Deep Direct Regression for Multi-Oriented Scene Text Detection

Wenhao He¹,²  Xu-Yao Zhang¹  Fei Yin¹  Cheng-Lin Liu¹,²
¹ National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences, Beijing, China
² University of Chinese Academy of Sciences, Beijing, China
Email: {wenhao.he, xyz, fyin, liucl}@nlpr.ia.ac.cn

Abstract

In this paper, we first provide a new perspective to divide existing high performance object detection methods into direct and indirect regressions. Direct regression performs boundary regression by predicting the offsets from a given point, while indirect regression predicts the offsets from some bounding box proposals. In the context of multi-oriented scene text detection, we analyze the drawbacks of indirect regression, which covers the state-of-the-art detection structures Faster-RCNN and SSD as instances, and point out the potential superiority of direct regression. To verify this point of view, we propose a deep direct regression based method for multi-oriented scene text detection. 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 bi-task outputs where one is pixel-wise classification between text and non-text, and the other is direct regression to determine the vertex coordinates of quadrilateral text boundaries. The proposed method is particularly beneficial to localize incidental scene texts. On the ICDAR2015 Incidental Scene Text benchmark, our method achieves the F-measure of 81%, which is a new state-of-the-art and significantly outperforms previous approaches.

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 significant achievements of deep convolutional neural network (CNN) based generic object detection in recent years, scene text detection also has been greatly improved by regarding text words or lines as objects. High performance methods for object detection like Faster-RCNN [19], SSD [14] and YOLO [18] have been modified to detect horizontal scene texts [27] [5] [21] [13] and gained great improvements. However, for multi-oriented text detection, methods like Faster-RCNN and SSD which work well for object and horizontal text detection may not be good choices. To illustrate the reasons, first we explain the definitions of indirect and direct regression in detection task.

Indirect Regression. For most CNN based detection methods like Fast-RCNN [3], Faster-RCNN, SSD, Multi-Box [2], the regression task is trained to regress the offset values from a proposal to the corresponding ground truth (See Fig.1.a). We call these kinds of approaches indirect regression.

Direct Regression. For direct regression based methods, the regression task directly outputs values for the position and size of an object from a given point (See Fig.1.b). Take DenseBox [7] as an instance, this model learns to directly predict offsets from bounding box vertexes to points in region of interest.

Indirect regression based detection methods may not be effective for multi-oriented text detection, even methods like Faster-RCNN and SSD have reached state-of-the-art performance for object detection and are also implemented for horizontal scene text detection. The reasons are mainly
in three folds. First, there are few robust methods to generate word-level or line-level proposals for multi-oriented text. Most previous methods could only provide proposals of character-level by extracting connected components. Second, anchor mechanism in Faster-RCNN may not be an effective solution to generate text proposals. The anchor mechanism can be deemed as rectangular proposals of various sizes and aspect ratios being evenly placed on an image, and setting proposals which have high overlap with ground truths as positive, otherwise as “NOT CARE” or negative. However, for multi-oriented scene texts which are long and heavily inclined, there may be no proper anchor that has sufficient overlap with them as shown in Fig. 2. Third, adopting anchor mechanism may cause the detection system less efficient. Taking horizontal scene text detection as instance, unlike generic objects, horizontal scene texts tend to have larger variation in sizes and aspect ratios, which requires more complicate design of anchors. The anchors used in [27] [13] are much more than traditional Faster-RCNN in both scale and aspect ratio. As to multi-oriented text detection, inclined text proposals may be generated by adopting multi-oriented anchors like [15], however, this will cost much more running time in the meanwhile and the proposal may not be an optimal choice. Based on the analysis above, direct regression based methods which need no proposals beforehand could be a better choice to produce the irregular quadrilateral boundaries for multi-oriented scene texts.

In this paper, we propose a novel multi-oriented text detection method based on direct regression. Our method is particularly beneficial to localize quadrilateral boundaries of incidental scene texts which are hard to identify the constituent characters and have large variations in scales and perspective distortions. On the ICDAR2015 Incidental Scene Text benchmark, we obtain F-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 reaches the state-of-the-art performance.

The proposed method has several novelties and advantages. First, this is the first direct regression based method for multi-oriented scene text detection. Second, the whole pipeline of the proposed method only has two parts in which one is a convolutional neural network and the other is a one-step post processing call Recalled Non-Maximum Suppression. Modules like line grouping and word partition are removed which saves much effort on tuning parameters. Third, since our method could predict irregular quadrilateral boundaries, it has great superiority in incidental texts detection task which needs to localize four vertexes of each word-level text.

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

2. Related Work

Scene Text Detection. Most scene text detection methods [26] [21] [8] [1] [17] 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 [23] [24] [10] 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 texts in the ICDAR2015 Incidental Scene Text Dataset [11], 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 [25] 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 [5] adopts the framework for object detection in [18], but the post-processing relies on the text sequentiality. The methods in [27] and [13] are based on Faster-RCNN [19] and SSD [14] respectively. They both attempt to convert text detection into object detection and the performance on horizontal text detection demonstrate their effectiveness. However, constrained by the deficiency of indirect regression, those two methods may not be suitable for multi-oriented scene text detection. The method in [15] rotates the anchors into more orientations and tries to find the best proposal to match the multi-oriented text. Deficiency of this method is that the best matched proposal may
not be an optimal choice since the boundary shape of scene texts is arbitrary quadrilateral while the proposal shape is parallelogram.

**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 direct and indirect regression. The direct regression based methods like [7] predict size and localization of objects straightforwardly. The indirect regression based methods like [3] [19] [2] [14] predict the offset from proposals to the corresponding ground truths. It should be noted that, the proposals here can be generated by either class-agnostic object detection methods like [22] or simple clustering [2], as well as anchor mechanism [19] [14].

Although most of the recent state-of-the-art approaches are indirect regression based methods, considering the wide variety of texts in scale, orientation, perspective distortion and aspect ratio, direct regression might have the potential advantage of avoiding the difficulty in proposal generation for multi-oriented texts. This is the main contribution of this paper.

**3. Proposed Methodology**

The proposed detection system is diagrammed in Fig.3. 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 $S$. This ensures the regression task could see long texts and give more accurate boundary prediction. Considering that the text feature is not as complicated as that of generic objects, our network tends to employ less parameters than models designed for ImageNet to save computation.

The feature fusion part referring to the design in [16] combine convolutional features from four streams to capture texts of multiple scales. However, to reduce computation, we only up-sample the fused feature to quarter size of the input image.

The multi-task part has two branches. The classification task output $M_{cls}$ is a $\frac{S}{4} \times \frac{S}{4}$ 2nd-order tensor and it can be approximated as down-sampled segmentation between text and non-text for input images. Elements in $M_{cls}$ with higher score are more likely to be text, otherwise non-text; The regression task output $M_{loc}$ is a $\frac{S}{4} \times \frac{S}{4} \times 8$ 3rd-order tensor. The channel size of $M_{loc}$ indicates that we intend to output 8 coordinates, corresponding to the quadrilateral vertexes of the text. The value at $(w, h, c)$ in $M_{loc}$ 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 $B(w, h)$ can be formulated as

$$B(w, h) = \begin{cases} L_{(w, h, 2n-1)} + 4w, L_{(w, h, 2n)} + 4h | n \in \{1, 2, 3, 4\} \end{cases}$$ (1)

By combining outputs of these two tasks, 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.4.

**3.2. Ground Truth and Loss Function**

The full multi-task loss $L$ can be represented as

$$L = L_{cls} + \lambda_{loc} \cdot L_{loc},$$ (2)

where $L_{cls}$ and $L_{loc}$ represent loss for classification task and regression task respectively. The balance between two losses is controlled by the hyper-parameter $\lambda_{loc}$.

**Classification task.** Although the ground truth for classification task can be deemed as a down-sampled segmentation between text and non-text, unlike the implementation in [26], we do not take all pixels within 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 “NOT CARE” boundary as transition from positive to negative (shown in Fig.5). The parameter $r$ is proportional to the short side of text boundaries and its value is 0.2.

Furthermore, text is taken as a positive sample only when its short side length ranges in $[32 \times 2^{-1}, 32 \times 2^{1}]$. If the short side length falls in $[32 \times 2^{-1.5}, 32 \times 2^{-1}] \cup [32 \times 2^{1}, 32 \times 2^{1.5}]$, we take the text as “NOT CARE”, otherwise negative. “NOT CARE” regions do not contribute to the training objective. Ground truths designed in this way reduce the confusion between text and non-text, which is beneficial for discriminative feature learning.

The loss function $L_{cls}$ chosen for classification task is the hinge loss. Denote the ground truth for a given pixel as $y_i^t \in \{0, 1\}$ and predicted value as $y_i$, $L_{cls}$ is formulated as

$$L_{cls} = \frac{1}{S^2} \sum_{i \in L_{cls}} \max(0, \text{sign}(0.5 - y_i^t) \cdot (y_i - y_i^t))^2$$ (3)

Besides this, we also adopt the class balancing and hard negative sample mining as introduced in [7] for better performance and faster loss convergence. Hence during training, the predicted values for “NOT CARE” region and easily classified negative area are forced to zero, the same as the ground truth.
Regression task. Considering that the ground truth values of regression task vary within a wide range, we use a Scale&Shift module (shown in Fig.4) for fast convergence. Scale&Shift takes the value $z$ from a sigmoid neuron as input and stretch $z$ into $\hat{z}$ by

$$\hat{z} = 800 \cdot z - 400, \quad z \in (0, 1)$$

Here we assume that the maximum positive text size is less than 400. We also use a sigmoid neuron to normalize the values before Scale&Shift for steady convergence.

According to [3], the loss function $\mathcal{L}_{loc}$ used in regression task is 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

$$\mathcal{L}_{loc} = \sum_{\forall \hat{y}_i^* > 0} \text{smooth}_{L_1}(z^*_i - \hat{z}_i),$$

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

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 getting the outputs produced by multi-task learning, each point of the output map is related with a scored quadrilateral. To filter the non-text region, we only preserve points with high score in classification task. However, there will be still densely overlapped quadrilaterals for a word or text line. To reduce the redundant results we propose a post-processing method called Recalled Non-Maximum Suppression.

The Recalled NMS is a trade-off solution for two problems: (i) when texts are close, quadrilaterals between two words are often retained because of the difficulty in classifying pixels within word space, (ii) if we solve problem (i) by simply retaining quadrilaterals with higher score, text region with relative lower confidence will be discarded and the overall recall will be sacrificed a lot. The Recalled NMS could both remove quadrilaterals within text spaces and maintain the text region with low confidence.

The Recalled NMS has three steps as shown in Fig.6.

• First, we get suppressed quadrilaterals $B_{sup}$ from densely overlapped quadrilaterals $B$ by traditional NMS.
• Second, each quadrilateral in $B_{sup}$ is switched to the one with highest score in $B$ beyond a given overlap. After this step, quadrilaterals within word space are changed to those of higher score and low confidence text region are preserved as well.
• Third, after the second step we may get dense overlapped quadrilaterals again, and instead of suppression, we merge quadrilaterals in $B_{sup}$ which are heavily overlapped with each other.

Figure 3. Overview of the proposed text detection method.

Figure 6. 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 closeness.
3.4. Network Implementation

The training samples of $320 \times 320$ are cropped from scaled images rotated randomly by $0$, $\pi/2$, $\pi$, or $3\pi/2$. The task balance index $\lambda_{loc}$ is raised from 0.01 to 0.5 after the classification task gets well trained. The network should learn what the text is first and then learn to localize the text. In testing, we adopt a multi-scale sliding window strategy in which window size is $320 \times 320$, sliding stride is 160 and multi-scale set is $\{2^{-5}, 2^{-4}, \cdots, 2^1\}$. Pixels on $M_{cls}$ 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 optimizing regression task. Lack-
ing of line level annotation and sufficient bi-lingual training data, we did not use the training set and instead, we 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 is introduced in [12].

### 4.2. Implementation Details

The network is optimized by stochastic gradient descent (SGD) with back-propagation and the max iteration is $2 \times 10^5$. We adopt the “multistep” strategy in Caffe [9] to adjust learning rate. For the first $3 \times 10^4$ iterations the learning rate is fixed to be $10^{-2}$ and after that it is reduced to $10^{-5}$ until the $10^3$th iteration. For the rest $10^5$ iterations, 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 $3 \times 10^4$th iteration. Weight decay is $4 \times 10^{-4}$ and momentum is 0.9. All layers except in regression task are initialized by “xavier” [4] and the rest layers are initialized to a constant value zero for stable convergence.

The model is optimized on training datasets from ICDAR2013 and ICDAR2015. 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.

### 4.3. Experimental Results

**ICDAR2015 Incidental Scene Text.** The results shown in Tab.1 indicate that the proposed method outperforms previous approaches by a large margin in both precision and recall. To demonstrate the effectiveness of Recalled NMS, we list the result adopting traditional NMS as the post-processing. From Tab.1 we can see the Recalled NMS give a higher precision mainly because of filtering quadrilaterals between text lines.

We also list the detection result running on the original image size, which demonstrate the necessity of testing on multiple scales. The precision value on single scale does not drop indicating the effectiveness by treating text in a proper scale range as positive samples.

Note that the method in [15] which ranks second is indirect regression based multi-oriented text detection and it also treats text detection as object detection. The large margin between our method and this method demonstrates our analysis on the deficiency of indirect regression and superiority of direct regression for multi-oriented text detection. Some examples of our detection results are shown in Fig.7.

**MSRA-TD500.** The results of our method on this dataset are shown in Tab.2, with comparisons to other representative results of state-of-the-art methods. It is shown that our method could reach the state-of-the-art performance. It should be noted that we did not adopt 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. Our model shows strong compatibility for both English and Chinese, however, 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.8.

**ICDAR2013 Focused Scene Text.** The detection results of our method on the ICDAR2013 dataset are shown in Tab.3. The performance of our method is also the new state-of-the-art. Apart from the precision, recall and F-measure, we also list the time cost of our method for per image. From the Tab.3 we can see our method is also competitively fast in running speed. 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.9.

---

**Table 1. Comparison of methods on ICDAR2015 Incidental Scene Text dataset.** R-NMS is short for Recalled NMS and T-NMS is short for traditional NMS. SS is short for single scale.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed (R-NMS)</td>
<td>0.82</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>Proposed (T-NMS)</td>
<td>0.81</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Liu et al. [15]</td>
<td>0.73</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td>Proposed (SS)</td>
<td>0.82</td>
<td>0.62</td>
<td>0.70</td>
</tr>
<tr>
<td>Tian et al. [21]</td>
<td>0.74</td>
<td>0.52</td>
<td>0.61</td>
</tr>
<tr>
<td>Zhang et al. [26]</td>
<td>0.71</td>
<td>0.43</td>
<td>0.54</td>
</tr>
<tr>
<td>StradVision2 [11]</td>
<td>0.77</td>
<td>0.37</td>
<td>0.50</td>
</tr>
<tr>
<td>StradVision1 [11]</td>
<td>0.53</td>
<td>0.46</td>
<td>0.50</td>
</tr>
<tr>
<td>NJU-Text [11]</td>
<td>0.70</td>
<td>0.36</td>
<td>0.47</td>
</tr>
<tr>
<td>AJOU [11]</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>HUST_MCLAB [11]</td>
<td>0.44</td>
<td>0.38</td>
<td>0.41</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.77</td>
<td>0.70</td>
<td>0.74</td>
</tr>
<tr>
<td>Zhang et al. [26]</td>
<td>0.83</td>
<td>0.67</td>
<td>0.74</td>
</tr>
<tr>
<td>Yin et al. [24]</td>
<td>0.81</td>
<td>0.63</td>
<td>0.71</td>
</tr>
<tr>
<td>Kang et al. [10]</td>
<td>0.71</td>
<td>0.62</td>
<td>0.66</td>
</tr>
<tr>
<td>Yao et al. [23]</td>
<td>0.63</td>
<td>0.63</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**Table 3. Comparison of methods on ICDAR2015 Focused Scene Text dataset.**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.92</td>
<td>0.81</td>
<td>0.86</td>
<td>0.9s</td>
</tr>
<tr>
<td>Liao et al. [13]</td>
<td>0.88</td>
<td>0.83</td>
<td>0.85</td>
<td>0.73s</td>
</tr>
<tr>
<td>Zhang et al. [26]</td>
<td>0.88</td>
<td>0.78</td>
<td>0.83</td>
<td>2.1s</td>
</tr>
<tr>
<td>He et al. [6]</td>
<td>0.93</td>
<td>0.73</td>
<td>0.82</td>
<td>–</td>
</tr>
<tr>
<td>Tian et al. [20]</td>
<td>0.85</td>
<td>0.76</td>
<td>0.80</td>
<td>1.4s</td>
</tr>
</tbody>
</table>
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 three aspects.

First, direct regression based detection structure avoids to generate proper proposals for irregular shaped multi-oriented texts and thus is more straightforward and effective for multi-oriented scene text detection.

Second, the restriction of positive text size guarantees the robustness of feature representation learned by deep convolutional neural networks. Features for small texts could fade a lot after the first down-sampling operations, and large texts would lose much context information causing the CNN could only see some simple strokes of the large texts. Texts within a proper scale range could contain both text textures and enough semantic context making the CNN learn more robust scene text features. Moreover, the classification task which is able to distinguish text and non-text regions providing a solid foundation for regression task.

Third, the end-to-end optimization mechanism to localize text is much more robust than rule based methods. Previous methods treating line grouping and word partition as post processing are 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. Comparison to Other Scene Text Detection Systems

Here we list and compare with some recent high performance scene text detection methods for better understanding on the superiority of our method. The listed methods are arranged by the time they are proposed.

**TextBoxes & DeepText.** TextBoxes [13] and DeepText [27] are based on SSD and Faster-RCNN respectively. They both take advantages from the high performance object detection systems and treat text word/line as a kind of generic object. Moreover, they both set anchors to have more varieties and can only detect horizontal scene texts. In our work, we perform the regression by a direct way and can tackle with multi-oriented text detection.

**DMPN.** DMPN [15] is an indirect regression based method and it also treats text detection as object detection. Unlike TextBoxes or DeepText, it introduces a multi-oriented anchor strategy to find the best matched proposal in parallelogram form to the arbitrary quadrilateral boundaries of multi-oriented texts. However, as [15] itself refers, DMPN relies on the man-made shape of anchors which may not be the optimal design and this fits well with our analysis on the drawbacks of indirect regression. The large margin of performance between DMPN and our method on ICDAR2015 Incidental Text benchmark also verify the significance of our work.

5. Conclusion

In this paper, we first partition existing object detection methods into direct and indirect regression, and analyze the pros and cons of both methods for irregular shaped object detection. Then we propose a novel direct regression based method for multi-oriented scene text detection. Our detection framework is straightforward and effective with only one-step post processing. Moreover it performs particularly well for incidental text detection. On the ICDAR2015 Incidental Scene Text benchmark, we have achieved a new state-of-the-art performance and outperformed previous methods by a large margin. Apart from this, we also analyze the reasons of the high performance and compare our method to other recent scene text detection systems. Future work will focus on more robust and faster detection structure, as well as more theoretical research on regression task.

Acknowledgment

This work has been supported by the National Natural Science Foundation of China (NSFC) Grant No.61411136002.

References


