Improved Chinese Complexly Arranged Scene Text Detection

Pan Hao¹,a, Yang Yunxiang¹,b,* and Pan Shuang²,c
¹School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei, China
²School of Economics, Wuhan University of Technology, Wuhan, Hubei, China
¹panhao163com, byangemail2@163com, c1131337572@qqcom
*Corresponding author

Abstract: At present, the existing text detection algorithms in natural scenes have made great progress, but most of them are for text line detection. On the one hand, they have limitation on the effective detection of complexly arranged (such as vertical arrangement, circular arrangement) text lines. On the other hand, the length of text lines varies greatly, which makes text detection very challenging. In this paper, an optimized SSD detection algorithm is proposed for single Chinese text detection. Convolution enhanced module is used to improve the detection of small target text. The introduction of deformable convolution improves the detection of deformable text. A text region connectivity algorithm is proposed to connect complexly arranged single text into readable text lines. In the natural scene Chinese text datasets on CTW and RCTW 17 datasets, mAP of 73.1% and F1 scores of 61.3% were achieved respectively.

1. Introduction
Although previous natural scene text detection⁴°³⁴ has made great progress, it is still a challenging problem. The main difficulties are the variety of text arrangement, text size and the complexity of background. On the whole, text detection can be divided into two categories: one is based on the object detection model (Faster-RCNN⁴, SSD⁵), such as CTPN⁶, TextBoxs⁷; the other is based on the semantic segmentation method such as MOTD⁸. The algorithm based on object detection model uses rectangular box to detect text lines, so it has a poor effect on multi-directional text detection of complex arrangement. The method based on semantic segmentation has a poor effect on small text detection due to the loss of some information in the process of pooling, and often produces overlapping areas when returning to a longer text box area.

In this paper, the authors adopts a single Chinese text detection and proposes a region connectivity algorithm to better detect arbitrary direction of the arranged text. By introducing the convolution enhanced module, the complete connection of the convolution layer is transformed into sparse connection, and features are extracted on multiple scales by using different sizes of filters, which increases the width of the network without increasing the parameters, and increases the adaptability of the network to multiple scales. The deformable convolution enables the network to better detect the deformed text region. Fusing the features of different convolution layers increases the feature map relationship between different layers and the number of feature maps. Therefore, this fusion method not only solves the problem of overlapping box existing in the original algorithm, but also solves the problem of small target detection.
2. Methodology

2.1. Network design
The main component of the proposed algorithm is a convolutional neural network, which is inspired by SSD object detection algorithm and generates fixed size boundary box and the probability score of corresponding region belonging to text. The idea of network design mainly includes the following points: 1. different degrees of convolution layer fusion are used to compensate for the information loss caused by pooling and reduce the occurrence of duplicate boxes caused by output prediction of individual convolution layer. 2. By introducing convolution enhanced module and deformable convolution, the detection accuracy of small-scale and deformable text is improved. 3. Resnet\(^9\) is used as the basic framework of the model, which can extract powerful semantic features. 4. By using the proposed connected region algorithm, the detected single text is connected to a complex array of text line regions. The overall structure of the model is shown in Fig1.

![Network model structure](image)

2.2. Convolution layer fusion
Most detection algorithms use top-level features to predict, such as Faster RCNN, which only uses the last-level features of the network. The lower-level feature semantic information is less, but the target location is accurate. The higher-level feature semantic information is rich, but the target location is rough. In addition, although some algorithms use multi-scale feature fusion methods such as SSD, the feature maps of different layers are independent input of the classification network, which is easy to see the same object being detected simultaneously by different sizes of frames\(^{10}\). In this paper, the authors propose a convolutional layer fusion structure with multi-scale features by lateral connection, which is used to increase the feature map links between different layers and reduce the appearance of overlapping boxes. The three lateral connections are used to detect small, medium and large text objects, which can effectively alleviate the learning difficulties of text/non-text classification and bounding box regression of network models, and make the model better able to handle multi-scale text. The third residual module convolutional layer fusion diagram is shown in Fig2.
Fig 2. Convolution layer fusion

Convolution layer fusion algorithm given by:

\[ \hat{C}_3 = W \ast C_3 + b \]  
\[ \hat{C}_4 = \text{UP}(C_4) \]  
\[ \text{result} = \begin{cases} \hat{C}_3 & \text{and} \\ \hat{C}_4 & \end{cases} \]  

W is the convolution kernel weight and b is the bias. UP is upsampling operation, and the size of \( \hat{C}_4 \) obtained after upsampling operation is the same as \( \hat{C}_3 \). The detection overlapping box is reduced by using the convolution layer fusion strategy as shown in Fig3. The result is shown in table 1.

![Fig 3. Comparison of before (left) and after (right) convolution layer fusion](image)

Table 1 AP comparison of text detection in different sizes before and after improvement

|          | Large (size >40px) | Medium (20px<size<40px) | Small (size< 20px) |
|----------|--------------------|-------------------------|-------------------|
| fusion   | 78.27%             | 76.02%                  | 58.96%            |
| No fusion| 77.15%             | 75.22%                  | 58.36%            |

2.3. Convolution enhanced module

In order to better deal with texts of different scales, the authors introduce the convolution enhanced module inspired by the GoogLeNet series\(^{[11-13]}\) network model. As shown in Fig4.
Three different convolution kernels are used in the convolution enhanced module: 1 x 1, 3 x 3, and 5 x 5. To handle text of different scales, text of different scales is activated in different convolution enhanced module. At the same time, further factorization n x n convolution into two convolutions, which is 1 x n convolution, followed by n x 1 convolution. The comparison of the effects before and after using the convolution enhanced module is shown in Fig.5 and Fig.6.
2.4. Deformable convolution
In convolution enhanced module, the authors introduce a deformable convolution layer to better cope with the deformation text. Fig.7 compares standard convolution and deformable convolution.

\begin{equation}
y(P_0) = \sum_{\Delta P_n \in \mathcal{R}} w(P_n) * x(P_0 + \Delta P_n)
\end{equation}

\begin{equation}
y(P_0) = \sum_{P_n \in \mathcal{R}} w(P_n) * x(P_0 + P_n)
\end{equation}

\(P_0\) is a point on the feature map, \(\mathcal{R}\) defines the size of the receptive field and the step size of the convolution. For a convolution kernel with a stride of 1 and a size of 3*3, \(\mathcal{R} = \{(-1, -1), (-1, 0), \ldots, (0, 1), (1, 1)\}\). \(\Delta P_n\) is learned from convolutional networks.

2.5. Regional connectivity algorithm
For the detected single text, it can be connected into a readable text line by the proposed region connectivity algorithm. The algorithm is based on the a priori assumption that the connected area...
characters are closely adjacent. The pseudo code of the algorithm is as follows:

```
Algorithm 1 Regional connectivity algorithm
Input: point /*center point of one box*/
points /*center point of all box*/
branch /*Connected box collection*/
Output: branch
1. branch<-point;/*this point enter branch*/
2. Twopoint=closetwo(point) /*The closest two points to point*/
3. if IsMid(point,Twopoint) /*point is the middle point of Twopoint*/
4. genbranch(Twopoint[0],points,branch)
5. genbranch(Twopoint[1],points,branch)
6. else:
7. if len(branch)==1:
8. Ltwop=closetwo(Twopoint[0])
9. Rtwop=closetwo(Twopoint[1])
10. if IsMid(Twopoint[0], Ltwop)
11. genbranch(Twopoint[0],points,branch)
12. end if
13. if IsMid(Twopoint[1], Rtwop)
14. genbranch(Twopoint[1],points,branch)
15. end if
16. end if
```

Fig 8. Comparison of before (left) and after (right) regional connectivity algorithm

3. Experiments detail

3.1. Image preprocessing
There are very small, blurred or heavily occluded characters in the image, which can not be recognized even by human eyes. Such characters are not obvious or even abnormal, so they do not participate in training. The specific method is that the text of short-side less than 10px is shielded in the original image and not trained as a text area. This design method reduces the confusion between text and non-text, and thus facilitates feature learning.

3.2. Expand data
The acquisition of label data is expensive, but a large amount of label data is necessary for deep learning models. Therefore, it is very valuable to artificially synthesize reasonable data in the natural scene conditions. The authors uses SynthText\textsuperscript{[15]} method to expand data and generate Chinese data sets, unlike the original method of generating English data sets.

3.3. Loss function
The overall objective loss function is a weighted sum of the localization loss (loc) and the confidence loss (conf), given by

$$L(x, c, l, g) = \frac{1}{N} (L_{conf}(x, c) + L_{loc}(x, l, g))$$

where $N$ is the number of matched default boxes. If $N = 0$, we set the loss to 0. The confidence loss is the softmax loss over multiple classes confidences $(c)$, given by

$$L_{conf}(x, c) = -\sum_{i} x_{ij} \log (\hat{c}_{ij}^{p}) - \sum_{i} \log (\hat{c}_{ij}^{0})$$

$$\hat{c}_{ij}^{p} = \frac{\exp(c_{ij})}{\sum_{p} \exp(c_{ij}^{p})}$$

The localization loss is a Smooth L1 loss between the predicted box $(l)$ and the ground truth box $(g)$ parameters. Similar to Faster R-CNN, the authors regress to offsets for the center $(cx; cy)$ of the default bounding box $(d)$ and for its width $(w)$ and height $(h)$.

$$L_{loc}(x, l, g) = \sum_{i} \sum_{m} x_{ij} L_{1}(l_{ij}^{m} - \hat{g}_{ij}^{m})$$

$$\hat{g}_{ij}^{cx} = (g_{ij}^{cx} - d_{ij}^{cx}) / d_{ij}^{w} \quad \hat{g}_{ij}^{cy} = (g_{ij}^{cy} - d_{ij}^{cy}) / d_{ij}^{h}$$

$$\hat{g}_{ij}^{w} = \log(\frac{g_{ij}^{w}}{d_{ij}^{w}}) \quad \hat{g}_{ij}^{h} = \log(\frac{g_{ij}^{h}}{d_{ij}^{h}})$$

$$L_{1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| & |x| \geq 1 \end{cases}$$

3.4. Training

The network is trained end-to-end using ADAM optimizer. To speed up learning, 512x512 crops are sampled from images to form a mini-batch of size 24. Learning rate of ADAM starts from 1e-3, decays to one-fifth every 5 epochs, and stops at 1e-5. The network is trained until performance stops improving.

3.5. Detect multi-scale text

![Fig 9. Result of image magnification](image)

With the deepening of the network model, image feature maps are pooling, which leads to the tendency of the network to detect larger sized text, and easy to ignore smaller sized text. One of the most important challenges in text detection is that the size of text changes greatly. In order to further improve the detection accuracy of small size text, the original image be enlarged and then sent to the network for detection. Because the original image is enlarged, the information of small size text can still be preserved after pooling. Thus, the detection accuracy of the whole model is improved.
However, the over-enlargement of the image will result in the decrease of detection accuracy. The experimental results are shown in Fig9.

4. Results and Analysis

4.1. CTW datasets

![CTW dataset detection samples](image)

Chinese Text in the Wild (CTW\cite{17}) contains 32,285 images with 1,018,402 Chinese characters, far beyond the previous datasets, which were taken from Tencent Street View and obtained from dozens of different cities in China without any special purpose. Due to its diversity and complexity, this database is extremely challenging. Fig10 shows the results of the test samples on the data set, Fig.11 shows the mAP curves for different sizes of text, and Fig.12 shows a comparison of the experimental results.

![mAP curves of different size text](image)

![Results on CTW](image)

4.2. RCTW-17 datasets
RCTW-17[18] consists of 12,263 pictures of natural scenes containing Chinese, most of which are taken directly by the camera or mobile phone, a small part of which is generated images, and each image contains at least one line of Chinese. The label is done by drawing the quadrilateral to mark a text line instead of the word unit, so the training data is generated by SynthText method and is marked in units of a single text. Fig.13 shows the test samples. Table2 show the results on RCTW-17.

| Method     | Recall  | Precision | F-score |
|------------|---------|-----------|---------|
| SegLink[19] | 0.404   | 0.760     | 0.527   |
| FTSN[20]   | 0.471   | 0.741     | 0.576   |
| Our method | 0.521   | 0.80      | 0.613   |

5. Conclusion and Future Work
In this paper, the authors propose an improved Chinese text detection method based on a single text. Its main advantages are: 1) By detecting a single text and using the region connectivity algorithm proposed in this paper, the text region can be more accurately located. 2) The convolution enhancement module improves the network's ability to detect multi-scale text, and the introduction of the deformable convolution enables the network to better extract the text features that are deformed in the natural scene at the micro level. 3) The use of convolutional layer fusion strategy reduces the occurrence of overlapping boxes. 4) Enlarging the original image can better detect small-sized text. The method proposed in this paper has been tested on different datasets and achieved better results. Future work will focus on faster detection structures and network structures that integrate detection and recognition.

Acknowledgement
This research was financially supported by the Major Special Projects of Hubei Science and Technology Plan (Grant NO. 2016AAA058).

References
[1] Cho H, Sung M, Jun B, Canny Text Detector: Fast and Robust Scene Text Localization Algorithm, IEEE Conference on Computer Vision and Pattern Recognition, pp.3566-3573, 2016.
[2] Gupta, Ankush, Andrea Vedaldi, and Andrew Zisserman, Synthetic data for text localisation in natural images., IEEE Conference on Computer Vision and Pattern Recognition, pp.2315-2324,2016.
[3] Zhou X, Yao C, Wen H, et al, EAST: an efficient and accurate scene text detector, IEEE Conference on Computer Vision and Pattern Recognition, pp. 2642-2651, 2017
[4] Ren, Shaoqing, et al, Faster r-cnn: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, pp. 1137-1149, 2017.

[5] Liu, Wei, et al, Ssd: Single shot multibox detector, *European conference on computer vision*. pp. 21-37, 2016.

[6] Tian, Zhi, et al, Detecting text in natural image with connectionist text proposal network, *European Conference on Computer Vision*, pp. 56-72, 2016.

[7] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, Textboxes: A fast text detector with a single deep neural network. *AAAI Conference on Artificial Intelligence*, pp. 4161-4167, 2017.

[8] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multi-oriented text detection with fully convolutional networks, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4159–4167, 2016.

[9] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[10] Jeong J, Park H, Kwak N. Enhancement of SSD by concatenating feature maps for object detection, 2017.

[11] Szegedy C, Liu W, Jia Y, et al, Going Deeper with Convolutions, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[12] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift, *International Conference on International Conference on Machine Learning*, pp. 430-439, 2015.

[13] Szegedy C, Vanhoucke V, Ioffe S, et al, Rethinking the Inception Architecture for Computer Vision, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818-2826, 2015.

[14] Dai J, Qi H, Xiong Y, et al, Deformable Convolutional Networks, *IEEE International Conference on Computer Vision*, pp. 764-773, 2017.

[15] Gupta, Ankush, Andrea Vedaldi, and Andrew Zisserman, Synthetic data for text localisation in natural images., *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2315-2324, 2016.

[16] Kingma D, Ba J, Adam: A Method for Stochastic Optimization, Computer Science, (2014)

[17] Yuan T L, Zhu Z, Xu K, et al. Chinese Text in the Wild, 2018.

[18] Shi B, Bai X, Belongie S. Detecting Oriented Text in Natural Images by Linking Segments, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 724-733, 2017.

[19] Shi B, Yao C, Liao M, et al. ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17), *IAPR International Conference on Document Analysis and Recognition*, pp. 1429-1434, 2017

[20] Dai Y, Huang Z, Gao Y, et al, Fused Text Segmentation Networks for Multi-oriented Scene Text Detection, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 534-544, 2017.