Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2021.017181
Article

Adaptive Multi-Scale HyperNet with Bi-Direction Residual Attention Module for Scene Text Detection

Junjie Qu, Jin Liu[*] and Chao Yu

College of Information Engineering, Shanghai Maritime University, Shanghai, China
[*]Corresponding Author: Jin Liu. Email: jinliu@shmtu.edu.cn

Received: 23 January 2021; Accepted: 29 March 2021

Abstract: Scene text detection is an important step in the scene text reading system. There are still two problems during the existing text detection methods: (1) The small receptive of the convolutional layer in text detection is not sufficiently sensitive to the target area in the image; (2) The deep receptive of the convolutional layer in text detection lose a lot of spatial feature information. Therefore, detecting scene text remains a challenging issue. In this work, we design an effective text detector named Adaptive Multi-Scale HyperNet (AMSHN) to improve texts detection performance. Specifically, AMSHN enhances the sensitivity of target semantics in shallow features with a new attention mechanism to strengthen the region of interest in the image and weaken the region of no interest. In addition, it reduces the loss of spatial feature by fusing features on multiple paths, which significantly improves the detection performance of text. Experimental results on the Robust Reading Challenge on Reading Chinese Text on Signboard (ReCTS) dataset show that the proposed method has achieved the state-of-the-art results, which proves the ability of our detector on both particularity and universality applications.

Keywords: Deep learning; scene text detection; attention mechanism

1 Introduction

In the process of video subtitle recognition, the task of text detection in video image frame is particularly important [1–5]. Although most of the film and television works in the subtitle have strict standards, there are still many cases, the video subtitles will appear some personalized customization. This kind of situation includes text skew, overlap, distortion, blur, subtitle text size, position, font and low contrast with the background color [6–13]. Under the interference of the above factors, image text line detection is still a huge challenge, so it is still of great significance to improve the performance of text line detection model. Video subtitle location detection is essentially a process that regards subtitle text line as a target, and then detects the target in digital image, so most of the existing target detection neural network models can improve and adapt to the text line detection task.

Existing research results show that using the neural network model of target detection can achieve better results [14–22] when applied to text detection tasks at the text level. And according to the existing research results, as long as the depth of the network is deep enough, the effect of the model can be better.

However, in the existing target detection models based on deep learning neural networks, most of them are modeled on networks such as VGG16 and VGG19, through stacking several convolution layers serially into blocks as image feature extractors and training. Due to the insufficient sensitivity of the extraction of target information and non-target information in shallow features, the model can only propagate all extracted feature information forward to the deeper convolution layer of image feature extractor for semantic information judgment. However, the spatial information extracted from the deep convolution layer is largely lost, resulting in a large amount of deviation in the final judgment of the target text localization area. Therefore, simply increasing the depth of the convolution neural network cannot effectively improve the accuracy of text detection at the line level, and the problems of gradient disappearance and model convergence make the training of the model more difficult.

In this paper, we propose a positive and negative two-way residual attention mechanism, and propose a subtitle text detection model which combines Hyper-Net and MSFCN structure. The structure of full convolution Network, residual attention mechanism, area generation network is used for text line location detection. In the residual attention module, soft mask branch is added with positive and negative mask images. The results are used to enhance the features of text objects and suppress the features of non text objects. Experiments show that the network has a great improvement in performance and running speed, and has a good effect on the Chinese video subtitle image data set. The major contributions of our research can be summarized as follows:

(1) We proposed a novel BRAM method that can effectively enhance the sensitivity of target semantics in shallow features.

(2) We proposed a novel AMSHN method that can significantly reduce the loss of spatial features due to downsampling and locate the target text area accurately.

(3) Our method achieved state-of-the-art performance on ReCTS dataset.

2 Related Work

According to the research object, the existing text detection research is divided into two categories: text line detection and character segmentation. Because the text line in the image has similar properties to other objects, the image object detection algorithm based on deep learning is widely used in the research task of text line detection.

In 2016, Zhang et al. [23] proposed an unconventional multi-directional text line detection framework model. The basic idea of the model is to integrate local clues and global clues in text blocks through coarse to fine feature extraction strategy. The image semantic segmentation result is generated by using full convolution network, and the result is divided into several candidate text blocks. Then the character is extracted from the text block, and the direction of the character is estimated by the projection of the character component. Finally, all the characters are merged into text lines again and the coordinates of the bounding box are calculated.

In 2017, Ma et al. [24] proposed another text line detection model for any direction. They proposed a Rotate Region Proposal Network (RRPN) based on the region generation network. The parameter of rotation angle was added in the region generation parameters, which made the network generate inclined candidate bounding boxes, in addition, they proposed a rotating region of interests pooling (RROI pooling) to map the candidate bounding boxes of RRPN network to the feature graph for pooling.

In 2017, Zhou et al. [25] proposed an efficient and accurate natural scene text line detector (EAST). The core part of the algorithm is a full convolution neural network, which outputs the probability that each pixel belongs to the text box, and the distance relative to the four boundaries of the bounding box, The rotation angle also has pixel offset values of four vertex coordinates to predict the existence of text instances and their geometric shapes. In this way, the intermediate steps such as candidate recommendation, text region formation and word division are omitted, and only the post-processing steps such as thresholding of predicted geometry and non maximum suppression algorithm are retained, so as to achieve efficient text line recognition.

3 Method

In order to make the text detection model sensitive to the target semantic information in the shallow features of the feature extractor, we introduce a computer vision attention mechanism. This mechanism adds a new softened mask branch to the shallow layer feature information in the original stacked serial convolution network model. Through a coder and decoder structure in the branch, we classify and understand the basic semantic information of the shallow features, and then generate the corresponding mask to enhance the eigenvalues in the target area of the main features. At the same time, in order to further weaken the impact of non target features, we add a negative soft mask branch, and also analyze the semantics of non target information in shallow features through an encoder and decoder structure, and partially weaken the eigenvalues of non target areas. In addition, in order to solve the problem that the spatial information in the features extracted from the deep convolution layer is largely lost, we borrowed the ideas of HyperNet, FPN and our previous research on MSFCN, and fused the shallow features with the deep features after scaling, by scaling the feature map obtained from the first three layers of convolution layer and the last three layers of convolution layer, it is transformed into a unified scale and concatenated to enhance semantic features while preserving morphological feature information.

The formal expression of the text line detection task is as follows, assuming that the input accepted by the text line detection model is an RGB image with a height of H and a width of W:

$$F_{in} = \{p_{x,y,c} \mid x \in [1,W], y \in [1,H], c \in [1,3]\} \tag{1}$$

where represents the pixel value of $P_{x,y,c}$ at (x,y) in the channel No. C. The final output of the model is:

$$F_{out} = ((x_1, y_1, w_1, h_1), (x_2, y_2, w_2, h_2), \ldots, (x_i, y_i, w_i, h_i)) \tag{2}$$

where x, y, w, h represent the abscissa, ordinate, width and length of the final output text line positioning box. Each set of arrays containing these four values can be uniquely represented as a single text box. The ultimate goal of the text detection task is to find a mapping H, which can match the input of the image with the output of the text box one by one:

$$F_{out} = H(F_{in}) \tag{3}$$

In order to increase the sensitivity of the network model to the target in the image, this paper introduces a new residual attention module in the original image feature extraction module, which uses the encoder decoder structure to obtain the location of the region of interest in the feature image and generate a mask to enhance the intermediate features. The overall structure of the text line detection model based on the attention mechanism of positive and negative two-way residuals is shown in the Fig. 1. The overall structure is divided into three parts: image feature extraction, region generation network and feature classification region regression network.

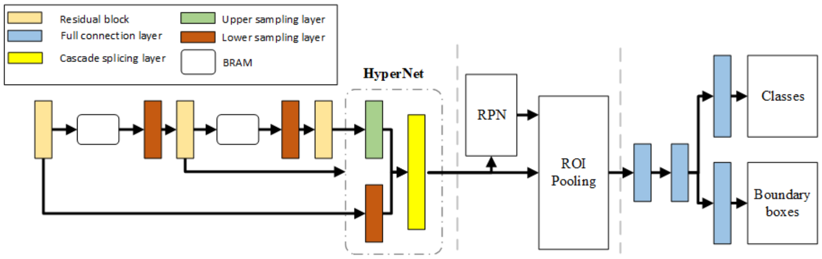

Figure 1: The overall structure of the text line detection model based on the attention mechanism of positive and negative two-way residuals

3.1 Feature Extraction Module

In this paper, we introduce a Bi-direction Residual Attention Module (BRAM) to strengthen the regions of interest in the image and weaken the regions of non-interest, so that we can improve the performance of the model in extracting the features of the text area without further increasing the depth of the model. The definition of the attention mechanism module is as follows:

$$H_{i,c}(x) = F_{i,c}(x) + PA_{i,c}(x) * F_{i,c}(x) - NA_{i,c}(x) * F_{i,c}(x) \tag{4}$$

$$H_{i,c}(x) = (1 + PA_{i,c}(x) - NA_{i,c}(x)) * F_{i,c}(x) \tag{5}$$

where x represents the feature input of the previous network layer, $H_{i,c}(x)$ represents the mapping relationship between the two network layers corresponding to the attention mechanism module, $F_{i,c}(x)$ represents the mapping relationship between the backbone branches, $PA_{i,c}(x)$ and $NA_{i,c}(x)$ respectively represents positive and negative bidirectional soft mask branch.

We also fuse the adaptive multi-scale feature representation structure in the multi-scale feature fusion network MSFCN structure based on the HyperNet and proposes a new Adaptive Multi-Scale Hyper Net (AMSHN), which increases the model's ability to extract semantic feature information while retaining the spatial feature information in the shallow features.

4 Experiment

We carried out our experiments on tReCTS datasets. The testing environment was conducted on one single Nvidia GTX 2080Ti graphic card with 16 GB memory, and an Intel(R) Core(TM) i7-7700 3.60 GHz. The ablation experiments of the model are shown in Tab. 1.

Table 1: The ablation experiments of the model

p	*t*	IoU = 0.3/0.7		
		P	*R*	*F*
2	1	0.90	0.80	0.84
2	2	0.90	0.80	0.84
3	1	0.91	0.80	0.84
3	2	0.92	0.80	0.85
4	1	0.92	0.80	0.85
4	2	0.92	0.80	0.85
5	1	0.93	0.81	0.86

As can be seen from the above table, we can find that the model's performance always increases with the number of residual blocks. However, after the hyperparameter *p* exceeds 3, the rising speed of the accuracy of the model prediction slows down, and the role of the hyperparameter *t* is almost dispensable. We know that with the increase of the parameter amount, the training difficulty and prediction time of the model will increase. Therefore, considering the above two factors, we finally choose the hyperparameter configuration of *p* = 3, *t* = 2 to carry out the default parameters of the performance comparison experiment of the final model.

We used the above hyperparameters to perform 300 rounds of iterative training on the model, recorded the four results of the RPN classifier loss value, RPN regression loss value, detector classifier loss value, and detector regression loss value in each iteration result. Plotted as a line chart, the results are shown in the figure below.

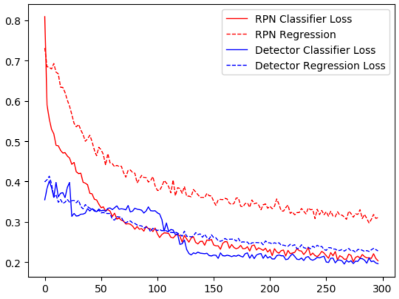

Figure 2: The result of training loss

The comparison of detection results is shown in Fig. 3.

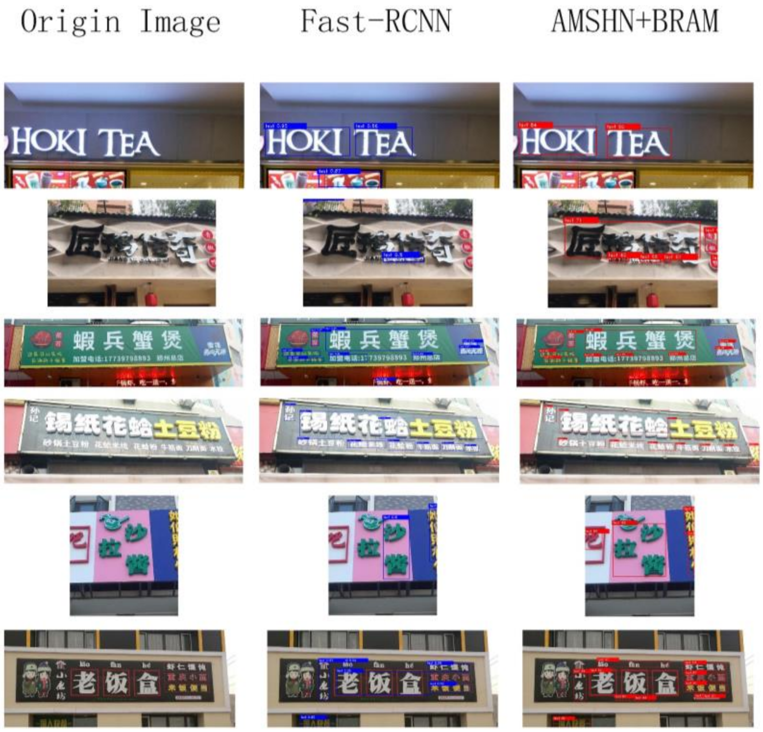

Figure 3: Comparison of detection results

Tab. 2 is the data display of the best performance that each method can achieve on our self-made data set. We finally selected the six best open source methods in the performance of the data set for comparison. We can see that our method perform better in the existing publicly published methods. The accuracy rate, recall rate, and *F* value of the model are higher than the existing methods, reaching the current best level.

Table 2: Comparison with other six detection methods

Method	*P*	*R*	*F*
Fast-RCNN	0.87	0.75	0.80
TH-TextLoc	0.80	0.73	0.77
TextFlow	0.86	0.76	0.81
VOCR	0.83	0.80	0.82
Text-CNN	0.91	0.74	0.82
CTPN	0.89	0.79	0.84
AMSHN + RPN	0.91	0.80	0.85

5 Conclusions

In this paper, we proposed an effective scene text detector for subtitle text detection. The proposed method aims at designing a fast and accurate scene text detector, in which a feature extraction module is added to enhance the sensitivity of target semantics in shallow features and reduce the loss of spatial features due to downsampling. Meanwhile, a Region Proposal Net and a regression module are designed to generate a certain number of feature candidate box areas. Finally, the outputs are sent to a Classifier

Net to produce the text boxes. Experimental result on the ReCTS show that the proposed method has achieved the state-of-the-art results, which proves the ability of our detector on both particularity and universality application. In the future, we will combine the detector with a subtitle text recognizer to solve the text detection problem caused by partial uneven illumination and partial specular reflection.

Acknowledgement: This work is supported by the National Natural Science Foundation of China (61872231, 61701297).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Tian, W. L. Huang, T. He, P. He and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," *European Conference on Computer Vision*, pp. 56, 2016.

[2] X. Y. Zhou, C. Yao, H. Wen, Y. Z. Wang, S. C. Zhou *et al.,* "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 2642–2651, 2017.

[3] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong *et al.*, "ICDAR 2003 robust reading competitions," in *7th Int. Conf. on Document Analysis and Recognition*, pp. 682, 2003.

[4] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura and L. P. D. L Heras, "ICDAR 2013 robust reading competition," in *12th Int. Conf. on Document Analysis and Recognition*, pp. 1484, 2013.

[5] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov *et al.*, "ICDAR 2015 competition on robust reading," in *13th Int. Conf. on Document Analysis and Recognition*, pp. 1156, 2015.

[6] M. Liao, B. Shi, X. Bai, X. Wang and W. Liu, "Text Boxes: A fast text detector with a single deep neural network," in *Thirty-First AAAI Conf. on Artificial Intelligence*, pp. 4161–4167, 2017.

[7] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao *et al.,* "Single shot text detector with regional attention," in *Proc. IEEE Int. Conf. Computer Vision (ICCV),* pp. 3047–3055, 2017.

[8] W. He, X. Y. Zhang, F. Yin and C. L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV),* pp. 745–753, 2017.

[9] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han *et al.,* "WordSup: Exploiting word annotations for character based text detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV),* pp. 4950–4959, 2017.

[10] J. Q. Ma, W. Y. Shao, H. Ye, L. Wang, H. Wang *et al.,* "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[11] M. Liao, B. Hi and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3676–3690, 2018.

[12] M. Liao, Z. Zhu, B. Shi, G. S. Xia and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 5909–5918, 2018.

[13] F. Wang, L. Zhao, X. Li, X. Wang and D. Tao, "Geometry-aware scene text detection with instance transformation network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 1381–1389, 2018.

[14] B. Shi, X. Bai and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 2550–2558, 2017.

[15] P. Lyu, C. Yao, W. Wu, S. Yan and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 7553–7563, 2018.

[16] Z. Liu, G. Lin, S. Yang, J. Feng, W. Lin *et al.,* "Learning Markov clustering networks for scene text detection," in *Proc. IEEE Intl. Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 6936–6944, 2018.

[17] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu *et al.,* "Multioriented text detection with fully convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 4159–4167, 2016.

[18] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou *et al., Scene Text Detection via Holistic, Multi-Channel Prediction.* 2016. [Online]. Available: https://arxiv.org/abs/1606.09002.

[19] Y. Wu and P. Natarajan, “Self-organized text detection with minimal post-processing via border learning,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 5010–5019, 2017.

[20] D. F. He, X. Yang, C. Liang, Z. H. Zhou, A. G. Ororbia *et al.*, “Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 474–483, 2017.

[21] C. K. Ch’ng and C. S. Chan, “Total-text: A comprehensive dataset for scene text detection and recognition,” in *Proc. 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR)*, vol. 1, pp. 935–942, 2017.

[22] C. Xue, S. Lu and F. Zhan, “Accurate scene text detection through border semantics awareness and bootstrapping,” in *Proc. European. Conf. Computer Vision*, 2018, pp. 355–372.

[23] J. Q. Ma, W. Y. Shao, H. Ye, L. Wang, H. Wang *et al.*, “Arbitrary-oriented scene text detection via rotation proposals,” *IEEE Transactions on Multimedia*, vol. 99, pp. 1, 2017.

[24] X. Y. Zhou, C. Yao, H. Wen, Y. Z. Wang, S. C. Zhou *et al.*, “EAST: An efficient and accurate scene text detector,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5551–5560, 2017.

[25] Z. Zheng, C. Zhang, S. Wei, Y. Cong and B. Xiang, “Multi-oriented text detection with fully convolutional networks,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4159–4167, 2016.