al. (2017))	0.804	0.727	0.764	8.3
EAST + PVANET2x (Zhou et al. (2017))	0.836	0.735	0.782	13.2
Zhang et al. (2016)	0.708	0.431	0.535	0.476
Yao et al. (2012)	0.723	0.587	0.648	1.61
SegLink (Shi et al. (2017))	0.731	0.768	0.750	-
Liu et al. (2018b)	0.72	0.80	0.76	-
SSTD (He et al. (2017b))	0.802	0.739	0.769	7.7
WordSup (Hu et al. (2017))	0.793	0.770	0.782	-
He et al. (2017a)	0.820	0.800	0.810	1.1
TextSnake (Long et al. (2018))	0.849	0.804	0.826	1.1
PixelLink (Deng et al. (2018))	0.855	0.820	0.837	3.0

ICDAR2013 Since the ICDAR2013 dataset is too small for training a deep neural network, we choose the images that contains English or Chinese from ICDAR2017 MLT training set for this task. During the test, we resize the short side to 600 while keeping the aspect ratio unchanged, which is the same as CTPN. As shown in Tab. 2, our method achieves a F-measure of 0.883 with ResNet50, which surpasses CTPN (0.877), EAST (0.873) and other methods on this dataset.

Table 2: Quantitative results on ICDAR2013.

Method	Precision	Recall	F-measure
Ours + ResNet50	0.948	0.826	0.883
Ours + VGG16	0.875	0.793	0.832
EAST + PVANET2x (Zhou et al. (2017))	0.926	0.826	0.873
CTPN (Tian et al. (2016))	0.930	0.830	0.877
Zhang et al. (2016)	0.88	0.78	0.83
SynthText (Liu et al. (2018a))	0.920	0.755	0.830
Holistic (Yao et al. (2016))	0.889	0.802	0.843
PixelLink (Deng et al. (2018))	0.864	0.836	0.845
SegLink (Shi et al. (2017))	0.877	0.83	0.853
Lyu et al. (2018a)	0.933	0.749	0.858
He et al. (2017a)	0.92	0.80	0.86

4.5. Qualitative Results

As mentioned above, our approach is able to detect the extremely long text, which is quite challenging for word-level approaches such as EAST. Since the images from common benchmark datasets only contain short word-level text, we conduct some experiments with long or large text. All the input images are resized to 720x360 and both the backbones of EAST and our method are ResNet50. Results are shown in Fig. 5. It is obvious that the proposed approach performs better on the larger or longer text while as accurate as EAST on the smaller ones.

Fig. 6 lists some detection results of CTPN and our method on ICDAR2015 from ICDAR Robust Reading Competition website. It can be seen that the proposed approach performs better on the smaller and oriented text.

Figure 5: Qualitative results of EAST (left) and proposed method (right). Detection results of EAST are retrieved from http://east.zxytim.com/.

4.6. Ablation Study

Influence of the feature map resolution and proposal width. We study the effect of the feature map resolution and proposal width by modifying the number of upsampling layers. Resolutions of output feature maps are set to 1/16 (no upsampling), 1/8, 1/4, 1/2 of the input image and the corresponding widths of proposals are 16, 8, 4, 2, respectively. All

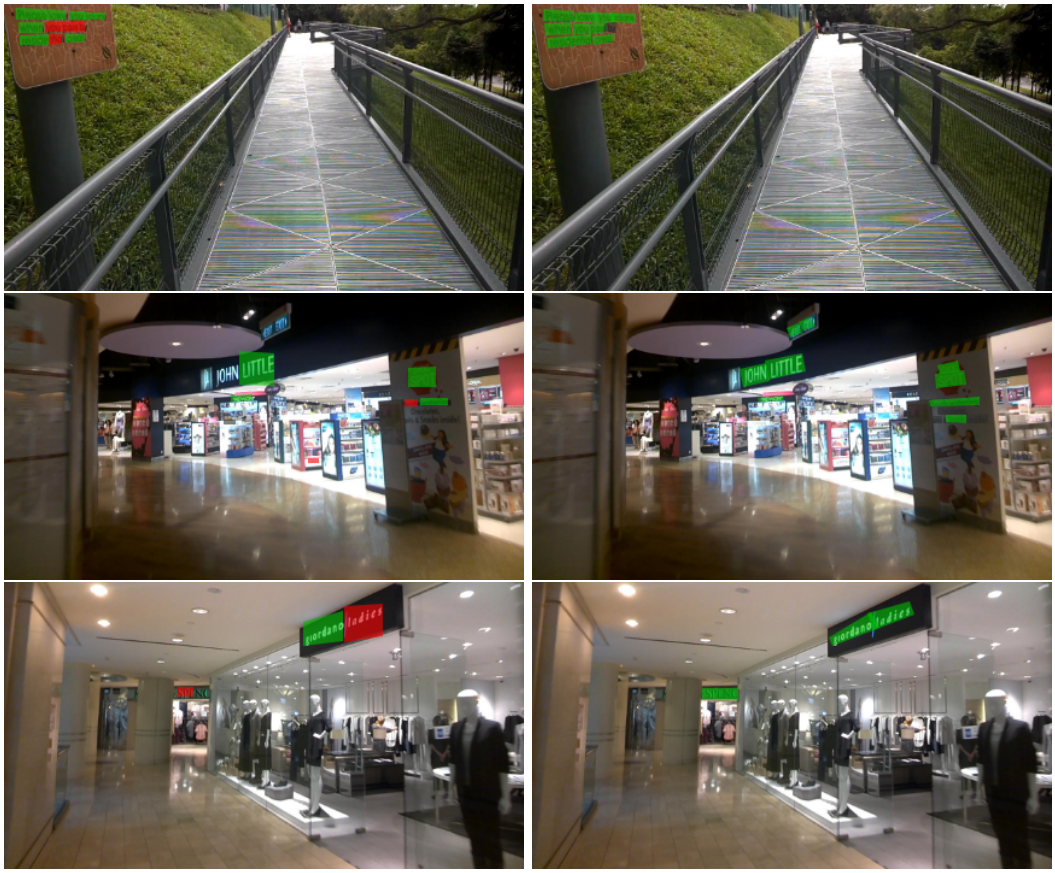


Figure 6: Qualitative results of CTPN (left) and the proposed method (right) on ICDAR2015 dataset. Detection results can be found on ICDAR Robust Reading Competition website. Green, red and gray regions stand for true positive, false positive and don't care, respectively.

the models are trained for 1000 epochs with VGG16 backbone and evaluated on ICDAR2015. Tab. 3 shows the experiment results, from which we can find that using higher resolution and smaller proposals results in better accuracy since it is difficult to separate the small words which are close to each other with lower resolution and larger proposals. Note that the model tends to split a word into characters if the proposals is too small, for example width of 2, which leads to worse performance. The visualization of different proposals is similar to Fig. 3 (corresponding to proposal width of 16 and resolution of 1/4).

5. Conclusion

In this paper, we have presented a fast yet accurate approach for arbitrary-length oriented text detection. By taking advantage of the locality and homogeneity of text, our method can overcome the limitation of receptive field while detecting the text instances with extreme aspect ratios. We have simplified the pipeline by removing the anchor mechanism and RNN

Table 3: Test results with different resolutions and proposals. '16s', '8s', '4s' and '2s' means the resolution of the feature map is 1/16, 1/8, 1/4 and 1/2 of the input image.

Method	Precision	Recall	F-measure
Ours-16s	0.582	0.432	0.496
Ours-8s	0.692	0.504	0.583
Ours-4s	0.827	0.760	0.792
Ours-2s	0.801	0.750	0.774

layer to increase the efficiency. Our method has achieved better or comparable results with less time consumption on the ICDAR2013 and ICDAR2015 benchmarks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Project 61175116.

References

Y. Dai, Z. Huang, Y. Gao, Y. Xu, K. Chen, J. Guo, and W. Qiu. Fused text segmentation networks for multi-oriented scene text detection. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3604–3609, Aug 2018. doi: 10.1109/ICPR.2018.8546066.

Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2963–2970, June 2010. doi: 10.1109/CVPR.2010.5540041.

Dafang He, Xiao Yang, Chen Liang, Zihan Zhou, Alexander G Ororbi, Daniel Kifer, and C Lee Giles. Multi-scale fcn with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3519–3528, 2017a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, and Xiaolin Li. Single shot text detector with regional attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3047–3055, 2017b.

Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Deep direct regression for multi-oriented scene text detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 745–753, 2017c.

Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, and Errui Ding. Wordsup: Exploiting word annotations for character based text detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4940–4949, 2017.

D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn, and L. P. de las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493, Aug 2013. doi: 10.1109/ICDAR.2013.221.

D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160, Aug 2015. doi: 10.1109/ICDAR.2015.7333942.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In Satinder P. Singh and Shaul Markovitch, editors, *AAAI*, pages 4161–4167. AAAI Press, 2017.

Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-Song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. pages 5909–5918, 06 2018. doi: 10.1109/CVPR.2018.00619.

S. Liu and W. Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, Nov 2015. doi: 10.1109/ACPR.2015.7486599.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46448-0.

Zichuan Liu, Yixing Li, Fengbo Ren, Wang Ling Goh, and Hao Yu. Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 7194–7201. AAAI press, 1 2018a.

Zichuan Liu, Guosheng Lin, Sheng Yang, Jiashi Feng, Weisi Lin, and Wang Goh. Learning markov clustering networks for scene text detection. pages 6936–6944, 06 2018b. doi: 10.1109/CVPR.2018.00725.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018.

Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018a.

Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. Multi-oriented scene text detection via corner localization and region segmentation. pages 7553–7563, 06 2018b. doi: 10.1109/CVPR.2018.00788.

J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11): 3111–3122, Nov 2018. doi: 10.1109/TMM.2018.2818020.

F. Milletari, N. Navab, and S. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, Oct 2016. doi: 10.1109/3DV.2016.79.

Lukas Neumann and Jiri Matas. A method for text localization and recognition in real-world images. In Ron Kimmel, Reinhard Klette, and Akihiro Sugimoto, editors, *Computer Vision – ACCV 2010*, pages 770–783, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19318-7.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2550–2558, 2017.

Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In Bastian Leibe, Jiri Matas, Nicu Sebe, and

Max Welling, editors, *Computer Vision – ECCV 2016*, pages 56–72, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46484-8.

Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2019.

C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1083–1090, June 2012. doi: 10.1109/CVPR.2012.6247787.

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.

Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 516–520, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3603-1. doi: 10.1145/2964284.2967274. URL http://doi.acm.org/10.1145/2964284.2967274.

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.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017.