EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

PSENet-based efficient scene text detection

Guanglong Liao, Zhongjie Zhu*, Yongqiang Bai*, Tingna Liu and Zhibo Xie

*Correspondence:
zhongjiezhu@hotmail.com;
byq-163@163.com
Ningbo Key Lab of Digital Signal Processing,
Zhejiang Wanli University,
Ningbo 315100, China

Abstract

Text detection is a key technique and plays an important role in computer vision applications, but efficient and precise text detection is still challenging. In this paper, an efficient scene text detection scheme is proposed based on the Progressive Scale Expansion Network (PSENet). A Mixed Pooling Module (MPM) is designed to effectively capture the dependence of text information at different distances, where different pooling operations are employed to better extract information of text shape. The backbone network is optimized by combining two extensions of the Residual Network (ResNet), i.e., ResNeXt and Res2Net, to enhance feature extraction effectiveness. Experimental results show that the precision of our scheme is improved more than by 5% compared with the original PSENet.

Keywords: Scene text detection, PSENet, Mixed Pooling Module

1 Introduction

As a key technique, text detection plays an important role in computer vision applications, which has been of great value to the description and understanding of scene contents. Hence, Text detection has been obtained the broad application, such as robot navigation, image search, industrial automation [1–4]. Due to the large variation in text patterns and background, detecting text in scene is still challenging.

In recent years, scene text detection algorithms can be roughly divided into the following three categories: regression-based algorithms, corner detection-based algorithms, and segmentation-based algorithms. Classical regression-based algorithms include the real-time object detection with region proposal networks [5], single-shot multi-box detector [6], you only look once [7], and DenseBox [8]. Subsequently, Liao et al. proposed a novel algorithm named TextBoxes, which modified the length-width ratio of the pre-selected box in the text area and adjusted the shape of the convolution kernel in the feature layer [9]. In order to detect text boxes in any direction, the authors further modified anchor boxes and convolution kernels [10]. Similarly, Zhou et al. directly predicted the score map, rotation angle and text box of each pixel by adopting the Full Convolution Network (FCN) to detect text in any direction [11]. Based on real-time object detection with region proposal networks, Ma et al. introduced a series of rotated candidate frames to solve the problem of rotating text, but the detection speed decreases [12]. In conclusion, these regression-based algorithms are easily limited by the text direction and text length–width ratio, and then the detection precision needs to be further improved. For the corner detection-based algorithms, Tychsen-Smith et al. adopted corner detection and directed sparse sampling in DENet, to replace the region proposal network portion of the RCNN model [13]. Pengyuan et al. detected the scene text by locating the corner of the text box and dividing the text area to the relative position [14]. Wang et al. combined the corner points, center points and regression points of the boundary of the text box into a network through a fully convolutional network [15]. The algorithm can solve the problem of documents inclination to a great extent, but the precision was reduced because the corner points were difficult to determine. For segmentation-based algorithms, Deng et al. proposed the PixelLink algorithm to solve the segmentation problem of adjacent text areas, which can predict the pixel links between different text blocks by adopting text secondary prediction and link secondary prediction algorithm [16]. Zhang et al. adopted the maximum stable extremum regions method to detect candidate characters from the extracted text regions, and divided the characters into words or text lines according to prior rules [17]. Xie et al. proposed a supervised pyramid context network, which introduced the power segmentation framework of Mask RCNN, and used context information to detect arbitrary shape text [18]. The Text Context Module and Re-Score Module proposed by the algorithm can effectively restrain false sample detection. Recently, Long et al. proposed the TextSnake algorithm based on semantic segmentation to solve the problem of curved text for the first time, by introducing a disc and the text centerline [19]. Although above-mentioned segmentation-based algorithms solve the problem of curved text, it is difficult to distinguish adjacent or overlapping texts accurately. Most of the previous scene text detection algorithms use multi-feature prediction methods such as feature pyramid networks or spatial pooling to solve complex scene problems. Among them, the traditional spatial aggregation is not well suited to the task of pixel-level prediction based on semantic segmentation due to the limitation that they all probe the input features map within square windows. Therefore, this paper proposes an efficient text detection method based on Mixed Pooling Module (MPM), which considers not only the regular shape of N × N, but also the long but narrow kernel, i.e., 1 × N or N × 1. The MPM with different pooling operations is designed to locate the scene text regions precisely.

Compared with above-mentioned segmentation-based algorithms, Wenhai et al. adopted Feature Pyramid Network (FPN) to convert images into feature maps of different scales, and then fuse features from different scales to achieve multi-scale prediction. In addition, the progressive scale expansion algorithm is adopted to solve the segmentation problem of adjacent text [20]. However, the single maximum pooling operation of PSENet resulted in the loss of some neighborhood feature information, which made it impossible to adapt to the text length of different scenes and capture the long and short features of the scene text object. To prevent the loss of adjacent feature information, various methods such as pyramid pooling and dilation convolution have been proposed in recent years [20, 21]. Enough experiments have proved that spatial pooling is an effective means to capture remote context information and perform pixel-level prediction tasks well. However, the above methods all use square windows to obtain input feature information. Because the square window can't adapt to the characteristics of text, it can't capture the anisotropic context information widely existing in real scenes flexibly. Scene text have no obvious boundary, and the characters in the text image have their own characteristics, such as English and Chinese, which are generally long, with horizontally dense and vertically sparse representations.

To sum up, the accurate location of text region relies not only on the large-scale saliency information, but also on the small-scale boundary feature. Hence, this paper presents an efficient scene text detection scheme based on the PSENet from two aspects as follows. Firstly, a Mixed Pooling Module (MPM) is designed to locate the scene text regions precisely. Then, the backbone network is optimized to enhance the effectiveness of multi-scale feature extraction. Specifically, the main contributions of this paper can be summarized as follows:

1. The MPM is designed with different pooling operations. Generally, the single pooling operation is adopted in the original PSENet. Its complexity is low, but for adjacent texts, it is easy to cause the problems of missed detection and false detection. Considering the diversity of text length in different scenes, the MPM with different pooling operations is designed to locate the scene text regions precisely. Through the context information extraction at different distances, we can capture the relevance of text information more effectively, and then extract the text shape features and locate the scene text regions.
2. The backbone network is optimized by the combination of ResNeXt and Res2Net. In original PSENet, the Residual Network (ResNet) is adopted as the backbone network, to solve the problem of gradient vanishing and further increase the network depth easily. But the ResNet cannot enhance the extraction capability of multi-scale features, which are very important for scene text detection. Therefore, the backbone network is optimized here by combining ResNeXt and Res2Net, to fuse different types of deep feature information and characterize the text features at a more fine-grained level, and further to improve the text detection precision.

The remainder of the paper is organized as follows: Sect. 2 describes the proposed scheme in detail. Experimental results are shown and analyzed in Sect. 3. Section 4 discusses the paper and Sect. 5 concludes the paper.

2 Proposed scheme

In order to achieve efficient and precise text detection, a scene text detection scheme based on PSENet is proposed. By designing the Mixed Pooling Module (MPM) and optimizing the backbone network, a detection model for scene text is trained, which can detect more accurately than the original PSENet model in the scene text. The diagram of the proposed scheme is shown in Fig. 1.

As shown in Fig. 2a, ResNet is adapted to the original PSENet backbone network to realize basic image feature extraction. The short connections are introduced to ResNet, which alleviated the problem of gradient vanishing disappearance and obtained deeper network structures at the same time. However, during the process of feature extraction, only four equivalent feature scales can be obtained in ResNet through different combinations of convolution operations. Multi-scale feature representations of backbone network are very important for vision tasks, because an effective backbone network needs to locate objects of different scales in the scene. Therefore, the backbone network is

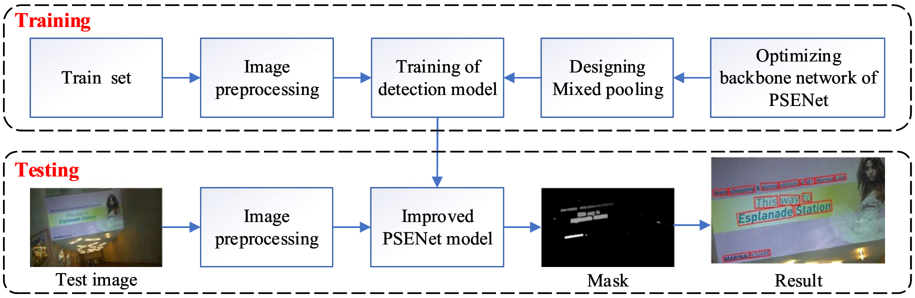

Fig. 1 Schematic diagram of the scheme

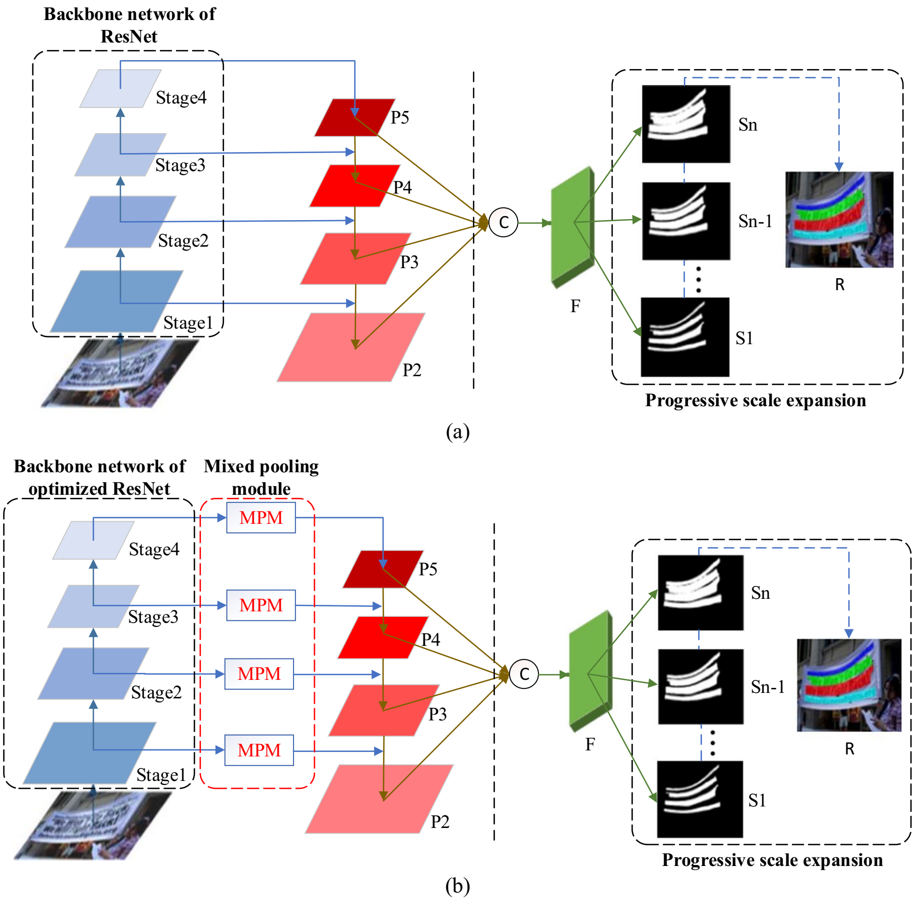

Fig. 2 Comparison between the original PSENet and the proposed scheme: **a** is the original PSENet; **b** is the proposed scheme

optimized by combining ResNeXt and Res2Net to enhance the feature extraction effectiveness shown in Fig. 2b. In addition, the MPM is embedded into the backbone network to capture the correlation between long distance and short distances between different locations. At the same time, the backbone network can extract the shape of the scene text better to improve the precision of text detection in the model.

2.1 Mixed pooling module

As shown in Fig. 3, with different pooling kernel sizes of MPM is designed to adapt to the characteristics of image text and avoid the noise impact caused by rectangular filtering. The MPM can be better adapted to the scene text features by different pooling kernel sizes and different dimensional pooling operations. As shown in Fig. 3a, the one-dimensional pooling operations in the MPM can effectively enhance the perceptual wildness of the backbone network in horizontal or vertical directions and further improve its long-range dependencies at the high-level semantic level. The pooling operations with square pooling kernels can enable the model to capture a large range of contextual information. In general, more local contextual information can be obtained in complex text scenarios by different pooling kernel shape operations. Figure 3a shows a simplified MPM diagram, and Fig. 3b provides detailed information about the design process of MPM. As shown in Fig. 3b, the MPM has two one-dimensional pooling layers, which can better adapt to the text features of the text image. Then, there are two spatial pooling layers and one original spatial information preserving layer, which can effectively capture the context information of dense text areas. Notice that the bin sizes of feature maps after each spatial pooling are 18 × 18 and 10 × 10, and the bin sizes after each one-dimensional pooling are N × 1 and 1 × N. Finally, all five sub-paths are combined by summing.

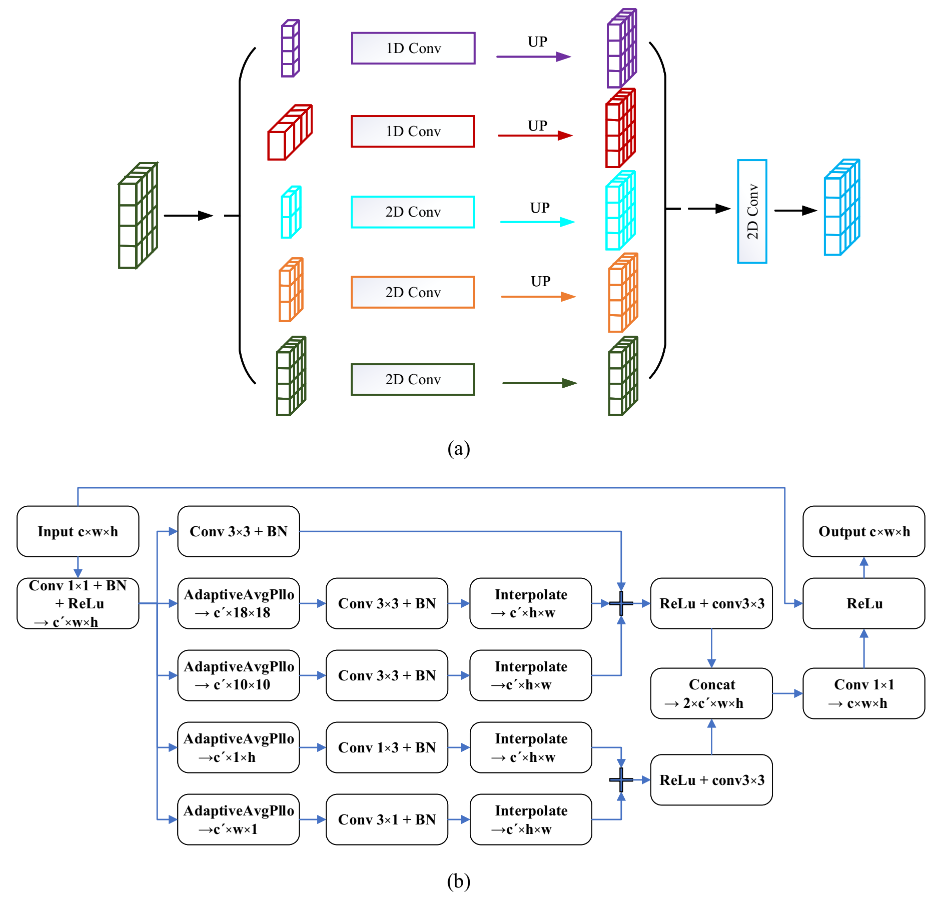

Fig. 3 The designed Mixed Pooling Module (MPM): **a** is the simplified MPM diagram; **b** is the detailed MPM structure

These MPMs are then embedded into the PSENet backbone network. It is noticeable that since the output of the backbone has 2,048 channels, the 1 × 1 convolutional layer is connected to the backbone first, to reduce the output channels from 2,048 to 1,024, and then two MPMs are embedded.

2.2 Optimizing the backbone network

As shown in Fig. 4, ResNet is optimized by combining two extensions of the residual network (ResNet), i.e., ResNeXt [22] and Res2Net [23]. Among them, the essence of ResNeXt is group convolution, which is composed of ResNet [24] and Inception [25], effectively reducing the number of parameters. The Res2Net increases the range of receptive fields of each network layer and improves the ability of multi-scale feature extraction by constructing hierarchical residual-like connections within one single residual block. Specifically: (1) the optimized ResNet is built in a similar way to the group convolution, where the number of groups is controlled by variable cardinality and the variable cardinality is set to 32. Compared with ResNet, the optimized ResNet can greatly reduce the amount of computation by controlling the variable cardinality, and thus improving running speed. The conclusion is proven in subsequent experiments. (2) the ResNet is optimized based on the idea of the Res2Net, adding a small residual block to each residual module in the ResNet. The attention mechanism module of Res2Net is removed and smaller groups of filters is adopted to replace a group of 3 × 3 filters, while connecting different filter groups in a hierarchical residual-like style. Different filter groups are connected in a similar residual layering way to obtain more feature information of different scales and effectively improve the model performance. Specifically, after a 1 × 1 convolution, the feature map is divided into s subsets, which are denoted

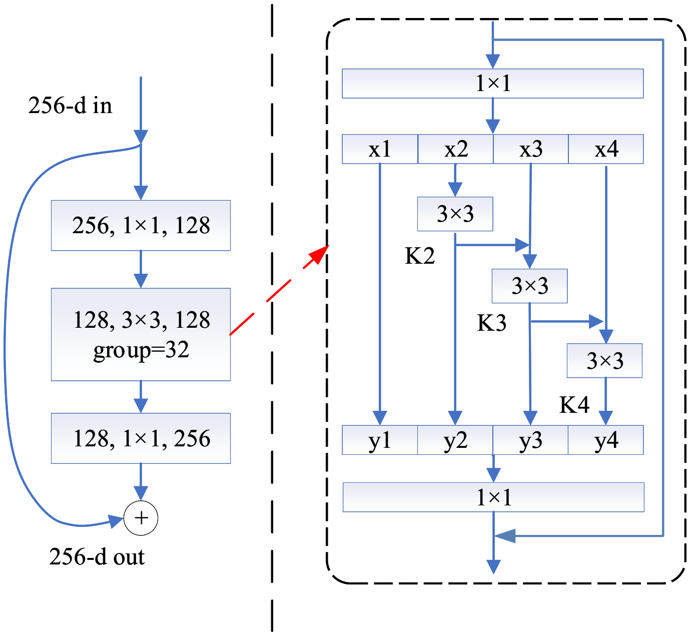

Fig. 4 The ResNet is optimized by combined ResNeXt and Res2Net (Variable cardinality $C = 32$, scale dimension $s = 4$)

as x_i, where $i \in \{1,2,\ldots,s\}$. The number of channel of each feature subset x_i is equal to $1/s$ of the input feature map, and their spatial size are the same. Except for x_1, every x_i has a corresponding 3×3 convolution, which is denoted by K_i (). Y_i is used to express the output of K_i () and the input of K_i () is the sum of the outputs of feature subset x_i and K_{i-1} () minus 1. And omitting a 3×3 convolution at x_1 is to increase s and reduce the number of parameters. Therefore, Y_i can be written as follows:

$$Y_i = \begin{cases} x_i & i = 1; \\ K_i & i = 2; \\ K_i(x_i + Y_i - 1) & 2 < i \leq s. \end{cases} \tag{1}$$

Note that the Res2Net performs multi-scale processing, and fuses different scale information through a 1×1 convolution, thus effectively processing the feature information. The optimized backbone network in this paper is conducive to extracting global and local information, and effectively improves the network's feature extraction ability, improving the text detection accuracy of the model.

3 Experimental results comparison and analysis

To evaluate the performance of the proposed scheme, experiments are conducted on ICDAR2015 dataset and ICDAR2017-MLT dataset. The precision, recall and F-measure are used for evaluation. The experimental development environment is as follows: CPU: i7-8700 3.20 GHz, RAM: 16 GB, GPU: NVIDIA GeForce GTX1060Ti 6 GB, and deep learning network framework: PyTorch.

When the proposed scheme is trained on ICDAR2015 dataset, 1,000 training ICDAR2015 images and 500 ICDAR2015 verification images are used to train the model. The batch is set to 2, and 600 epochs are performed on a single GPU. The initial learning rate is 1×10^{-4} and the final learning rate is 1×10^{-7}. Between the 200 epochs and 400 epochs, the attenuation rate is set to 5×10^{-4}. In the above training process, the loss balance is set to 0.7, the online hard example mining to 3, the kernel size to 0.5, the number of kernels to 6, and the aspect ratio of the input image to the output image to 1 [20]. The affine transform is adopted to process the training data. The details are given below. (1) the pictures and text boxes are randomly scaled according to the ratio of {0.5,1,2,3}. (2) random rotation is within the amplitude range of [− 10°,10°]. (3) the picture is flipped horizontally and vertically. (4) gaussian additive noise is added to the image.

3.1 Performance comparison of mixed pooling

To evaluate the MPM performance, experiments are conducted, where the ResNet and ResNet-MPM are used for comparison on the ICDAR2015 dataset. As shown in Table 1,

Table 1 Performance comparison of the MPM on the backbone network. "MPM" means adding mixed pooling to the ResNet network, and bold indicates that it ranks first in related performance

Backbone	Precision (%)	Recall (%)	F-Measure (%)	FPS	Model scale (MB)
ResNet	81.13	77.03	79.03	1.61	342.1
ResNet-MPM	**86.32**	**80.55**	**83.34**	**1.65**	343.9

when the MPM is embedded into the PSENet backbone network, the network performance has been greatly improved. The precision of model detection has increased by more than 5%, and the recall and F-measure have also increased, but the scale of the network model has not increased. The experimental results demonstrate that the MPM can not only significantly improve the performance of the model, but also hardly need additional model parameters.

3.2 Performance comparison of the backbone network

To further verify the performance of optimized PSENet backbone network, relevant experiments are conducted on the basis of the previous experimental steps. Compared with the original PSENet model, the proposed scheme training model has made some progress in precision, recall, F-measure, frames per second (FPS) and model scale shown in Table 2.

In addition, experiments are conducted to compare the performance between optimized ResNet-MPM-50 and ResNet-152 shown in Fig. 5. As shown in Fig. 5a, when the number of epochs reaches 540, the ResNet-152 based model has an 85.51% on precision, an 80.69% on recall, an 83.03% on F-measure, and 0.3720 on loss. Figure 5b shows the optimized ResNet-MPM-50 based model reaches the optimal value when the number of epochs reaches 350, with the precision of 87.26%, the recall of 80.79%, the F-measure of 84.00%, and the loss of 0.3275. Experimental results demonstrate that the performance of the proposed scheme training model is even better than that of the ResNet-152 based model. Moreover, the scale of the proposed scheme training model is about half of that based on ResNet-152. It is worth noting that the proposed scheme can effectively improve the performance of model without deepening the network.

3.3 Comparison with classical scene text detection algorithms

To evaluate the performance of the proposed scheme, experiments are conducted on ICDAR2015 dataset, in which the original PSENet is compared with several other classical scene text detection algorithms. As shown in Table 3, the detection precision of the proposed scheme has a 5.76% improvement on precision, 0.24 on FPS compared with the original PSENet. It is worth noting that the proposed scheme has the highest detection precision among the following scene text detection algorithms. In order to evaluate the performance of the proposed scheme in multi-directional text, experiments are conducted on ICDAR2017-MLT dataset. As shown in Table 4, the precision, recall and F-measure of the proposed scheme are 77.67%, 69.98% and 73.83%, respectively, and the precision is higher than that of the original PSENet by more than 3%.

Table 2 Performance comparison of the optimized backbone network. Bold indicates that the performance ranks first

Backbone	Precision (%)	Recall (%)	F-measure (%)	FPS	Model Scale (MB)
ResNet	81.13	77.03	79.03	1.61	343.90
ResNet-152	85.51	80.69	83.03	1.38	760.20
Optimized ResNet with MPM	87.26	80.79	84.00	1.84	333.40

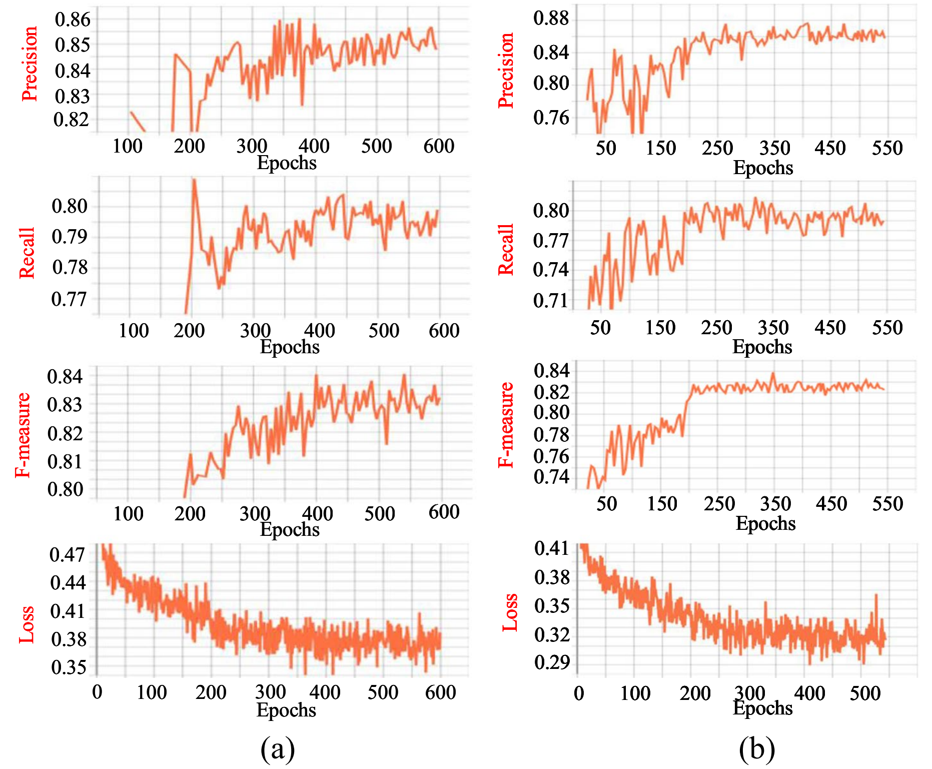

Fig. 5 Performance comparison between ResNet-152 and the optimized ResNet-MPM-50: **a** is PSENet backbone network of ResNet-152; **b** is PSENet backbone network of the optimized ResNet-MPM-50

Table 3 Performance of test results on ICDAR2015, the top three related performance rankings are in bold. "–" indicates that the reference information in relevant references is incomplete

Algorithms	**ICDAR2015**			
	Precision (%)	**Recall (%)**	**F-measure (%)**	**FPS**
CTPN [26]	74.20	51.60	60.90	7.10
SegLink [27]	73.10	76.80	75.00	–
WordSup [28]	79.30	77.00	78.20	–
EAST [11]	83.60	73.50	78.20	13.20
RRPN [12]	82.00	73.00	77.00	–
PixelLink [16]	82.90	81.70	82.30	7.30
TextSnake [19]	84.90	80.40	82.60	1.10
PSENet	81.50	79.70	80.60	1.60
Proposed scheme	87.26	80.79	84.00	1.84

Table 4 Performance of test results on ICDAR2017-MLT

Algorithms	**ICDAR2017-MLT**		
	Precision (%)	Recall (%)	F-measure (%)
PSENet	73.77	68.21	70.88
Proposed scheme	**77.67**	**69.98**	**73.83**

To further intuitively show that the performance of the proposed scheme is better than the original PSENet, two sets of experimental results are shown in Fig. 6, and the effectiveness of the proposed scheme is analyzed. As shown in Fig. 6 (a1, a2), the original PSENet has the phenomenon of missed detection and false detection. The missed detection or false detection of text objects often leads to failure in text recognition tasks. Because the text cannot be accurately detected and recognized, the final semantic information will not be understood. Compared with the original PSENet, the proposed scheme can precisely identify the text without missed detection or false detection and precisely locate the scene text regions and object boundaries in an image shown in Fig. 6 (b1, b2). In conclusion, the proposed scheme can precisely locate the scene text regions and object boundaries in an image and it has higher precision and a lower false detection rate than the original PSENet.

4 Discussion

In this paper, the scene text detection is optimized by both the Mixed Pooling Module (MPM) and the fusion networks (i.e., ResNeXt, and Res2Net). The idea of this scheme is to collect more context information with the different pooling operations. Furthermore, the fusion mechanism is used to enhance the multi-scale feature extraction ability of the backbone network and we discuss the following aspects in detail.

1. The role of the MPM. In this study, we analyze the effect of context information extraction ability of MPM on model detection precision. The results show that the

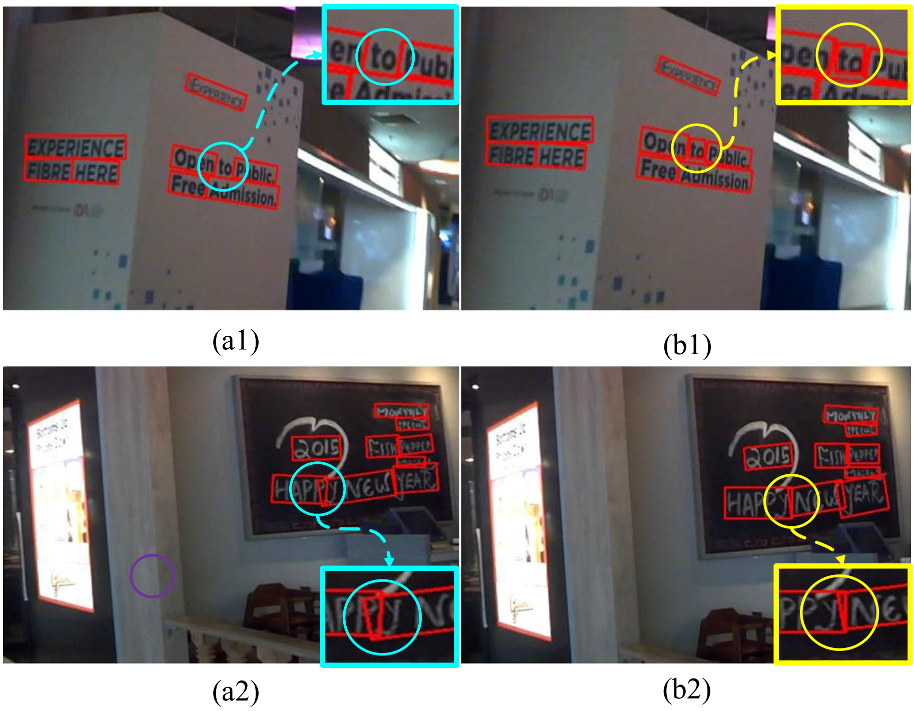
(a1)
(b1)
(a2)
(b2)

Fig. 6 Examples of test results: (**a1** and **a2**) are the missed detection and false detection results of the original PSENet, respectively; (**b1** and **b2**) are the results of the proposed scheme

designed MPM can improve the precision and reduce the missed detection and false detection of our scheme. The real reason is that, more comprehensive information can be extracted efficiently with the targeted pooling module for different text scenes, inspired by the assumption of the pooling strategies in [29, 30].

2. The effect of the fusion networks. Related studies have shown that the ResNet is mainly used to solve the problem of gradient vanishing in deep neural network, but its performance is still needed to be further improved for the diversity of scene text feature. Hence, the backbone network is optimized by the fusion networks to enhance the multi-scale feature extraction capability. The experimental results illustrate the effectiveness of this fusion networks on the ICDAR2015 dataset.

However, the complex post-processing steps and heavy network lead to the scene text detection algorithm still does not meet the requirement of real-time detection. At the same time, text detection is only the first key step for scene text recognition and the end-to-end text spotting framework is the further hot topic and study trend [31–33]. Therefore, our future research focuses on how to simplify the post-processing steps and build lightweight networks, aiming to improve the model speed and further build an end-to-end detection and recognition framework.

5 Conclusion

An efficient scene text detection scheme based on PSENet is proposed to solve the problem of missed detection and false detection existing in most scene text detection algorithms. The Mixed Pooling Module can capture the dependency between different text positions and collect context information, and precisely locate the scene text regions and object boundaries in an image. Additionally, the backbone network is optimized by combining ResNeXt and Res2Net to further improve its multi-scale feature extraction ability. Experimental results have demonstrated that, compared with common scene text detection algorithms, the proposed scheme has lower missing detection rate and higher detection precision. Specifically, the precision of the proposed scheme is improved by more than 5% compared with the original PSENet.

Abbreviations

PSENet: Progressive Scale Expansion Network; ResNet: Residual Network; MPM: Mixed Pooling Module; FPN: Feature Pyramid Network; FPS: Frames Per Second.

Acknowledgements

The authors would like to thank the help of all the people who influenced the paper. Specifically, they want to acknowledge the editor and anonymous reviewers for their valuable comments.

Authors' contributions

GL: Ideas; Creation of models; Software; ZZ: Conceptualization; Investigation; Review & Editing; Funding acquisition; YB: Methodology; Formal analysis; Funding acquisition; Validation; Review & editing; TL: Evidence collection; Original draft; Review & editing; ZX: Visualization; Data curation; Review & editing. All the authors read and approved the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61671412), Zhejiang Provincial Natural Science Foundation of China (Grant No. LY19F010002, LY21F010014), Natural Science Foundation of Ningbo, China (Grant No. 2018A610053, 202003N4323), Ningbo Municipal Projects for Leading and Top Talents (Grant No. NBLJ201801006), General Scientific Research Project of Zhejiang Education Department (Grant No. Y201941122), Commonweal Projects of Zhejiang Province (Grant No. LGN20F010001), General Project of Zhejiang Education Department(Grant No.Y201940951), the School Level Scientific Research and Innovation Team Project, Fundamental Research Funds for Zhejiang Provincial Colleges and Universities, College Students Innovative Training Project (Grant

No. 2021R420017), General scientific research projects of Zhejiang Provincial Department of Education (Grant No.Y202045430).

Availability of data and materials

The datasets used and/or analysed during the current study are available in the Robust Reading Competition repository, https://rrc.cvc.uab.es/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 30 April 2021 Accepted: 5 October 2021
Published online: 16 October 2021

References

1. W. Kazmi, I. Nabney, G. Vogiatzis, P. Rose, A. Codd, An efficient industrial system for vehicle tyre (Tire) detection and text recognition using deep learning. IEEE Trans. Intell. Transp. Syst. **22**(2), 1264–1275 (2021). https://doi.org/10.1109/TITS.2020.2967316
2. Y. Liu, L. Jin, C. Fang, Arbitrarily shaped scene text detection with a mask tightness text detector. IEEE Trans. Image Process. **29**, 2918–2930 (2020). https://doi.org/10.1109/TIP.2019.2954218
3. P. Cheng, Y. Cai, W. Wang, "A direct regression scene text detector with position-sensitive segmentation. IEEE Trans. Circuits Syst. Video Technol., 30(11): 4171–4181 (2020). https://doi.org/10.1109/TCSVT.2019.2947475
4. P. N. C. a. w. P. Shivakumara, R. Raghavendra, S. Nag, U. Pal, T. Lu, D. Lopresti, "An episodic learning network for text detection on human bodies in sports images," In IEEE Transactions on Circuits and Systems for Video Technology, 1–1 (2021). https://doi.org/10.1109/TCSVT.2021.3092713
5. S. Ren, K. He, R. Girshick, J. Sun, Faster RCNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell **39**(6), 1137–1149 (2017). https://doi.org/10.1109/tpami.2016.2577031
6. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A.C. Berg, "SSD: single shot multibox detector," In European Conference on Computer Vision (ECCV), 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2
7. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: unified, real-time object detection," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
8. L. Huang, Y. Yang, Y. Deng, Y. Yu, "DenseBox: unifying landmark localization with end to end object detection," arXiv preprint arXiv:1509.04874(2015)
9. M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A Fast Text Detector with a Single Deep Neural Network," In The National Conference on Artificial Intelligence (AAAI), 4161–4167 (2017).
10. M. Liao, B. Shi, X. Bai, "TextBoxes++: a single-shot oriented scene text detector," IEEE Trans. Image Process., 3676–3690 (2018). http://dx.doi.org/https://doi.org/10.1109/TIP.2018.2825107.
11. X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, "EAST: an efficient and accurate scene text detector," In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2642–2651 (2017). https://doi.org/10.1109/CVPR.2017.283
12. J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. Multimedia **20**(11), 3111–3122 (2018). https://doi.org/10.1109/TMM.2018.2818020
13. L. Tychsen-Smith, L. Petersson, "DeNet: scalable realtime object detection with directed sparse sampling," In IEEE International Conference on Computer Vision (ICCV), 428–436 (2017)
14. L. Pengyuan et al., "Multi-oriented scene text detection via corner localization and region segmentation," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7553–7563 (2018). https://doi.org/10.1109/CVPR.2018.00788
15. X. Wang, K. Chen, Z. Huang, C. Yao, W. Liu, "Point linking network for object detection," arXiv preprint arXiv: 1706.03646 (2017)
16. D. Deng, H. Liu, X. Li, D. Cai, "PixelLink: detecting scene text via instance segmentation," In The National Conference on Artificial Intelligence (AAAI), 6773–6780 (2018)
17. Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, X. Bai, "Multi-oriented text detection with fully convolutional networks," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4159–4167 (2016). https://doi.org/10.1109/CVPR.2016.451
18. X. Enze et al., "Scene text detection with supervised pyramid context network," In The National Conference on Artificial Intelligence (AAAI), 9038–9045 (2019)
19. L. Shangbang et al., "TextSnake: a flexible representation for detecting text of arbitrary shapes," In European Conference on Computer Vision (ECCV), 20–36 (2018). https://doi.org/10.1007/978-3-030-01216-8_2
20. W. Wenhai et al., "Shape robust text detection with progressive scale expansion network," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 9336–9345 (2019). https://doi.org/10.1109/CVPR.2019.00956

21. H. Qibin et al., "Strip pooling: rethinking spatial pooling for scene parsing," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4003–4012 (2020). https://doi.org/10.1109/CVPR42600.2020.00406
22. S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, "Aggregated residual transformations for deep neural networks," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5987–5995 (2017). https://doi.org/10.1109/CVPR.2017.634
23. S. Gao, M. Cheng, K. Zhao et al., Res2Net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 652–662 (2021). https://doi.org/10.1109/TPAMI.2019.2938758
24. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
25. C. Szegedy, W. Liu, Y. Jia et al., "going deeper with convolutions," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
26. T. Zhi, W. Huang, H. Tong, et al., "Detecting text in natural image with connectionist text proposal network," In European Conference on Computer Vision (ECCV), 56–72 (2016). https://doi.org/10.1007/978-3-319-46484-8_4
27. B. Shi, X. Bai, S. Belongie, "Detecting oriented text in natural images by linking segments," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3482–3490 (2017). https://doi.org/10.1109/CVPR.2017.371
28. H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, E. Ding, "WordSup: exploiting word annotations for character based text detection," In IEEE International Conference on Computer Vision (ICCV), 4950–4959 (2017). https://doi.org/10.1109/ICCV.2017.529
29. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, "Pyramid scene parsing network," In CVPR, 6230–6239 (2017)
30. J. He, Z. Deng, L. Zhou, Y. Wang, Y. Qiao, "Adaptive pyramid context network for semantic segmentation," In CVPR, 7519–7528 (2019)
31. Y. Liu, J. Yan, Y. Xiang, "Research on license plate recognition algorithm based on ABCNet," In IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE), 465–469 (2020). https://doi.org/10.1109/ICISCAE51034.2020.9236855
32. M. Liao, P. Lyu, M. He, C. Yao, W. Wu, X. Bai, Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 532–548 (2021). https://doi.org/10.1109/TPAMI.2019.2937086
33. W. Feng, W. He, F. Yin, X. Zhang, C. Liu, "TextDragon: an end-to-end framework for arbitrary shaped text spotting," In IEEE International Conference on Computer Vision (ICCV), 9075–9084 (2019), https://doi.org/10.1109/ICCV.2019.00917

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.