Arbitrarily Oriented Scene Text Detection by Predicting Quadrilaterals

Anonymous CVPR submission

Paper ID 1366

Abstract

Treating texts as oriented long shaped objects, we propose a non-proposal based scene text detection method which can detect arbitrarily oriented texts by predicting quadrilaterals. Our detection framework is simple and effective with a fully convolutional network and one-step post processing. The fully convolutional network is optimized in an end-to-end way and has a bi-task output in which one is for pixel-wise classification between text and background, and the other is for regression to determine the vertex coordinates of quadrilaterals. The proposed method is particularly beneficial for localizing incidental scene texts, which are hard to identify the constituent characters. On the ICDAR2015 Incidental Scene Text benchmark, our method achieves the F1-measure of 81%, which is a new state-of-the-art and significantly outperforms previous approaches. On other standard datasets with focused scene texts, our method also performs competitively.

1. Introduction

Scene text detection has drawn great interests from both computer vision and machine learning communities because of its great value in practical uses and the technical challenges. Owing to the intensive research in the past years, especially the adoption of deep convolutional neural networks (CNNs), focused horizontal text detection has been greatly improved during the past two years [1] [2] [3]. However, most CNN-based approaches still assume horizontal or nearly horizontal text lines, and a general framework to detect multi-oriented texts is not well studied. The detection of incidental texts, which have abundant variations of blurring, scale, direction, aspect ration and font style, remains a bigger challenge.

Text detection methods can be divided into two groups according to the comprehension of what is text. Most traditional methods [4] [5] [6] [7] [8] regard text as a character composite [9] and as a result, those methods follow a "character to line" strategy. In other words, they first localize characters or components and then group them into a word or text line. It works well on clean images where characters or components are salient, but is insufficient for images with blurred or low-resolution texts. The other kind of methods simply treat text words or lines as objects [2] [10] [11] regardless of the composition attribute. This strategy takes advantage of the holistic textual properties of texts and their distinction from background, and performs well when trained with scene text data, especially when using deep neural network models such as convolutional neural network.

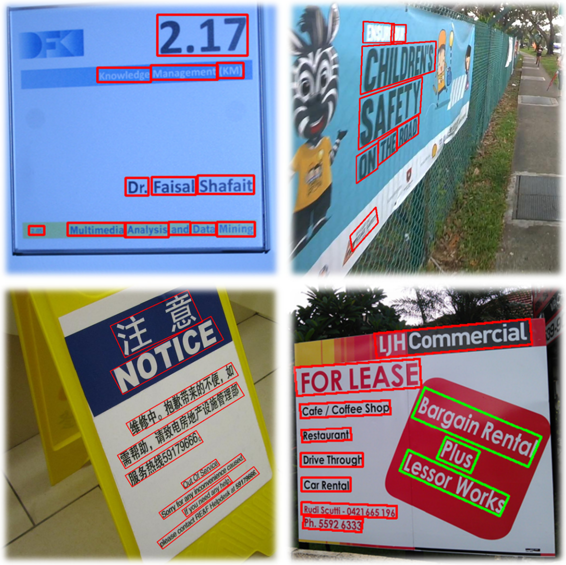


Figure 1. Natural scene texts in various of scale, orientation, perspective distortion and aspect ratio.

The dramatic improvements in generic object detection are mainly attributed to deep convolutional feature and end-to-end structure design. In proposal based methods, the work in [12] performs classification based on cropped proposals beforehand, while works like [13] [14] try to integrate proposal stage with the network without clipping images. After that, proposals are replaced by "anchors" devised in [15], and so, proposal-based object detection totally realizes global training and testing. Non-proposal based methods like [16] [17] [18] [19] [20] were also proposed, and some of them are similar with [15] in sense of adopting anchor mechanism. The detection of incidental text (as shown in Fig.1), however, is greatly different from generic object and the popular anchor mechanism cannot be directly implemented for long and heavily inclined text.

In this paper, we propose a novel architecture based on fully convolutional network (FCN) [21] for word level arbitrarily oriented text detection. Our non-proposal based detection framework is able to output arbitrarily quadrilaterals without utilizing the sequentiality or anchor mechanism. On the ICDAR2015 Incidental Scene Text benchmark, we obtain F1-measure of 81%, which is a new state-of-the-art and surpass the second placed method by a large margin. On other popular datasets of focused images, the proposed method also performs competitively.

The proposed method has several novelties and advantages. First, it generalizes CNN-based generic object detection method by regressing arbitrary quadrilaterals, so as to detect texts of arbitrary orientation. In our framework, all the parameters except the post processing module can be optimized in an end-to-end way. Moreover, when regarding a word as an object, our framework can perform word level detection without word partition procedure in post processing. For post processing, we also propose a recalled nonmaximum suppression (NMS), which is modified from NMS for object detection to partition the texts that are too close.

The rest of this paper is organized as follows: In Section 2 we give a brief review of detection for scene text and generic object, in Section 3 we introduce details of our proposed method, in Section 4 we present the results on benchmarks, rationality analysis as well as time consumption estimation and, in Section 5 we conclude this paper.

2. Related Work

Scene Text Detection. Most scene text detection methods [1] [3]-[8] treat text as a composite of characters, so they first localize character or components candidates and then group them into a word or text line. Even for multi-oriented text, methods like [22] [23] [24] [25] also follow the same strategy and the multi-oriented line grouping is accomplished by either rule based methods or more complex graphic model. However, for incidental text in the ICDAR2015 dataset, some blurred or low resolution characters in a word could not be well extracted, which hinders the performance of localization.

Recently, some text detection methods discard the text composition and take text words or lines as generic objects. The method in [10] makes use of the symmetric feature of text lines and tries to detect text line as a whole. Despite the novelty of this work, the feature it uses is not robust for cluttered images. The method in [2] adopts the framework for object detection in [18], but the post processing relies on the text sequentiality. The method in [11], based on Faster R-CNN [15], attempts to convert text detection into object detection and the performance on horizontal texts demonstrates its effectiveness. However, constrained by the anchor mechanism and wide variety of text shapes, as well as the Intersection-over-Union (IoU) criterion in training, the Faster R-CNN based multi-oriented text detection requires many modifications and sacrifice the efficiency.

Generic Object Detection. Most generic object detection frameworks are multi-task structure with a classifier for recognition and a regressor for localization. According to the distinction of regressor, we divide these methods into two groups. The first one is non-proposal based regression method [16] [18] [20] and the regressor is trained to directly predict bounding boxes of objects. The second one is proposal based regression method [13] [14] [15] [17] [19] [26] and regressor is trained to predict the offset from proposals to corresponding ground truths. The proposals here can be generated by either low-level features [27] [28] [29] or simple clustering [17], as well as anchor mechanism [15].

Although most of the recent state-of-the-art approaches are proposal based regression method, considering the wide variety of texts in scale, orientation, perspective distortion and aspect ratio, non-proposal based regression has the potential advantage of avoiding the difficulty of proposal generation in complex images.

3. Proposed Methodology

The proposed detection system is diagrammed in Fig.2. It consists of four major parts: the first three modules, namely convolutional feature extraction, multi-level feature fusion, multi-task learning, together constitute the network part, and the last post processing part performs recalled NMS, which is an extension of traditional NMS.

3.1. Network Architecture

The convolutional feature extraction part is designed so that the maximum receptive field is larger than the input image size S. Considering the text feature is not as complicated as that of generic objects, our network tends to employ less parameters than models designed for ImageNet. The feature fusion part refers to the design in [21] and here we combine convolutional features from four streams. However, for reducing computation, we only up-sample the fused feature to quarter size of the input image. The multi-task part has two branches: First, the classification task output $\mathcal{M}_{cls}$ is a $\frac{S}{4} \times \frac{S}{4}$ 2nd-order tensor and it can be approximated as down-sampled segmentation (indicating text/background) for input images. Elements in $\mathcal{M}_{cls}$ with higher score are likely to be text, otherwise background; Second, the localization task output $\mathcal{M}_{loc}$ is a $\frac{S}{4} \times \frac{S}{4} \times 8$ 3rd-order tensor. The channel size of $\mathcal{M}_{loc}$ indicates that we intend to output 8 coordinates, corresponding to the quadri-

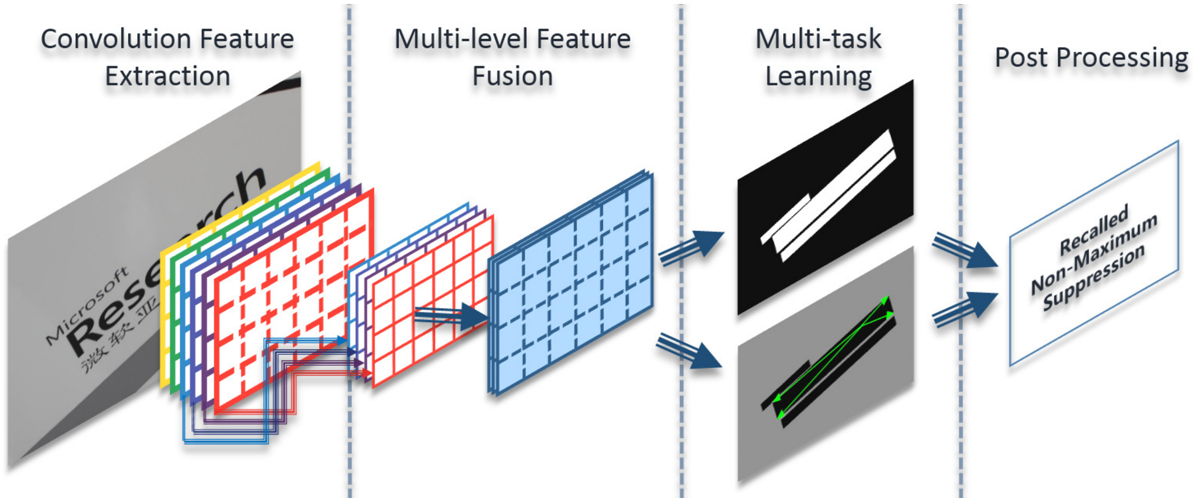

Convolution Feature Extraction
Multi-level Feature Fusion
Multi-task Learning
Post Processing
Recalled Non-Maximum Suppression

Figure 2. Overview of the proposed text detection method.

lateral vertexes of the text. The value at (w,h,c) in $\mathcal{M}_{loc}$ is denoted as $L_{(w,h,c)}$, which means the offset from coordinate of a quadrilateral vertex to that of the point at $(4w,4h)$ in input image, and therefore, the quadrilateral $\mathcal{B}(w,h)$ can be formulated as

$$\mathcal{B}(w,h)=\left\{L_{(w,h,c)}+(4w,4h)\,\middle|\,c\in\{1,2,\cdots,8\}\right\} \quad (1)$$

By combining outputs of multi-task, we predict a quadrilateral with score for each point of $\frac{S}{4}\times\frac{S}{4}$ map. More detailed structure and parameterized configuration of the network is shown in Fig.3. Note that in our approach, both classification and localization are conducted at $1/4$ resolution of original image.

3.2. Ground Truth and Loss Function

The full multi-task loss $\mathcal{L}$ can be represented as

$$\mathcal{L}=\mathcal{L}_{cls}+\lambda_{loc}\cdot\mathcal{L}_{loc}, \quad (2)$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{loc}$ represent loss for classifier and regressor respectively. The balance between two tasks is controlled by λ_{loc}.

Classification task. Although the ground truth for classification task can be deemed as a down-sampled segmentation between text and background, unlike the implementation in [1], we do not take all pixels of text region as positive, instead, we only regard pixels around the text center line within distance r as positive and enclose positive region with an ignored boundary as transition from positive to negative (shown in Fig.4). The parameter r is proportional to the short side of text with ratio 0.2.

Furthermore, text is taken as a positive sample only when its short side length ranges in $\left[32\times2^{-1},32\times2^{1}\right]$. If the short side length falls in $\left[32\times2^{-1.5},32\times2^{-1}\right)\cup\left(32\times2^{1},32\times2^{1.5}\right]$, we take the text as ignored, otherwise negative. Ground truths designed in this way reduce the confusion between text and non-text, which is beneficial for discriminative feature learning.

The loss function $\mathcal{L}_{cls}$ chosen for classification task is the hinge loss. Denote the ground truth for a given pixel as $y_i^*\in\{0,1\}$ and predicted value as $\hat{y}_i$, $\mathcal{L}_{cls}$ is formulated as

$$\mathcal{L}_{cls}=\frac{2}{S^2}\sum_{i=1}^{\frac{S^2}{4}}\max\left(0,\mathrm{sign}\left(0.5-y_i^*\right)\cdot\left(\hat{y}_i-y_i^*\right)\right)^2 \quad (3)$$

Besides this, we also adopt the class balancing and hard negative sample mining as introduced in [20] for better performance and faster convergence. Hence during training, the predicted values for ignored region and easily classified negative area are forced to zero.

Localization task. According to [14], the loss function $\mathcal{L}_{loc}$ used in localization task can be defined as follows. Denote the ground truth for a given pixel as z_i^* and predicted value as $\hat{z}_i$, $\mathcal{L}_{loc}$ is formulated as

$$\mathcal{L}_{loc}=\sum_{i}^{\frac{S^2}{4}}\left[y_i^*>0\right]\cdot\mathrm{smooth}_{L_1}\left(z_i^*-\hat{z}_i\right), \quad (4)$$

$$\mathrm{smooth}_{L_1}(x)=\begin{cases}0.5x^2 & \text{if } |x|<1,\\ |x|-0.5 & \text{otherwise.}\end{cases} \quad (5)$$

We choose smooth L_1 loss here because it is less sensitive to outliers compared with L_2 loss. During training stage, smooth L_1 loss need less careful tuning of learning rate and decreases steadily.

3.3. Recalled Non-Maximum Suppression

After we get the outputs produced by multi-task learning, each point with high score in classification task is related with a quadrilateral, however, there will be densely

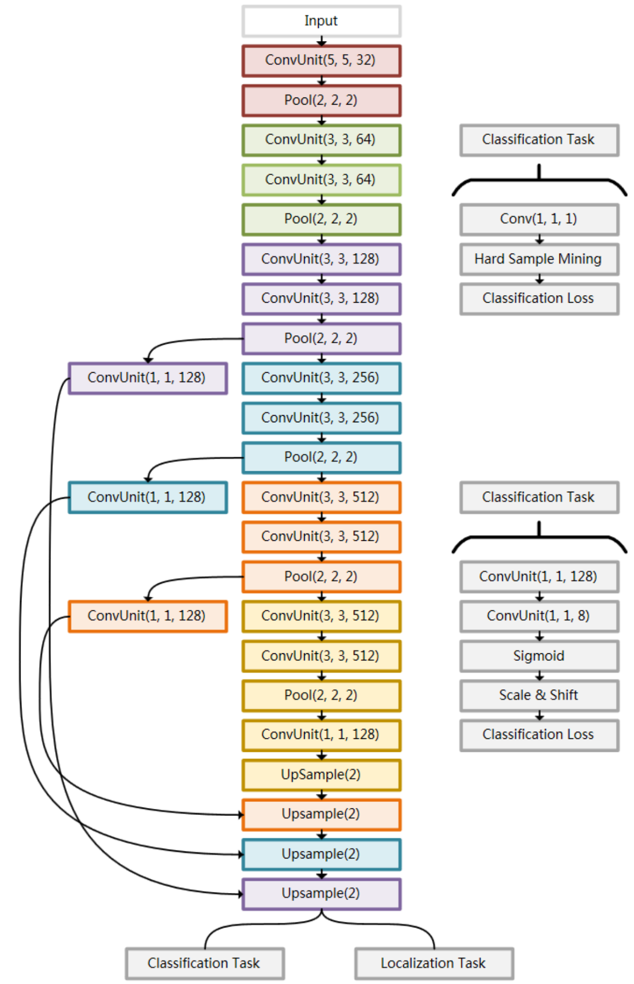
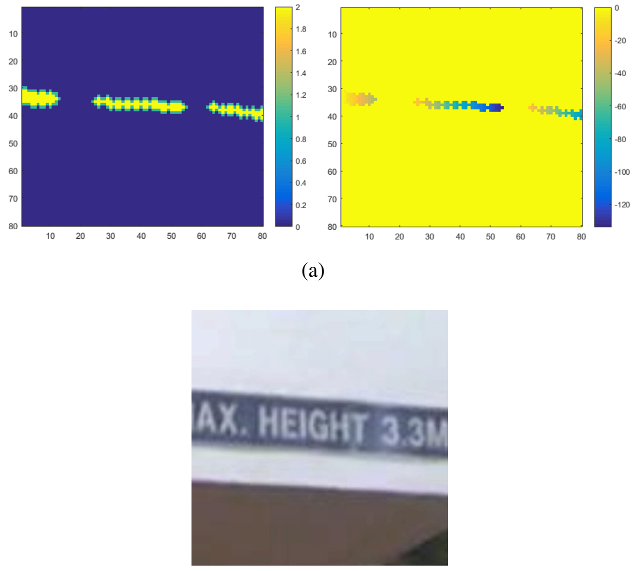


Figure 3. Structure of the network. Left: Detailed components of the convolutional feature extraction and multi-level feature fusion. The "ConvUnit(w, h, n)" represents a convolutional layer of n $w \times h$ kernels, connected by a batch normalization layer and a ReLU layer. Right: The design of multi-task module.

overlapped quadrilaterals for a word or text line. To reduce the redundant results we propose a method called recalled non-maximum suppression which is specialized for text detection. The recalled NMS mainly solves the problem that when texts are close, quadrilaterals between two words are often retained because of the difficulty in classifying pixels in word space.

The recalled NMS has three steps as shown in Fig.5. Firstly, we get suppressed quadrilaterals $\mathcal{B}_{sup}$ from densely overlapped quadrilaterals $\mathcal{B}$ by traditional NMS. Secondly, each quadrilateral in $\mathcal{B}_{sup}$ is switched to the one with highest score in $\mathcal{B}$ beyond a given overlap. After this step, quadrilaterals in word space are changed to quadrilaterals of higher score. Thirdly, after the second step we may get dense overlapped quadrilaterals again, and instead of suppression, we merge quadrilaterals in $\mathcal{B}_{sup}$ which are close to each other.



(a)

(b)

Figure 4. Visualized ground truth of multi-task. (a) The left map is the ground truth for classification task, where the yellow region are positive, enclosed by ignored region. The right map is the ground truth of top-left offset for localization task where values grow smaller from left to right within a word region. (b) The corresponding input image taken by the network.

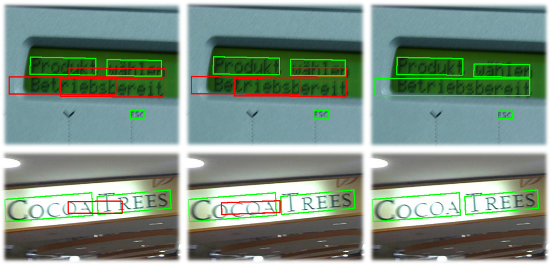


Figure 5. Three steps in recalled NMS. Left: results of traditional NMS (quadrilaterals in red are false detection). Middle: recalled high score quadrilaterals. Right: merging results by overlap ratio.

3.4. Network Implementation

The training samples of 320 × 320 are cropped from scaled images rotated randomly by 0, $\pi/2$, π, or $3\pi/2$. The task balance index λ_{loc} is raised from 0.01 to 0.5 after the classification task gets well trained. In testing, when images are larger than 320 × 320, we adopt a multi-scale sliding window strategy in which window size is 320 × 320, sliding stride is 160 and multi-scale set is $\{2^{-5}, 2^{-4}, \cdots, 2^{1}\}$. Pixels on $\mathcal{M}_{cls}$ are deemed as text if their values are higher than 0.7. In post processing, the only parameter, overlap ratio, in recalled NMS is 0.5.

4. Experiments

We evaluate our method on three benchmarks: ICDAR2015 Incidental Scene Text, MSRA-TD500 and ICDAR2013. The first two datasets have multi-oriented texts and the third one has mostly horizontal texts. For fair comparison we also list recent state-of-the-art methods on these benchmarks.

4.1. Benchmark Description

ICDAR2015 Incidental Scene Text. This dataset is recently published for ICDAR2015 Robust Reading Competition. It contains 1000 training images and 500 test images. Different from previous scene text datasets where texts are well captured in high resolution, this dataset contains texts with various scales, resolution, blurring, orientations and viewpoint. The annotation of bounding box (actually quadrilateral) also differs greatly from previous ones which has 8 coordinates of four corners in a clock-wise manner. In evaluation stage, word-level predictions are required.

MSRA-TD500. This dataset contains 300 training images and 200 test images, where there are many multi-oriented text lines. Texts in this dataset are stably captured with high resolution and are bi-lingual of both English and Chinese.

The annotations of MSRA-TD500 are at line level which casts great influence on training localization task. Lacking of line level annotation and sufficient bi-lingual training data, we did not use the training set and instead, utilize the generalization of our model trained on English word-level data.

ICDAR2013 Focused Scene Text. This dataset lays more emphasis on horizontal scene texts. It contains 229 training images and 233 test images which are well captured and clear. The evaluation protocol which allows grouping words into a text line is more flexible than that of ICDAR2015, but is more strict in Intersection-over-Union criterion.

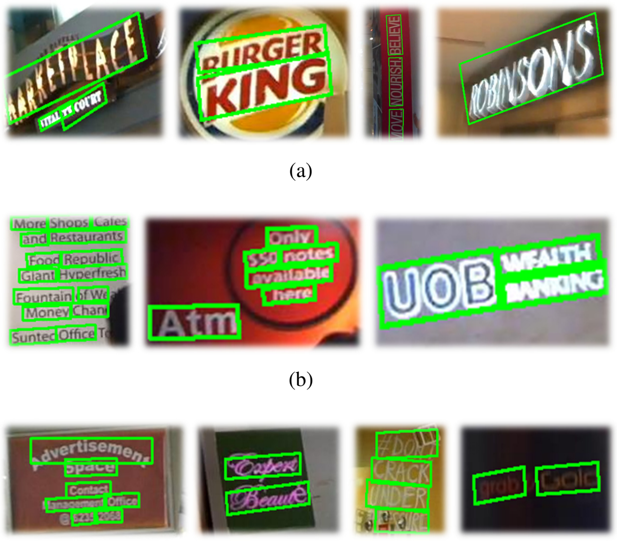
4.2. Implementation Details

The network is optimized by stochastic gradient descent (SGD) with back-propagation and the max iteration is 2×10^5. We adopt the "multistep" strategy in Caffe [30] to adjust learning rate. For the first 3×10^4 iterations the learning rate is fixed to be 10^{-2} and after that it is reduced to 10^{-3} until the 10^5th iteration. For the rest 10^5 iteration, the learning rate keeps 10^{-4}. Apart from adjusting learning rate, the hard sample ratio mentioned in Sec.3.2 is increased from 0.2 to 0.7 at the 3×10^4th iteration. Weight decay is 4×10^{-4} and momentum is 0.9. All layers except in localization task are initialized by "xavier" [31] and the rest layers are initialized to a constant value 0 for stable convergence.

The model is optimized on training datasets from ICDAR2013 and ICDAR2015, as well as 200 negative images (scene images without text) collected from the Internet. The whole experiments are conducted on Caffe and run on a workstation with 2.9GHz 12-core CPU, 256G RAM, GTX Titan X and Ubuntu 64-bit OS.

Table 1. Comparison of methods on ICDAR2015 dataset.

Algorithm	Precision	Recall	F-measure
Proposed	**0.85**	**0.77**	**0.81**
Tian *et al.* [3]	0.74	0.52	0.61
Zhang *et al.* [1]	0.71	0.43	0.54
StradVision2 [32]	0.77	0.37	0.50
StradVision1 [32]	0.53	0.46	0.50
NJU-Text [32]	0.70	0.36	0.47
AJOU [32]	0.47	0.47	0.47
HUST_MCLAB [32]	0.44	0.38	0.41



(a)

(b)

(c)

Figure 6. Detection examples of our model on ICDAR2015 Incidental Scene Text benchmark. (a) detection of multi-oriented texts. (b) automatic word partition within clustered texts. (c) detection for curved text, curlicue font, occlusion and low resolution.

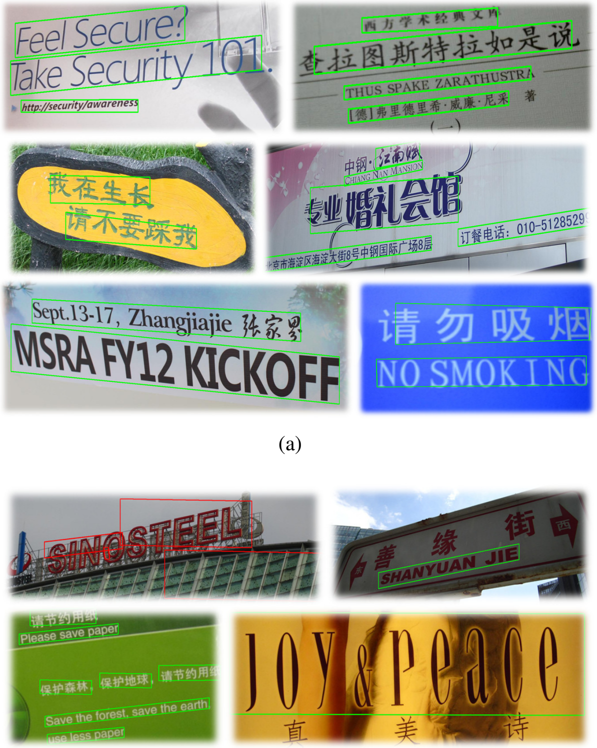
4.3. Experiment Results

ICDAR2015 Incidental Scene Text. The results shown in Tab.1 indicates the proposed method outperforms previous best one by a large margin in both precision and recall. To make a more meaningful comparison, we also list the top five F-measure of results in ICDAR2015 competition for extensive comparison. Some examples of our detection results are shown in Fig.6.

MSRA-TD500. The results of our method on this dataset are shown in Tab.2, with comparison with representative results of state-of-the art methods. It is shown that our method could reach the state-of-the-art performance even

Table 2. Comparison of methods on MSRA-TD500 dataset.

Algorithm	Precision	Recall	F-measure
Proposed	**0.77**	**0.70**	**0.74**
Zhang *et al.* [1]	0.83	0.67	0.74
Yin *et al.* [23]	0.81	0.63	0.71
Kang *et al.* [24]	0.71	0.62	0.66
Yin *et al.* [7]	0.71	0.61	0.65
Yao *et al.* [22]	0.63	0.63	0.60



(a)

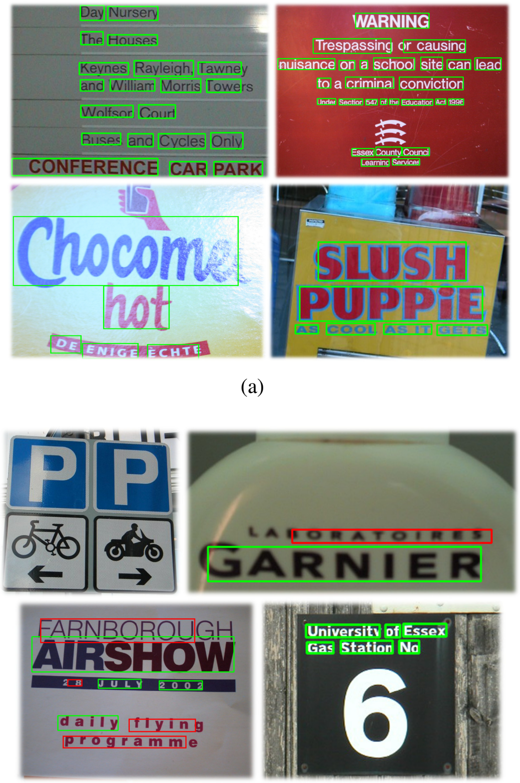
(b)

Figure 7. Detection examples of our model on MSRA-TD500. (a) Chinese text lines can also be detected due to the model generalization. (b) failure cases for complicated background or wide character space.

without adopting the provided training set or any other Chinese text data. Since our method could only detect text in word level, we implement line grouping method based on heuristic rules in post processing. Although our model shows strong compatibility for both English and Chinese, we still fail to detect Chinese text lines that have wide character spaces or complex background. Part of our detection results are shown in Fig.7.

Table 3. Comparison of methods on ICDAR2013 dataset.

Algorithm	Precision	Recall	F-measure
Tian *et al.* [3]	0.93	0.83	0.88
Proposed	**0.92**	**0.81**	**0.86**
Zhang *et al.* [1]	0.88	0.78	0.83
He *et al.* [33]	0.93	0.73	0.82
Tian *et al.* [8]	0.85	0.76	0.80



(a)

(b)

Figure 8. Detection examples of our model on ICDAR2013. (a) word level results for clustered texts. (b) failure cases for single character text and losing characters at either end.

ICDAR2013. The detection results of our method on the ICDAR2013 dataset are shown in Tab.3. The performance of our method is very competitive to the best previous result on this dataset. Failed cases are mainly caused by single character text and the inability to enclose letters at either end. Part of our detection results are shown in Fig.8.

4.4. Rationality of High Performance

The proposed method is intrinsically able to detect texts of arbitrary orientation, and able to partition words automatically. The tremendous improvements in both precision and recall for incidental text is mainly attributed to two aspects. First, the powerful feature representation learned by deep convolutional neural network guarantees high recall. Blurred or low resolution texts that can be hardly extracted by MSER [6], SWT [4] or other low-level feature based methods [5] are well detected by the classification task. Moreover, the classifier is also able to distinguish text-like regions providing a solid foundation for regression. Sec-

Table 4. Time cost comparison between different input sizes. *STD*,*LRG*, *GNT* refer the 320×320, 640×640, 1280×720 input size, respectively. *Feature*, *Fusion*, *Task*, *Patch*, *Image*, *Loose* represent the feature extraction, multi-level fusion, multi-task, patch level, image level and loose condition respectively.

Stage	*STD*	*LRG*	*GNT*
Feature	0.017	0.037	\
Fusion	0.005	0.013	\
Task	0.001	0.001	\
Patch	0.024	0.051	\
Image	1.4	1.1	\
Loose	\	\	0.2

ond, the learned mechanism to localize text is much more robust than rule based methods and this ensures the high precision. Treating word partition as a post processing is prone to lose much useful information and rely on thresholds chosen, but integrating localization into the network for end-to-end training could well solve the mentioned issues above.

4.5. Time Consumption Analysis

Differing from the timing analysis in generic object detection, for scene text which varies much in scales, a multi-scale detection strategy is essential. Therefore, we analyze the time consumption in both patch level and image level.

For patch level, we compare the time cost of each subnetwork for two input sizes: 320×320 and 640×640. Timing statistics are listed in Tab.4 and it is obvious to find that the time cost is a little more than double as the input image size increases twice.

For image level, set the image size 1280×720, the same as that in ICDAR2015 dataset and perform the multi-scale sliding window test strategy as illustrated in Sec.3.4 with stride equal to the window size. The multi-scale set remains unchanged and then according to the patch level time cost listed in Tab.4, we can easily estimate the consumed time for the whole image: for patch size 320×320 the total time cost is $1.4s$ and for 640×640 the cost is reduced to $0.9s$. Thus, enlarging the window size would be an effective way for speeding up.

However, if we consider a slightly loose condition than the ICDAR2015 dataset that we only need to detect text whose short side is larger that 20 pixels, we can directly enlarge the input size to 1280×720 and the time cost would be greatly reduced to about $0.2s$.

5. Conclusion

In this paper, we proposed a novel method for word level arbitrarily-oriented text detection based on the viewpoint that text is a special object in detection task and can be detected as a whole without generating character proposals. Without character detection, line grouping and word partition in traditional text detection pipeline, our framework is optimized in an end-to-end manner and directly outputs the quadrilateral boundary of each word benefiting from the multi-task design. On the ICDAR2015 Incidental Scene Text benchmark, our method achieved the state-of-the-art performance and outperformed previous methods by a large margin. Time consumption analysis also shows that our method is fast.

References

[1] 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*, 2016. 1, 2, 3, 5, 6

[2] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2

[3] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *Proceedings of the European Conference on Computer Vision*, pages 56–72. Springer, 2016. 1, 2, 5, 6

[4] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. Detecting text in natural scenes with stroke width transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2963–2970. IEEE, 2010. 1, 6

[5] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu. A hybrid approach to detect and localize texts in natural scene images. In *Proceedings of the IEEE Transactions on Image Processing*, volume 20, pages 800–813. IEEE, 2011. 1, 6

[6] Huizhong Chen, Sam S Tsai, Georg Schroth, David M Chen, Radek Grzeszczuk, and Bernd Girod. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *Proceedings of the 18th IEEE International Conference on Image Processing*, pages 2609–2612. IEEE, 2011. 1, 6

[7] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. Robust text detection in natural scene images. In *Proceedings of the IEEE transactions on Pattern Analysis and Machine Intelligence*, volume 36, pages 970–983. IEEE, 2014. 1, 6

[8] Shangxuan Tian, Yifeng Pan, Chang Huang, Shijian Lu, Kai Yu, and Chew Lim Tan. Text flow: A unified text detection system in natural scene images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 1, 2, 6

[9] Qixiang Ye and David Doermann. Text detection and recognition in imagery: A survey. In *Proceedings of the IEEE transactions on Pattern Analysis and Machine Intelligence*, volume 37, pages 1480–1500. IEEE, 2015. 1

[10] Zheng Zhang, Wei Shen, Cong Yao, and Xiang Bai. Symmetry-based text line detection in natural scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2558–2567, 2015. 1, 2

[11] Zhuoyao Zhong, Lianwen Jin, Shuye Zhang, and Ziyong Feng. Deeptext: A unified framework for text proposal generation and text detection in natural images. In *arXiv preprint arXiv:1605.07314*, 2016. 1, 2

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proceedings of the European Conference on Computer Vision*, pages 346–361. Springer, 2014. 1, 2

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 1, 2, 3

[15] 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. 1, 2

[16] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proceedings of the International Conference on Learning Representations*, 2014. 1, 2

[17] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1, 2

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2016. 1, 2

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott Reed. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, 2016. 1, 2

[20] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. In *arXiv preprint arXiv:1509.04874*, 2015. 1, 2, 3

[21] 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. 2

[22] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1083–1090. IEEE, 2012. 2, 6

[23] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang, and Hong-Wei Hao. Multi-orientation scene text detection with adaptive clustering. In *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 37, pages 1930–1937. IEEE, 2015. 2, 6

[24] 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. IEEE, 2014. 2, 6

[25] Chucai Yi and YingLi Tian. Text string detection from natural scenes by structure-based partition and grouping. In *Proceedings of the IEEE Transactions on Image Processing*, volume 20, pages 2594–2605. IEEE, 2011. 2

[26] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1134–1142, 2015. 2

[27] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. In *Proceedings of the International Journal of Computer Vision*, volume 104, pages 154–171. Springer, 2013. 2

[28] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3286–3293, 2014. 2

[29] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision*, pages 391–405. Springer, 2014. 2

[30] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678. ACM, 2014. 5

[31] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 9, pages 249–256, 2010. 5

[32] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *Proceedings of the 13th International Conference on Document Analysis and Recognition*, pages 1156–1160. IEEE, 2015. 5

[33] Tong He, Weilin Huang, Yu Qiao, and Jian Yao. Text-attentional convolutional neural network for scene text detection. In *Proceedings of the IEEE Transactions on Image Processing*, volume 25, pages 2529–2541. IEEE, 2016. 6