Scene Text Detection with Text Statistical Characteristics and Deep Neural Network

Yanyun Qu, Xiaodong Yang(✉), and Li Lin

Xiamen University, Xiamen, China
yyqu@xmu.edu.cn, 4490736262@qq.com, 379484437@qq.com

Abstract. Scene text recognition is a hot topic in the field of computer vision. Inspired by the success of the Single Shot Multibox Detection (SSD) on generic object detection, the architecture of SSD is implemented on scene text detection. SSD does not do well on text detection, because scene text as an object is usually smaller than a generic object and SSD cannot detect small objects well. Thus, the statistic analysis for scene text is made. Based on statistic characteristics of scene text, we propose a method named Text-SSD to detect scene text. Moreover, in order to boost the detection accuracy, multi-scale image are used to learn the multi-scale models. The voting based non-maximum suppressing is made for a candidate text region. The experimental results show that our method achieved the state-of-the-art performance on the benchmark dataset ICDAR2013 in the detection accuracy. Moreover, when using a single model, our method achieves the fastest speed compared with several latest text detection method based on deep neural network. Thus, experimental results demonstrate our method is efficient on scene text detection.

Keywords: Scene text detection · Text statistical characteristic · SSD

1 Introduction

Due to its widely applications in image retrieval, visual navigation, and scene understanding, etc., scene text detection has attracted more and more attentions [10–12]. Though tremendous efforts have been devoted to text detection, scene text detection in a wild is still a challenging task because of unconstrained environments. Moreover, scene text is flexible distributed with the changes of fonts, style and layout, and so on. Until now, the text detection accuracy is not high, which will greatly influence the successive text recognition.

Great progresses have been witnessed in object detection in recent years. Inspired by the successes of deep learning on image classification and speech recognition, object detection is solved in an end-to-end way based on a convolutional neural network, which is different from the pipeline scheme of traditional object detection methods. Especially, Single Shot Multibox Detector (SSD) has

J. Yang et al. (Eds.): CCCV 2017, Part III, CCIS 773, pp. 245–254, 2017.
https://doi.org/10.1007/978-981-10-7305-2_22

achieved the breakthrough performance in detection accuracy. And SSD is suitable for generic object detection such as dogs, bikes, persons, etc. However, it cannot do well on text detection if it is implemented directly on scene text detection. The failure is due to the following two reasons: (1) Scene text is usually smaller than generic objects. In ILSVRC and VOC, an object often makes up no smaller than 10% of the whole image. However, scene text often makes up much smaller than a generic object. (2) The aspect ratio of scene text is different from a generic object. Scene text is usually like a horizontal thin bar, while a generic object can be bounded by a rectangle. In order to make SSD correctly detect text, we treat scene text as an object, and we mine the text statistical characteristics and design a text-specified default box according to text statistical characteristics in the SSD framework, which is named Text-SSD. The text-specified default box can efficiently get the discriminative features of scene text, and make significant improvements of scene text detection in accuracy and speed.

The remainder of this paper is organized as follows. We introduce related work in Sect. 2. In Sect. 3, we detail how to mine the text statistical characteristics and how to design the default boxes. Experimental results are presented in Sect. 4, and we conclude in Sect. 5.

2 Related Work

In recent years, efforts have been devoted to scene text detection [1–8]. There are two typical classes of text detection methods, one of which are traditional methods and the other of which are deep learning based methods. Most of the traditional methods solve the text detection problem in a pipeline way. Yao et al. [11] used SWT to extract the connected components and then filtered out non-character regions by using a Random Forest classifier combined with character-level features. After that, they connected the candidate character regions into strings according to the similarity of geometric structures, spatial layouts, etc. Finally, they filtered out non-text regions by using a string-level classier. Neumann and Matas [10] extracted External Regions (ER) of an image as candidate regions. After that, the incremental features of an ER are extracted, and then, a two-stage classifier is learned from the training data to filter out the non-text candidate ER. In the first stage, a real AdaBoost classifier formed by decision trees, is implemented. In the second stage, a SVM classifier with RBF kernel is implemented. Finally, the exhaustive search [9] is used to find out the scene text regions. This method is more efficient and robust compared with other text detection methods before it.

With the upsurge of deep learning, deep learning based text detection methods have surged. They can be divided into two classes according to their framework: pipeline methods and end-to-end methods. The former uses the traditional bottom-up text detection framework, in which CNN is used to extract the text features instead of hand-crafted features [1,4,16]. The latter adopts the end-to-end framework which are used for generic object detection or segmentation

[14,15,17,18] for scene text detection. Compared to pipeline methods, they have significantly improved scene text detection in both accuracy and speed. Huang et al. [1] adopted CNN combined with MSER for text detection. In this method, MSER-tree was firstly constructed according to which candidate regions were extracted. After that, CNN was implemented to filter out the non-text regions. Furtherly, text lines were constructed based on simple features such as intensity, color, height and width etc. This method improves the robustness of MSER-based text detection methods. Jaderberg et al. [16] used CNN together with sliding window scheme for scene text detection. Each sliding window was scored by the CNN classifier and a saliency map is computed. According to the saliency map, text lines are formed. In addition, CNN can be extended to solve the 62-class classification problem. In other word, this method can not only be used for text detection, but also for text recognition in an end-to-end way.

He et al. [17] firstly transforms the text detection problem to a segmentation problem. Fully Connected Network [24] is used for text detection. The method used two CNNs, the first CNN was used to detect a text block, and the second CNN was used to extract the text lines in each text block.

Yao et al. [18] also regarded the text detection as a semantic segmentation problem which was solved by a deep neural network named HED [19]. And it can be used for multi-task text recognition, such as text proposal detection, single character recognition, and multiple character detection.

Tian et al. [15] proposed the Connectionist Text Proposal Network (CTPN) combined with the Recurrent Neural Network (RNN) for text detection. CTPN firstly used improved RPN (region proposal network) network to select the character candidate, and then used anchors with fixed width to detect the candidate area of text. After that, the feature vectors corresponding to the anchors in the same row are concatenated into a sequence, which is then put into RNN for text recognition. At last, the fully connection layer was used for classifying and regression, and the correct candidate regions are merged into a text line.

Liao et al. [14] modified the Single Shot MultiBox Detector (SSD) [13] for text detection. The convolutional filters and the aspect ratio of the sliding window are adjusted. Furthermore, a multi-scale inputs were used for improving the detection effect, but the speed is greatly reduced. Liao's method is the closest related work to ours. However, their method does not design the default box depending on the text statistical characteristics but depending on the empirical results. Thus, the model can be improved for text detection.

3 Multi-scale Text-SSD

In this section, we introduce how to modify the original SSD for scene text detection. As we know, the text scale is a very important factor for text detection, and any single model of text detection cannot do well on all text scales. Thus, in order to boost the detection accuracy, a multi-scale SSD model is designed for scene text detection, which is named muti-scale Text-SSD. As shown in Fig. 1, the proposed approach has three steps: (1) An input image is rescaled into

three scales: 512 * 512, 700 * 512, 700 * 300. (2) For each scale, a SSD model is designed for text detection. (3) A voting fusion strategy is used to obtain the text candidate regions.

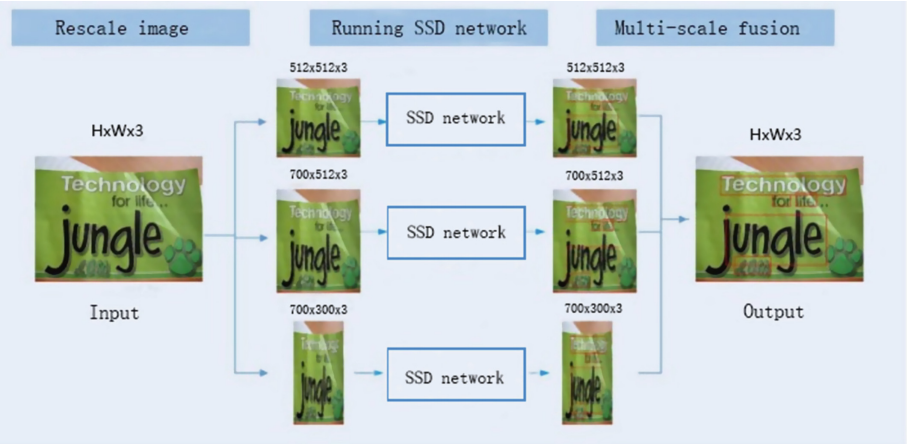

Fig. 1. The framework of multi-scale Text-SSD

3.1 Brief Review of SSD

SSD includes five parts: (1) The first part includes the first five convolution modules (conv1 to conv5_3) of VGG-16. (2) In the second part, the two fully connected layers fc_6, fc_7 of VGG-16 are modified as full convolution layers. (3) In the third part, four convolution modules from conv8 to conv11 are added, each of which contains 1×1 convolutions and 3×3 convolutions. (4) In the fourth part, the classification decision and its responding regression boxes of the last four convolution layers from conv8 to conv11 are output, in which 3×3 convolution modules are used to predict the classification label and the position of candidate regions. (5) Finally, the non-maximal suppression is performed to obtain the final result. SSD uses the sliding window scheme on a feature map instead of the original image. In addition, SSD uses the idea of Anchor in Faster R-CNN, so that the location of each sliding window corresponds to an anchor with different scales and aspect ratios, which is called default box. Default boxes are used in the six convolution layers in SSD: conv4_3, conv7, conv8_2, conv9_2, conv10_2, and conv11_2, and in each of the last four convolution layers, the default boxes are only used in the 3 * 3 convolution for class label prediction. There are five initial aspect ratios in SSD: $a_r = \{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$. The parameter s_k is denoted the scale of the kth convolution layer. In the shallowest convolution layer, s_{min} is set to be 20%, and in the deepest convolution layer, s_{max} is set to be 90%. In the middle convolution layers, the scale is formulated as,

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1}(k - 1), k = 1, 2, ..., m \quad (1)$$

where k is the order of the convolution layer in which default boxes are implemented. There are six default boxes in each position of a feature map, and the width and height are formulated as, $w_k = s_k\sqrt{a_r}$ and $h_k = s_k/\sqrt{a}$.

3.2 Limit of SSD on Text Detection

Here, we firstly find out what result in the failure of SSD on text detection. We implement SSD with its default settings which are shown in Table 1 on text detection. Some results are shown in Fig. 2. The bounding boxes signed in green are results of the original SSD detection, and the ground truth boxes are signed in red. It shows that SSD do not work well on text detection because SSD cannot deal with small objects.

This empirical results demonstrates that the inappropriate setting of the default box lead to a high missed rate and error rate. Thus, the original SSD can not be implemented directly on scene text detection, and default box setting should be modified to adapt to scene text detection.

Table 1. Default box settings in scales for each layer on ICDAR2013

Name	Conv4_3	Fc7	Conv6_2	Conv7_2	Conv8_2	Conv9_2
SSD	10–20	20–37	37–54	54–71	71–88	88–100
Text-SSD	5–10	10–25	25–40	40–55	55–70	70–85

Fig. 2. Results of the original SSD on text detection

3.3 Text Statistical Characteristics

Default boxes in SSD have two parameters: scales and aspect ratios. In order to be able to set reasonable parameters of default boxes for text detection, the distribution of text sizes and aspect ratios are investigated, which guide us to design the default box.

We make statistics for scene text on training data of ICDAR2013: the distributions of text width, the distribution of text height, the distribution of text area and the distribution of the text aspect ratio. Figure 3(a) and (b) shows the histograms of text widths and heights, respectively, Fig. 3(c) shows the histogram of text area, and Fig. 3(d) shows the histogram of text aspect ratios. Observing

the statistical analysis of scene text, we can see that most text regions is relatively small, and the text shape is prone to long strip. The text areas averagely make up 4% of an image. The average of the width ratio between the text and the image is 26.9% and the height of the text is smaller than 10% of the height of an image. Thus, the initial default boxes settings exist three problems: (1) In the shallow layer (conv4_3) of neural network, the scale is too small to cover small object. (2) In the deep layer of neural network (conv8_2, conv9_2), the scale is too large, and too many backgrounds fall within the scale range. Thus, it introduces noises and results in the low overall accuracy. (3) Most text regions are detected successfully owing to the default box setting in the middle layers (fc7, conv6_2, conv7_2).

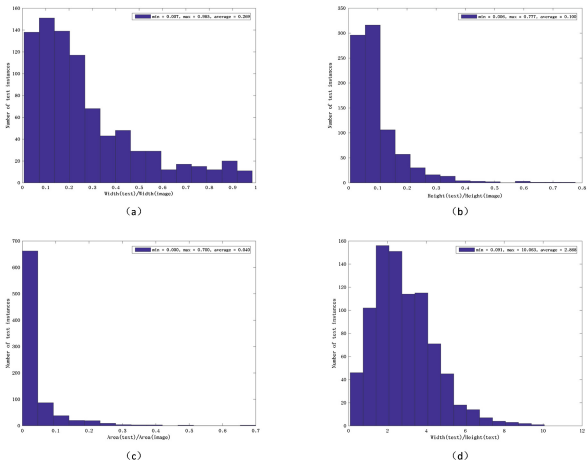

Fig. 3. Text statistical characteristics. (a) Text width ratio histogram. (b) Text height ratio histogram. (c) Text area ratio histogram. (d) Text aspect ratio histogram

We design the default box with the text-driven setting for text detection. In the text-driven setting, we modify the initial default box setting and reduce the scale of the default box in the shallow layer and make the width and height of the default boxes in the middle layers change linearly with the indicator of their layers, in order that the candidate text regions can be detected uniformly in all the layers. In Table 1, we give the scale of the default box. The first row shows the original default box setting, and the second row shows the scale setting according to the text-driven setting. Compared with the original SSD setting, there are fewer default boxes in the text-driven setting. We also set the aspect ratio of the default box as $a_r = \{1, 4, \frac{1}{4}\}$ according to the text statistical characteristics. In our experiment, we use nine default boxes with the three scales and three aspect ratios at each position.

3.4 Multi-scale Detection Fusion

When a query image is tested, it will be input in the three different Text-SSD models with three different scales. We will fuse the three results from three Text-SSDs based on non-maximum suppression. The algorithm is divided into the four steps: (1) We sort the detected bounding boxes in a descending order according to their confidence values. (2) We calculate the Jaccard overlap for each pairwise bounding boxes. (3) We remove the bounding box with a low confidence score and accumulate the voting scores for all overlapping bounding boxes. (4) After updating the score, the bounding box with the score of confidence lower than threshold is eliminated. Compared with the original non-maximal suppression, we can see that the voting based multi-scale fusion algorithm is more robust to noise.

4 Experimental Results

In this section, we implement the proposed multi-scale Text-SSD on ICDAR2013. ICDAR2013 is the competing text detection dataset and contains 229 training images and 233 testing images. We evaluate the text localization performance of Text-SSD by the standard criteria used in the competitions of ICDAR2013: the recall, the precision, and the F-score and we use the F-score as the final evaluation score. There are two standard evaluation methods used in ICDAR2013: IC13 and ICDet. We compare Text-SSD with 12 state-of-the-art methods, the six methods were published in 2015, and the other six methods are published in 2016. We also compare the single scale Text-SSD and multi-scale Text-SSD. As shown in Table 2, multi-scale Text-SSD ranks the first in the criteria of IC13 and ranks the second in the criteria of ICDet in the compared text detection methods. Multi-scale SSD is higher by about 2% than Gupta's method [5] which ranks the second in the criteria of IC13 and lower by about 1.2% than CPTN [15] which ranks the second in the criteria of ICDet. However, multi-scale Text-SSD is faster than Gupta's method. Single scale Text-SSD ranks the fifth in both the criteria of IC13 and ICDet, but it is the fastest in all the compared methods. Single scale Text-SSD achieve 11.6 Fps.

We also compare Textbox [14] which is the closest work to Text-SSD. The difference between ours and Textbox is that Text-SSD use text-specified default box setttings according to text statistical characteristics. We use fewer default boxes in SSD than Textbox. As shown in Table 3, Text-SSD is superior to TextBox in both single scale and multiple scales. Single scale Text-SSD is higher by 1.29% and 2.18% than single scale Textbox in terms of F-score in IC13 and ICDet, respectively. Multi-scale Text-SSD is higher by 1.24% and 0.83% than multi-scale Text-SSD in terms of F-score in IC13 and ICDet, respectively. Moreover, the speed of Text-SSD is faster than Textbox in both single scale and multiple scales. To sum up, Text-SSD is effective and robust in scene text detection.

Table 2. Comparison of state-of-the-art text detection methods on ICDAR2013

Name	Year	IC13			ICDet			Speed (FPS)
		Recall	Precision	F-score	Recall	Precision	F-score	
Text_Flow[3]	2015	75.89	85.15	80.25				<1
Neumann [21]	2015	72.4	81.8	77.1				3
Neumann [22]	2015	71.3	82.1	76.3				3
Busta [2]	2015	69.3	84	76.8				6
Zhang [20]	2015				74	88	80	<0.1
Yin [4]	2015	65.11	83.98	73.35				3
Zhang [8]	2016				78	88	83	<1
Gupta [5]	2016	76.4	**93.8**	84.2	75.5	92	83	<1
CTPN [15]	2016	74	93	82	83	**93**	**88**	7.1
SSD [13,14]	2016	60	80	69	60	80	69	10
Cho [23]	2016	78.45	86.26	82.17				7.69
He [6]	2016	73	93	82				<1
Single Text-SSD		76.62	86.57	81.29	77.1	87.98	82.18	**11.6**
Multi-scale Text-SSD		**82.83**	89.95	**86.24**	**83.18**	90.82	86.83	3.75

Table 3. Comparison between Text-SSD and TextBoxes on ICDAR2013

Method	IC13			ICDet			Speed (FPS)
	Recall	Precision	F-score	Recall	Precision	F-score	
Fast Textboxes [14]	74	86	80	74	88	80	11.1
TextBoxes [14]	83	88	85	83	89	86	1.37
Single Text-SSD	76.62	86.57	81.29	77.1	87.98	82.18	**11.63**
Multi-scale Text-SSD	**82.83**	**89.95**	**86.24**	**83.18**	**90.82**	**86.83**	3.75

5 Conclusion

In this paper, Text-SSD is proposed to detect scene text in which the default box is designed for scene text according to the text statistical characteristics. Moreover, in order to boost the performance of text detection, multi-scale Text-SSDs is used and the output are fused based on voting. Text-SSD is implemented on ICDAR2013, the experimental results demonstrate that the proposed method is superior to the state-of-the-art methods not only in the detection accuracy but also in the running time.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant 61373077.

References

1. Huang, W., Qiao, Y., Tang, X.: Robust scene text detection with convolution neural network induced MSER trees. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 497–511. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_33
2. Busta, M., Neumann, L., Matas, J.: FASText: efficient unconstrained scene text detector. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1206–1214 (2015)
3. Tian, S., Pan, Y., Huang, C., et al.: Text flow: a unified text detection system in natural scene images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4651–4659 (2015)
4. Yin, X.-C., Pei, W.-Y., Zhang, J., et al.: Multi-orientation scene text detection with adaptive clustering. IEEE Trans. Tattern Anal. Mach. Intell. **37**(9), 1930–1937 (2015)
5. Gupta, A., Vedaldi, A., Zissermana.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315–2324 (2016)
6. He, T., Huang, W., Qiao, Y., et al.: Text-attentional convolutional neural network for scene text detection. IEEE Trans. Image Process. **25**(6), 2529–2541 (2016)
7. Jaderberg, M., Simonyan, K., Vedaldi, A., et al.: Reading text in the wild with convolutional neural networks. Int. J. Comput. Vis. **116**(1), 1–20 (2016)
8. Zhang, Z., Zhang, C., Shen, W., et al.: Multi-oriented text detection with fully convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4159–4167 (2016)
9. Neumann, L., Matas, J.: Text localization in real-world images using efficiently pruned exhaustive search. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 687–691. IEEE (2011)
10. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3538–3545. IEEE (2012)
11. Yao, C., Bai, X., Liu, W., et al.: Detecting texts of arbitrary orientations in natural images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1083–1090. IEEE (2012)
12. Huang, W., Lin, Z., Yang, J., et al.: Text localization in natural images using stroke feature transform and text covariance descriptors. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1241–1248 (2014)
13. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
14. Liao, M., Shi, B., Bai, X., et al.: TextBoxes: a fast text detector with a single deep neural network. arXiv preprint arXiv:161106779 (2016)
15. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 56–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_4
16. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 512–528. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_34

17. He, T., Huang, W., Qiao, Y., et al.: Accurate text localization in natural image with cascaded convolutional text network. arXiv preprint arXiv:160309423 (2016)
18. Yao, C., Bai, X., Sang, N., et al.: Scene text detection via holistic, multi-channel prediction. arXiv preprint arXiv:160609002 (2016)
19. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403 (2015)
20. Zhang, Z., Shen, W., Yao, C., et al.: Symmetry-based text line detection in natural scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2558–2567 (2015)
21. Neumann, L., Matas, J.: Efficient scene text localization and recognition with local character refinement. In: 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 746–750. IEEE (2015)
22. Neumann, L., Matas, J.: Real-time lexicon-free scene text localization and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1872–1885 (2016)
23. Cho, H., Sung, M., Jun, B.: Canny text detector: fast and robust scene text localization algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3566–3573 (2016)
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)