A Parallel Backbone Networks Structure for Scene Text Detection

Linyuan Li*
International College, Beijing University of Posts and Telecommunications, Beijing, China

*Corresponding author email: lly72824328@163.com

Abstract. Text detection in complex scenes is very hard realize by the diversification of text distribution, direction, and typesetting. This paper proposes one scene text detection method with end-to-end structure with parallel backbone network and region segmentation. With multiple deformable convolutions and extracting features of multi-dimensional text regions, multiple candidate regions of different sizes are generated and corresponding states are further given. Experiments show that compared with baseline, this method can further adapt to the problem that the different shapes and angles of the target in the image lead to the decrease of accuracy.

1. Introduction
Thanks to the rapid development of technology, Optical Character Recognition is rapidly developed with widely used in offices as a very natural human-machine interface [1]. OCR generally refers to the recognition of possible text information in an image through processing and analysis. Scene Text Recognition also belongs to the field of OCR technology in a broad sense, which is to recognize text from natural scene pictures.
As early as the 1960s, OCR technology began to be paid attention to by many researchers. The earliest characters that can be recognized are only 0-9, mainly serving postal codes [2], helping post offices to distribute letters. In 1986, China's Chinese character recognition technology entered the product stage, but the recognition rate and product performance of these products were insufficient to meet actual requirements.
OCR method based on machine learning and deep learning has gradually emerged. With end-to-end system, it can deal with sequences of any length was proposed in literature [3]. It is based on 111-5K, street view text and ICDAR database to generate a smaller model. DCNN cannot generate variable length label sequences because RNN is usually used for fixed size input and output. RNN is designed to process sequences and does not require the position of each element. The model is named CRNN, which combines DCNN and RNN. LSTM operates on the sequence, and the last module is transcription, which converts each frame prediction made by the BNN into a label sequence. It uses conditional probability to remove repeated labels and blanks, and can use the pre-post algorithm to efficiently calculate the conditional probability formula. It can perform end-to-end training on image and sequence pairs.
The paper [4] was the method of text detection based on the character and affinity between them. The framework referred to CRAFT and designed with a convolution network score. Based on VGG16 and the architecture of Unet, the network is framed as for the labels of training set, Character Boxes and Affinity Boxes are formed. Adding the 2D Gaussian, Region score and Affinity Score are produced.
Using the label and images, the model is trained. The next step is the segmentation of words. After the image passing through the network, region score and affinity score are achieved and then have the character box. At the inference stage, the output should be word-level bounding boxes QuadBox.

CountourNet is aimed to handle these questions [5]: 1. FP in the text representation; 2. Too much scale variance is hard to learn. Its properties, which will lead to false prediction, should be limited. Heatmap of two direction on vertical and horizontal can be merged into a result with higher accuracy. The model consists three modules, Adaptive Region Proposal Network, LOTM and PR. LOTM uses conv 1*1 to get the convolution of 1*k and k*1, and then the latent pics of two directions are generated. PRA focus on inhibiting the FPs us NMS (Non-Maximum Suppression).

The variation of the scale of pictures challenges the robust and function of text detection model. GNN is developed to solve this problem. GNN consists of Backbone, GNM and shared Text Detection Header. The characteristic pictures are extracted from the backbone and pas to GNM, which is made of SNU and ONU. SNU extract scale-specific features and ONU extract oriented-specific features. Based on the combinations of SNU and ONU, GNM will generate distinct feature maps and they delivered to detection headers. The core of this network is GNM [6].

At present, OCR has been the main research field in various commercial applications such as banking, publishing and navigation. The more common one in consumer applications is information translation, such as the conversion of certain languages. However, the text in natural images has an infinite number of styles, such as shapes, styles, and sizes. Therefore, it is very difficult to attempt. In addition, noise, occlusion, uneven light, etc. will cause a rapid decline in recognition accuracy. This paper proposes an end-to-end STD method, which simultaneously achieve deformable convolution and extract the character of text in image by multiple dimensions through the parallel Backbone Network. Its purpose is to generate multiple matching the target length and width ratio Candidate area reference frame. Some experiments show that it can improve further it accuracy and execution speed of image recognition tasks with low computational overhead.

The rest of this paper is: the second part gives a general method for natural scene text recognition based on CNN; the third part proposes the end-to-end scene text detection structure; the fourth part is the experimental results; the final is conclusion and next steps.

2. CNN-BASED Scene Text Detection

The general OCR network model needs to complete two parts: area detection and recognition. Region detection methods include many structures, such as first generating a large number of text candidate frames, using a random forest classifier to remove errors and redundancy, and finally adjusting the frame size through CNN. In addition, it uses the YOLO target detector as the basis to implement the FCN regression network, where the input image is divided into a specific size and the transform invariant predictor on the VGG-16 predicts the position, the rotation angle and the confidence rate for each region. In terms of recognition, you can regard text recognition and use a classifier for converting text images into a dictionary; there are also methods to treat text as a sequence and use two-way LSTM and CTC FCN. In addition, there is the end-to-end model network that combines them.

As shown in Fig.1, The output is bounding box. The model consists of a large, multilayer network in unsupervised feature learning together with off-the-shell methods to achieve end-to-end lexicon-driven detection [7]. Firstly, a 32*32 pixel window is designed to detect whether the window contains a central character. Then, CNN is used for text detection. First layer use unsupervised algorithms and another conv layer is staked on it to obtain the second layer response map. These outputs are fully connected to the classification. Its process of the end-to-end pipeline involves using window detection to obtain a set of candidate lines of text and these response can be used to estimate the space in line; integrating the response with candidate spacing and beam search to obtain end-to-end results.

Figure 1. Example of the output bounding box [7].
Mask R-CNN [8] can robustly detect text in natural scene. The Pyramid Attention Network is used as the backbone for text detection. The network includes four modules: PAN backbone for computing a multi-scale convolution features; RPN generate rectangular proposal; Fast RCNN classify and output bounding boxes.

In general, the text detection methods are separated two categories: pixel-based vs. anchor-based. However, pixel-based is high precision, low recall due to small text and anchor-based has high recall, it suffers from “Anchor Matching Dilemma”. In addition, the existing method have bad performance in detect long text. To cater these problems, Pixel-anchor method was developed. In pixel-based module, image is the input and the output is attention heat map, which stand for possibility of whether a pixel belongs to text. The output will be delivered to anchor module [9].

HMCP increases the accuracy of HED and HMCP’s detect box will detect the polygon and other forms of quadrilateral. Once we input a picture, three map and the detection will be generated according the feature maps. The backbone of HMCP is VGG16 and the detection boxes of Text is generated [10].

Scribble lines text detection use a simpler method [11]. This method includes weakly labeling method and weakly supervised scene text detection. Several annotated points are set near the centerline. As for the weakly supervised scene text detection. The network includes a backbone to extract the features, RPN to get the candidate samples, Text-line segmentation to get the line and boundary reconstructions modules.

However, several problems exist, such as hard to detect irregular text field and hard to segment single word. The text field use VGG16 as the backbone and then visualize and calculate it in magnitude and direction information [12]. The magnitude vector is used to distinct text/non-text pixels and direction is used for post processing. Text super pixel segmentation of the candidate text pixel is generated according to orientation vector. Based on the super pixel, the generate candidate instance is produced. After generating the instances, some filters are adopted for the non-text.

![Figure 2. Textfield pipeline [12].](image)

3. Parallel Backbone End-to-end Scene Text Detection

Recently many methods of scene text recognition have been affected by sequence-to-sequence tasks such as machine translation, and CTC and LSTM have been introduced to process sequence-to-sequence deep neural network structures [13]. In fact, the feature sequence extraction in the network has an important impact on the performance of detection and recognition. Even the simple introduction of Resnet can significantly improve the performance. However, the feature sequence extraction part of the network is usually also the most computationally expensive place, so its structure needs to be improved to achieve efficiency optimization.

The overall architecture of the backbone network method [14] includes four components: FPN is the network structure of the backbone network, RPN is used as a text candidate network structure, Fast RCNN regression boundary, and mask is used for segmentation. RPN is responsible for generating multiple text candidates. The output of RPN is used as the input of Fast R-CNN. The calculation goal is to select the correct candidate and segmentation result. Further refine the multi-scale semantic features, and use the top-down backbone network fusion method.

ESIR is a new End-to-end method [15], which is based on the Tensorflow framework to achieve ESIR scene text recognition and uses adaptive and entropy loss ADADELTA optimization. The rectification
method proposed by ESIR iteratively corrects perspective and distortion, and the corrected result is sent to the recognition network. Iterative rectification training is a way of backward propagation, which solves several typical constraints for scene problems, robust distortion correction is realized. Evaluating ESIR needs million iterations. A variety of models are used in ESIR, including VGG/ResNet as the network backbone, and tested on different datasets.

As pointed out above, it is hard to check some arbitrary text in natural scenes. This paper proposes an end-to-end parallel backbone network scene text detection method, the purpose is to realize the deformable convolution at the same time, and realize the multi-dimensional extraction of the feature vector of the text region in the image through parallel, and finally output the reference of the multiple matching candidate regions. The frame is shown in Fig.3.

Figure 3. End-to-End parallel backbone scene text detection.

The principle of this structure is that the backbone network uses Region Proposal Networks of multiple sizes to adapt to different shapes of text, at the same time, it judges the probability that the proposal corresponds to the text and regresses to adjust the position of the positive sample proposal. The text box proposal can either pre-set the anchor box or directly predict it. In terms of text segmentation, the parallel Resnet Network is used to distinguish the candidate boxes output by the parallel Region Proposal Network, and the target and connection status of each region are given. The purpose is to solve the problem of different shapes and angles.

4. Experimental Results

The experiment is based on the ICDAR2015 [16], which is one useful dataset in natural scene text detection. ICDAR2015 mainly contains natural images taken randomly and most of the images appear severely distorted and blurred. Our model is pre-trained and fine-tune on the ICDAR2015 dataset. We train on 3 Titan GPUs with a batch size of 4. The data in the training process are adjusted successively according to different proportions. We evaluated them of this parallel backbone network on a standard benchmark. The test data is collected independently from the network, covering multi-directional text and curved text.

Fig.4 shows the results of precision/recall/f-measure on the test dataset using baseline and parallel backbone networks. Compared with the baseline method, these scores imply that the parallel backbone network method that provides different size area proposals can better adapt to different shapes of targets. Fig.5 shows the P/R/F results of the parallel backbone network and the final proposed network on the test dataset. It can be seen that the final proposed with parallel text segmentation. The parallel backbone network candidates are further distinguished and the corresponding status is given to further solve the accuracy drop caused by different angles.

Figure 4. Experimental scores on testing dataset.
5. Conclusions

The text detection technology in natural images has always been the focus and difficulty of the text area detection research. The reason is that the shape and size of the text itself are irregular, and the occlusion and noise in the image greatly interfere with the text area detection. This paper proposes one parallel backbone to achieve end-to-end STD. Its feature is to generate some candidate regions matching the target aspect ratio by implementing deformable convolution and having the characters of the text region from multi-dimensional image.

References

[1] Mori S, Suen C Y, Yamamoto K. Historical review of OCR research and development [J]. Proceedings of the IEEE, 1992, 80(7): 1029-1058.
[2] Alginahi Y M, Siddiqi A A. A proposed hybrid OCR system for Arabic and Indian numerical postal codes[C]. 2009 International Conference on Computer Technology and Development. IEEE, 2009, 2: 400-404.
[3] 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 [J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11): 2298-2304.
[4] Baek Y, Lee B, Han D, et al. Character region awareness for text detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9365-9374.
[5] Wang Y, Xie H, Zha Z J, et al. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11753-11762.
[6] Xu Y, Duan J, Kuang Z, et al. Geometry normalization networks for accurate scene text detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9137-9146.
[7] Wang T, Wu D J, Coates A, et al. End-to-end text recognition with convolutional neural networks[C]. Proceedings of the 21st international conference on pattern recognition (ICPR2012). IEEE, 2012: 3304-3308.
[8] Huang Z, Zhong Z, Sun L, et al. Mask R-CNN with pyramid attention network for scene text detection[C]. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019: 764-772.
[9] Li Y, Yu Y, Li Z, et al. Pixel-anchor: A fast oriented scene text detector with combined networks[J]. arXiv preprint arXiv:1811.07432, 2018.
[10] Yao C, Bai X, Sang N, et al. Scene text detection via holistic, multi-channel prediction[J]. arXiv preprint arXiv:1606.09002, 2016.
[11] Zhang W, Qiu Y, Liao M, et al. Scene Text Detection with Scribble Lines [J]. arXiv preprint arXiv:2012.05030, 2020.
[12] Xu Y, Wang Y, Zhou W, et al. Textfield: Learning a deep direction field for irregular scene text detection[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5566-5579.
[13] Zhang H, Yao Q, Yang M, et al. Efficient backbone search for scene text recognition [J]. arXiv preprint arXiv:2003.06567, 2020.

[14] Lyu P, Liao M, Yao C, et al. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 67-83.

[15] Zhan F, Lu S. Esir: End-to-end scene text recognition via iterative image rectification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2059-2068.

[16] Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. ICDAR 2015 competition on robust reading[C]. 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015: 1156-1160.