A Weakly Supervised Text Detection Based on Attention Mechanism

Lanfang Dong(✉), Diancheng Zhou, and Hanchao Liu

School of Computer Science and Technology,
University of Science and Technology of China, Hefei, China
lfdong@ustc.edu.cn, {zdc0803, lhanchao}@mail.ustc.edu.cn

Abstract. In this paper, we propose a new method for natural image text detection under a weakly supervised data set. Currently, most of the text detection models are based on bounding box label training data. However, the cost of the bounding box label training data is very high. In order to solve this problem, we propose an attention mechanism that can be trained on image-level labels data and roughly identifies text regions via an automatically learned attentional map based on a convolutional neural network. There are three main steps: firstly, a VGG model is trained using image-level labels data to score the likelihood that a text region exists in the picture; secondly, the region of interest is extracted by means of the attention mechanism and the extracted region is evaluated using the network trained in the first step to getting the text region and finally, the text line is extracted in the text region using the MSER algorithm. Trained with the weakly supervised data which is only with image-level labels, our model can generate bounding boxes for the text line in the image. The results of our model are very close to those of the models using bounding box label training data on the text detection benchmark sets of MSRA-TD500, ICDAR2013, and ICDAR2015.

Keywords: Weakly supervised · Text detection · Attention mechanism · MSER algorithm

1 Introduction

Text information in images is of indispensable value in semantic visual understanding. Reading the text in a natural scene, however, compared to traditional OCR, is still a challenging problem. With which application of deep neural networks, the performance of text detection has been improved rapidly. However, training most of the detection models in the current research requires a large number of images that are labeled with bounding boxes. In fact, the cost of obtaining these strong supervised images is very high. And it is even impossible to get a lot of strong supervised data in some special tasks. On the other hand, the acquisition cost of weakly supervised image-level labels data which only marks whether the text is in or not in the image is far from lower than that of strong supervised data. Besides, the research on weakly supervised text

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11901, pp. 406–417, 2019.
https://doi.org/10.1007/978-3-030-34120-6_33

detection is still scarce at present. So developing a text detection model that can be trained with weakly supervised data is very valuable.

In recent years, people have found that human cognition of things is not a one-time focus on the entire scene, but gradually draws attention to different regions of the scene while extracting relevant information [1,2]. By quickly scanning the global image, humans obtain the target area that needs to be focused on and then invest more attention resources in this target area to obtain more attention targets. At present, the attention-based models have achieved good results on many challenging tasks, such as machine translation [1,3], question and answer system [4,5], image description [2], and so on. In these tasks, the attention mechanism reflects the excellent key area positioning ability. So we hope to use this excellent positioning ability to achieve the task of weakly supervised text detection.

Because the shape of the text is generally fixed, the size is moderate, and the colors of the continuous text fields are very similar. These features of the text allow the positioning ability of the attention mechanism to be fully utilized to perform the detection tasks we expect. And the results of our experiments have proved that the text detection model based on the attention mechanism has a very good effect in weakly supervised text detection task.

In this paper, we designed an attention-based text detection model that can be trained by images with only image-level labels data. We first trained a VGG network to classify whether texts are in the images or not. Than attention mechanism is applied to extract the interest regions and we use the trained VGG network to get the text regions. Finally, the MSER algorithm is applied to generate the bounding boxes of the text regions. In summary, our contributions are in two folds:

(1) We propose a text detection model under weakly supervised data. This is mainly achieved by training an evaluation model using image-level label data to evaluate the localization performance of another localization model.
(2) We apply the attention mechanism to the task of weakly supervised text detection and design a novel loss function for the training of the text detection model with an attention mechanism.

The remainder of this paper is organized as follows: In Sect. 2, we briefly review the previous related work. In Sect. 3, we describe the proposed method in detail, including the construction of the evaluation network based on weakly supervised data, the application of attention mechanisms and location the text line from the text region. Experimental results are described in Sect. 4. Finally, conclusive remarks and future work are given in Sect. 5.

2 Related Work

Deep learning technologies have significantly advanced the performance of text detection in the past years [6–9]. These approaches essentially work in a sliding window fashion, with two key developments: (i) they leverage deep features, jointly learned with a classifier, to enable the strong representation of text; (ii) sharing a convolutional mechanism was applied for reducing the computational cost remarkably. With these two improvements, a number of Fully Convolutional Network [11] based approaches have been proposed [8–10]. They compute pixel-wise semantic estimations of text or non-text, resulting in a fast text detector able to explore rich regional context information.

Recently, some methods cast the previous character based detection into direct text region estimation, avoiding multiple bottom-up post-processing steps by taking word or text-line as a whole. Tian et al. [12] modified Faster-RCNN [13] by applying a recurrent structure on the convolution feature maps of the top layer horizontally. The algorithm proposed by He et al. [14] was inspired from single shot multi box detector [15]. They both explored the framework from generic objects and convert to scene text detection by adjusting the feature extraction process to this domain specific task. However, these methods are based on bounding boxes, which need to be carefully designed in order to fulfill the requirements for training.

Methods of direct regression for inclined bounding boxes, instead of offsets to fixed bounding boxes, have been proposed recently. EAST [16] designed a fully convolutional network structure that outputs a pixel-wise prediction map for text/non-text and five values for every point of text region, i.e., distances from the current point to the four edges with an inclined angle. He et al. [17] proposed a method to generate arbitrary quadrilaterals by calculating offsets between every point of text region and vertex coordinates.

However, almost all text detection models are based on bounding box labels, but the high cost of bounding box labels in some cases has no way to obtain suitable data. In this paper, we train a text detection model only through image-level labels, which solves the problem that the text detection model relies on bounding box labels.

3 Methodology

The entire methodology consists of three main parts. First, we build and train a VGG network for classification and feature extraction. Then, the construction of attention maps based on the results of the VGG network. Finally, the MSER algorithm was used to locate the text block with the attention map.

3.1 Text/Non-text Classification Network

In order to get a text-detection model with only image-level labels, we have to classify whether the texts exist in the image. In the first step of our method, a VGG network [18] is trained to distinguish whether there is text in the image. This network will provide a discriminator for subsequent models for positioning accuracy evaluation.

The network, which is denoted as Discriminant Network, is inspired by VGG network [18]. However, there are several improvements in our networks to distinguish whether there is text in the image compared with the original VGG network. Because most of the text appears in lines, the shape is more like a flat rectangle, we adjusted the scale of the convolution kernel in some layers, 5×5 convolution kernels that are more suitable for extracting text features from complex backgrounds, we also use 1×2 max-pooling layer as in [19], which reserves more information along the horizontal axis and benefits the detection of narrow shaped. In order to make the effects of 5×5 convolution kernels and 1×2 max-pooling layer clear, we have designed four networks with similar architecture and the details of these networks are shown in Table 1. The experiments in Sect. 4 also prove that our adjustments are very effective.

Table 1. Network configurations (shown in columns). The detailed differences are shown in the contents of this section.

Network A	Network B	Network C	Network D
Input: 512×512 RGB images			
Conv - 3×3 - 64			
Conv - 3×3 - 64			
MaxPooling - k: 2×2 - s: 2×2			
Conv - 3×3 - 128			
Conv - 3×3 - 128			
MaxPooling - k: 2×2 - s: 2×2			
Conv - 3×3 - 256			
Conv - 3×3 - 256			
MaxPooling - k: 2×2 - s: 2×2		MaxPooling - k: 1×2 - s: 1×2	
Conv - 3×3 - 512	Conv - 3×3 - 512	Conv - 3×3 - 512	Conv - 3×3 - 512
Conv - 3×3 - 512	Conv - 5×5 - 512	Conv - 3×3 - 512	Conv - 5×5 - 512
Conv - 3×3 - 512	Conv - 3×3 - 512	Conv - 3×3 - 512	Conv - 3×3 - 512
MaxPooling - k: 2×2 - s: 2×2		MaxPooling - k: 1×2 - s: 1×2	
Conv - 3×3 - 512	Conv - 3×3 - 512	Conv - 3×3 - 512	Conv - 3×3 - 512
Conv - 3×3 - 512	Conv - 5×5 - 512	Conv - 3×3 - 512	Conv - 5×5 - 512
Conv - 3×3 - 512	Conv - 3×3 - 512	Conv - 3×3 - 512	Conv - 3×3 - 512
MaxPooling - k: 2×2 - s: 2×2		MaxPooling - k: 1×2 - s: 1×2	
FC - 4096			
FC - 512			
FC - 2			
Softmax			

In Table 1, ‘Conv’ stands for Convolutional layers, with kernel size and output channels presented. The stride and padding for convolutional layers are all set to ‘1’. For Maxpooling layers, ‘k’ means kernel size, and ‘s’ represents stride.

It is worth mentioning that we are required to perform related operations on the input data of the model to make our model work. We will use the discriminator network in the subsequent steps to score the possibility of the presence of text in the picture, and the scored picture is generated by combining the processed picture with the attention map like Fig. 1(c), so we need to expand the input data by randomly smearing the training image make the picture more similar to the picture that combines the attention map. This method allows the discriminator to adapt to the pictures in the subsequent steps.

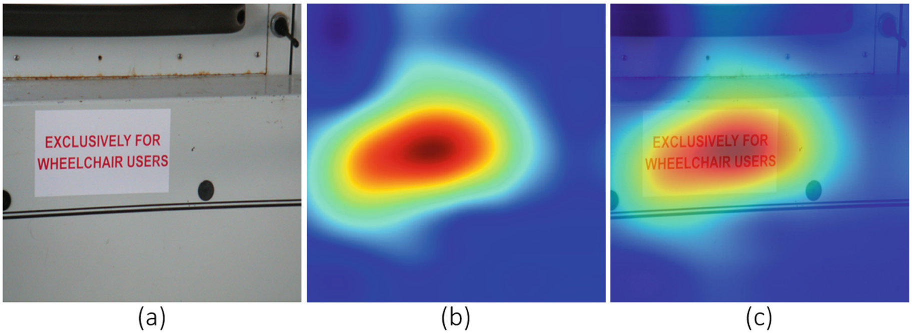

Fig. 1. The results of our method. (a) An input image; (b) The attention map, which is generated by the attention mechanism. Since the attention map is actually a matrix of the $[0, 1]$ interval, visualization is performed here; (c) Visualize the results of the combining the attention map with the input image.

3.2 Attention Mechanism

Our attention module is designed to learn rough spatial regions of text from the convolutional features automatically. It generates a pixel-wise probability heat map that indicates the text probability at each pixel location like Fig. 1(b). This probability heat map is referred to like the attention map which has an identical size of an input image and will be downsampled for each prediction layer.

We use the same convolution structure as that in Table 1, and inherit all the weight data from step one. This method of inheriting parameters is essentially a pretraining method that can make full use of the features of positioning that has been learned by convolutional neural networks.

The specific implementation of the attention mechanism is as follows: Denote by $L^s = \{l_1^s, l_2^s, \ldots, l_n^s\}$ the set of feature vectors extracted at a given convolutional layers s. Here, each l_i^s is the vector of output activations at the spatial location i of n total spatial locations in the layer. The global feature g is the feature of the fully connected layer of the previous model before the final output. We use the dot product between g and l_i^s as a measure of their compatibility:

$$c_i^s = \langle l_i^s, g \rangle, i \in \{1, \dots, n\} \tag{1}$$

In this case, the relative magnitude of the scores would depend on the alignment between g and l_i^s in the high-dimensional feature space and the strength of activation of l_i^s. For each of one or more layers s, the set of compatibility scores $C(\hat{L^s}, g) = \{c_1^s, c_2^s, \dots, c_n^s\}$, where $\hat{L^s}$, is the image of L^s under a linear mapping of the l_i^s to the dimensionality of g. The compatibility scores are then normalised by a softmax operation:

$$a_i^s = \frac{exp(c_i^s)}{\sum_j^n exp(c_j^s)}, i \in \{1, 2, \dots, n\} \tag{2}$$

The normalised compatibility scores $A^s = \{a_1^s, a_2^s, \dots, a_n^s\}$ is attention map which we will use it for location.

The attentional information is learned automatically in the training process, we can construct a loss function to describe whether the attention points are focusing on the right location. We have designed a new loss function by drawing on the idea of generative adversarial net [20].

$$Loss = -\lambda_{text} log(P_{text}) - \lambda_{non-text} log(1 - P_{non-text}) \tag{3}$$

λ_{text} and $\lambda_{non-text}$ is two constants used to adjust the attention mechanism, and we make it 0.5 in our model. P_{text} is a score for the possibility of text in the image which combined original image and attention map, we hope that this value is as large as possible. $P_{non-text}$ is similar to the definition of P_{text}, except that $P_{non-text}$ uses the attention map opposite to P_{text}. The meaning is the probability that there is text in the area that is not noticed, and we hope that this value is relatively small.

3.3 Bounding Boxes Generation

Although the text region detected by the attention map provides coarse localizations of text lines, they are still far from satisfactory. We borrowed from the methods in the paper [8] to further extract accurate bounding boxes of text lines.

At first, we extract the character components within the text blocks by MSER [21], since MSER is insensitive to variations in scales, orientations, positions, languages, and fonts. We use a constraint to reduce the wrong character components, the minimum area ratio of character components needs to be greater than 1% of the text region. After testing, in this way, most of the false components are excluded.

Then, we assume that text lines from the same text region have a substantially uniform spatial layout, and characters from one text line are in the arrangement of straight or near straight line. We use the statistical method to calculate the slope of the line passing through the most character components obtained by the MSER algorithm and record this slope as θ. In order to facilitate the calculation, the actual statistics are at intervals of $\frac{\pi}{24}$.

Finally, we combine the character components extracted by the MESR to text line candidate generation. We divide the components into groups. A pair of the components (A and B) within the text block α are grouped together if they satisfy the following conditions:

$$\frac{H(A)}{H(B)} \in \left[\frac{4}{5}, \frac{6}{5}\right] \tag{4}$$

$$O(A,B) - \theta \in \left[-\frac{\pi}{12}, \frac{\pi}{12}\right] \tag{5}$$

where $H(A)$ and $H(B)$ represent the heights of A and B, $O(A,B)$ represents the orientation of the pair.

For one group $\beta = \{c_i\}$, c_i is i-th character components, we draw a line l along the orientation θ passing the center of β. The point set ρ is defined as:

$$\rho = \{p_i\}, p_i \in l \cap \mathbb{B} \tag{6}$$

where $\mathbb{B}$ represents the boundary points of text region.

The minimum bounding box bb of β is computed as a text line candidate:

$$bb = \beta \cup \rho \tag{7}$$

where bb denotes the minimum bounding box that contains all points and components.

4 Experiments

Our methods are evaluated on three standard benchmarks, the MSRA-TD500 [22], ICDAR 2013 [23] and ICDAR 2015 [24]. The effectiveness of each proposed component is investigated by producing exploration studies. Full results are compared with the state-of-the-art performance on the three benchmarks.

4.1 Datasets

The following datasets are used in our experiments:

MSRA-TD500. The MSRA-TD500 dataset [22] is a multi-orientation text dataset including 300 training images and 200 testing images. The dataset contains text in two languages, namely Chinese and English. This dataset is very challenging due to the large variation in fonts, scales, colors and orientations. Here, we followed the evaluation protocol employed by [22], which considers both of the areas overlap ratios and the orientation differences between predictions and the ground truth.

ICDAR2013. The ICDAR2013 [23] consists of 229 training images and 233 testing images, with word-level annotations provided. It is the standard benchmark for evaluating near-horizontal text detection.

ICDAR2015. The ICDAR2015 [24] was collected by using Google Glass and it has 1,500 images in total: 1,000 images for training and the remained 500 images for testing. Different from the previous ICDAR competition, in which the text is well-captured, horizontal, and typically centered in images, these datasets focus on the incidental scene where text may appear in any orientation and any location with a small size or low resolution. This dataset is more challenging and has images with arbitrary orientations, motion blur, and low-resolution text. We evaluate our results based on the online evaluation system [24].

4.2 Implementation Details

The training of our model is divided into two steps: the first step is to train a classification network, and the second step is to train a regional extraction network based on the attention mechanism.

In the first step, we want to train a classification network to distinguish whether there is text in the image. However, the above three standard data sets are used for text detection, in which all pictures have the presence of text. So we use all the images of the above three datasets as a classification to represent images with text, and then randomly select the same number of images from the ImageNet [25] dataset as the classification without text. Our classification network is initialized by pretraining a model for ImageNet classification. The weights of the network are updated by using a learning rate of 10^{-3} for the first 100k iterations and 10^{-4} for the next 100k iterations, with a weight decay of 5×10^{-4} and a momentum of 0.9.

The results of our different networks are shown in Table 2. These results give strong evidence that the usage of 5×5 convolution kernels and 1×2 max-pooling layer can get better accuracies.

Table 2. Precision, Recall and F-measure of our networks in text/non-text classification task

Network	Precision	Recall	F-measure
A	0.81	0.86	0.83
B	0.83	0.85	0.84
C	0.86	0.89	0.87
D	**0.86**	**0.90**	**0.88**

It can be found that Network D works best, so in the subsequent steps we are all based on Network D. In the subsequent steps, we generate an attention map by combining the information of the fully connected layer with the information of the convolutional layer as described in Sect. 3.2. The weights of the network are updated by using a learning rate of 10^{-4} with a weight decay of 5×10^{-4} and a momentum of 0.9.

4.3 Experimental Results

MSRA-TD500. As shown in Table 3, although our method uses image-level labels, the results are still slightly better than other methods on MSRA-TD500 datasets. The proposed method achieves precision 0.81, recall 0.65 and f-measure 0.72. Compared to [26], our method obtains improvements on recall 0.02 and f-measure 0.01.

Table 3. Performance comparisons on the MSRA-TD500 dataset. The results are reported in the terms of Precision, Recall and F-measure

Method	Precision	Recall	F-measure
Yao et al. [22]	0.63	0.63	0.60
Yin et al. [27]	0.71	0.61	0.65
Kang et al. [28]	0.71	0.62	0.66
Yin et al. [26]	0.81	0.63	0.71
Proposed method	**0.81**	**0.65**	**0.72**

ICDAR2013. We also test our method on the ICDAR2013 dataset, which is the most popular for horizontal text detection. As shown in Table 4, the proposed method achieves 0.83, 0.73, 0.78 in precision, recall, and f-measure. Although the accuracy of our method is not as good as other state-of-the-art methods, it is worth emphasizing that our method is based on a weakly supervised data set. It is also very meaningful to train such a challenging result in a weakly supervised data set.

Table 4. Performance comparisons on the ICDAR2013 dataset.

Method	Precision	Recall	F-measure
SSD [15]	0.80	0.60	0.68
Yin et al. [26]	0.84	0.65	0.73
FASText [29]	0.84	0.69	0.77
TextBoxes [30]	0.86	**0.74**	0.80
Yin et al. [27]	0.88	0.66	0.76
CTPN [12]	**0.93**	0.73	**0.82**
Proposed method	0.82	0.71	0.76

ICDAR2015. As this dataset has been released recently for the competition in ICDAR2015, there is no literature to report the experimental result on it.

Table 5. Performance of different algorithms evaluated on the ICDAR2015 dataset. The comparison results are collected from ICDAR 2015 Competition on Robust Reading [24]

Method	Precision	Recall	F-measure
HUST_MCLAB [24]	0.44	0.38	0.41
AJOU [24]	0.47	0.47	0.47
Deep2Text-MO [24]	0.50	0.32	0.39
StradVision1 [24]	0.53	0.46	0.50
NJU-Text [24]	0.70	0.36	0.48
Yao et al. [10]	0.72	**0.57**	**0.64**
CTPN [12]	0.74	0.52	0.61
StradVision2 [24]	**0.77**	0.37	0.50
Proposed method	0.59	0.33	0.42

Therefore, we collect competition results [24] as listed in Table 5 for comprehensive comparisons. Our approach is less than ideal under this data set. After the analysis, we found that because the background of this data set is the most complicated and the size of the text is too small, the MSER algorithm will cause errors when extracting the bounding boxes from the attention area, affecting the final result.

The effectiveness and versatility of the proposed method can be proved by the above experiments. Besides the quantitative experimental results, several detection examples under various challenging cases of the proposed method on the MSRA-TD500 and ICDAR2013 datasets are shown in Fig. 2.

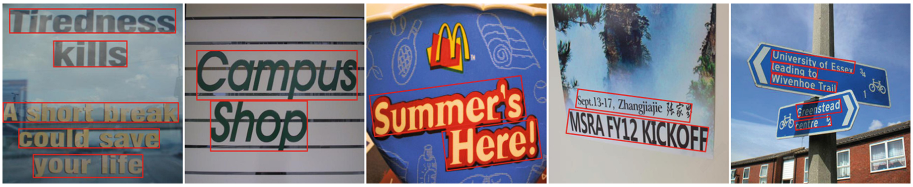

Fig. 2. Detection results by the proposed weakly supervised text detection model on the MSRA-TD500 and ICDAR2013 datasets.

5 Conclusion

In this paper, we have elaborately designed a model based convolutional neural network with an attention mechanism for weakly supervised text detection. In experiments, our model used image-level labels on MSRA-TD500, ICDAR2013, and ICDAR2015 datasets to compare the other state-of-the-art approaches which were trained in bounding box labels to show the validity and feasibility of our model for the text detection task. However, due to the limitations of the attention mechanism and the MSER algorithm. Our model does not have a beneficial effect on small text detection in complex backgrounds. This is also our future research direction.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Xu, K., Ba, J., Kiros, R., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
3. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
4. Ye, Y., Zhao, Z., Li, Y., et al.: Video question answering via attribute-augmented attention network learning. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 829–832 (2017)
5. Chen, Q., Hu, Q., Huang, J.X., et al.: Enhancing recurrent neural networks with positional attention for question answerin. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 993–996 (2017)
6. 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
7. Wang, T., Wu, D.J., Coates A., et al.: End-to-end text recognition with convolutional neural networks. In: Proceedings of the 21st International Conference on Pattern Recognition, pp. 3304–3308 (2012)
8. Zhang, Z., Zhang, C., Shen, W., et al.: Multi-oriented text detection with fully convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4159–4167 (2016)
9. He, T., Huang, W., Qiao, Y., et al.: Accurate text localization in natural image with cascaded convolutional text network. arXiv preprint arXiv:1603.09423 (2016)
10. Yao, C., Bai, X., Sang, N., et al.: Scene text detection via holistic, multi-channel prediction. arXiv preprint arXiv:1606.09002 (2016)
11. 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)
12. 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
13. Ren, S., He, K., Girshick, R., et al.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
14. He, P., Huang, W., He, T., et al.: Single shot text detector with regional attention. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3047–3055 (2017)

15. Liu, W., et al.: 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
16. Zhou, X., Yao, C., Wen, H., et al.: EAST: an efficient and accurate scene text detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5551–5560 (2017)
17. He, W., Zhang, X.Y., Yin, F., et al.: Deep direct regression for multi-oriented scene text detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 745–753 (2017)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017)
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
21. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3538–3545 (2012)
22. Yao, C., Bai, X., Liu, W., et al.: Detecting texts of arbitrary orientations in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1083–1090 (2012)
23. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., et al.: ICDAR 2013 competition on robust reading. In: 12th International Conference on Document Analysis and Recognition, pp. 1484–1493 (2013)
24. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., et al.: ICDAR 2015 competition on robust reading. In: 13th International Conference on Document Analysis and Recognition, pp. 1156–1160 (2015)
25. Deng, J., Dong, W., Socher, R., et al.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
26. Yin, X.C., Pei, W.Y., Zhang, J., et al.: Multi-orientation scene text detection with adaptive clustering. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1930–1937 (2015)
27. Yin, X.C., Yin, X., Huang, K., et al.: Robust text detection in natural scene images. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 970–983 (2014)
28. Kang, L., Li, Y., Doermann, D.: Orientation robust text line detection in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4034–4041 (2014)
29. 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)
30. Liao, M., Shi, B., Bai, X., et al.: Textboxes: a fast text detector with a single deep neural network. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)