Applicable Scene Text Detection Based on Semantic Segmentation

Shifu Gu¹ and Fusheng Zhang²*

¹ School of Electrical Engineering and Electronic Information, Xihua University, Sichuan 610039, China
² Bytedance, Inc., Beijing, China
Email: Zhangfusheng.cheung@bytedance.com

Abstract. In the past few years, scene text detection problem has attracted much attention and widely studied for many applications. Segmentation based methods are quite popular in this filed, it was not only for the simple post-processing procedure, but also for the ability of handle the various shape of text. In this manuscript, a method named CFPM-IDB was proposed which based on a new decoder module Consolidated Feature Pyramid Module (CFPM) and a light weight segmentation head Improved Differentiable Binarization (IDB). The CFPM module originated from Feature Pyramid Networks and achieved an idea trade-off between local details and macro semantic information. IDB module performed binarization process by threshold map and probability map. Combining them results in CFPM-IDB, which improved the performance of scene text detection, both precision and efficiency. Experimental results on two different datasets proved the promising advance in the accuracy and speed of this model.

1. Introduction

In the Internet era, text, images and videos are three of the main forms of information dissemination. Text is often embedded in pictures and videos that accounts for the most network traffic. In order for the computer to understand and process these contents, it is necessary to locate and recognize the text in them to understand videos and pictures. As a basic and important problem in the field of computer vision, text detection has a wide range of applications. It can be applied to the navigation control of autopilot vehicle to identify road signs, can be used in bill and form recognition, and can be utilised in accomplishing translation by mobile devices.

The traditional OCR technology is suitable for the images obtained by scanning documents with optical scanners. There are many advantages of neat printed text such as uniform camera angle, clean background and homogeneous illumination, and its post processing steps and methods of locating text are simple and effective, which has been widely used in large-scale commercialization. While modern OCR technology is facing plenty of challenges, especially, images are token from open nature scene, these images are affected by these disturbances: (1). Multiple oriented text. (2). Irregular text and image distortion. (3). Variation of text shapes. (4). Fuzzy and wrong annotations. (5). Complex background interference. (6). Uneven illumination. (7). Low resolution images. (8). Overlapped text and long text. In some application, the integrity and tightness of detection are essential to recognition contents of...
images. Considering these defects in images, text detection problem is quite knotty. Figure 1 shows some examples of these situations.

In addition to the complex image problems, another problem that plagues the academic community is the high cost of achieving annotated training samples. Therefore, the current datasets popularly leveraged by the researchers are usually several thousand. For example, icdar2015 [1], a text detection dataset published by ICDAR which holds a robust reading academic competition, training set contains only 1000 images and 500 images in testing set. How to train a better model with these limited data is a problem worth studying.

Figure 1. Four examples indicate the tough situation that hinders text detection.

Recent scene text detection techniques based on CNNs can be roughly classified into three categories: regression-based methods, segmentation-based methods and component-based methods.

Regression-based methods: These methods are usually improved from the general object detection algorithms and can be further subdivided into one stage method [2, 3] or two stage method [4, 5]. The one stage method directly regresses the coordinates of the text boxes, and the two stage method includes the stage of generating candidate text area and the stage of refining the text box. The two stage method is usually able to achieve high accuracy, but also endures the high demand for capacity of calculation. The text boxes location is different from the general object detection. The text appears in the image with a variety of irregular shapes and length width ratios. In response to this situation, Liao et al. proposed the textboxes algorithm [6] in 2017, which adapts to the text area with variable shapes by modifying the convolution kernel and the shape of the anchor. Furthermore, in 2018, Liao et al. proposed the rotation sensitive regression detector (RSDD) algorithm [7]. In order to make full use of the rotation invariant features in the image, the algorithm actively uses the convolution kernel of rotation to enhance the detection effect of text in various directions, but it also has the drawback that the shape of text in all open environments cannot be exhausted.

Segmentation-based methods: This method is usually inspired by the general instance segmentation algorithm, which aims to find the text area at the pixel level and infer the candidate text box in the segmented image. Compared with the regression-based method, the segmentation-based method can detect any shape of text more easily. This method leverages pixel level prediction results and post-processing algorithm steps to get the text boxes. In 2016, Zhang et al. First applied full convolution networks (FCN) to the text line detection task [8], extracted the candidate text area using FCN, then used MSER algorithm to obtain the text information, and inferred the coordinates of the text area through post-processing algorithm. In 2017, He et al. proposed the Mask R-CNN for the general object detection task [9] and improved the performance in the instance segmentation task by adding an object mask branch. In 2018, Lyu applied Mask R-CNN to the text line detection task and proposed the Mask Spotter [10]. For the text area without fixed shape, it predicted the probability mask in character level, and then divided it as a sample. In 2018, Deng et al. proposed PixelLink [11], starting from predicting the relationship between each pixel and its adjacent pixels in eight directions,
combining pixels belonging to the same instance. In 2019, Wang et al. proposed PSE net [12], which separated different text instances with progressive scale expansion fashion, avoiding overlapped text areas being fused into a text example. For the case that images contain dense text and curved text instances, PSE net achieved significant effect.

Component-based methods: This is a bottom-up method, which first locates a single portion or character, and then combines them into a character or word through a series of post-processing measures. In 2016, Tian et al. proposed CTPN [13], which first uses CNN and bidirectional LSTM learning spatial and sequential features, then adopts region proposal mechanism similar to Fast R-CNN [4] to generate a large number of dense candidate regions, and then selects candidate boxes with the maximum probability value contained in each anchor, and merges these candidate boxes with text to form the final text region. Compared with the Fast R-CNN, the CTPN enhances the sequence of features, while the CTPN uses a fixed number of anchors with fixed direction, it is unable to detect the irregular quadrilateral text area and curved text. In 2017, Hu et al. proposed WordSub algorithm [14], aiming at the current situation of lack of character level annotation, this algorithm utilised weak supervised learning method to obtain the position coordinates of each character, and after the text structure analysis method including character clustering, it can obtain words or text lines after processing. In 2017, Shi et al. proposed the Seglink [15], which is similar to the SSD, firstly, predicted the connection relationship between multiple slices containing text characters and slices. These slices may contain one or more characters, and then merge the slices into a text instance or split them into different text instances based on the connection relationship. Compared with CTPN, Seglink can output the rotation angle information of non-horizontal text area, but it also has the imperfection that it can’t deal with curved text and deformed text.

2. Methodology
In this study, a segmentation based text region detection algorithm is adopted. The feature extraction backbone network adopts the Consolidated Feature Pyramid Module (CFPM) network structure. The segmentation head design adopts the Improved Differentiable Binarization IDB. The overall structure is shown in figure 2.

![Figure 2](image_url)

**Figure 2.** Architecture of the proposed CFPM-IDB method.

In figure 2, the probability map and threshold map are marked as $P$ and $T$ respectively, and their sizes are both $H \times W \times 1$, which are the same as the original image, but there is only one channel. After extracting features from the input image by two stacked CFPM modules, the output features include four feature maps, the sizes of which are 1/4, 1/8, 1/16 and 1/32 of the original image, and the number of channels is 128. In other words, according to the general practice, the 1/8, 1/16 and 1/32 feature maps are enlarged to 1/4 of the original image, and the nearest neighbour difference method is used to reduce the amount of computation. Then the probability map $P \in \mathbb{R}^{H \times W}$ and threshold map $T \in \mathbb{R}^{H \times W}$ are predicted by the feature map $F$, and then the approximate binary map $B \in \mathbb{R}^{H \times W}$ is generated by the probability map $P$ and threshold map $T$. In the process of training, probability map,
threshold map and approximate binary map all have training labels, in which probability map and approximate binary graph share one training label; in reasoning stage, the boundary box of text area can be obtained by approximate binary map or probability map. The boundary box of the text region can be obtained from the approximate binary map or probability map through the connected domain expansion algorithm.

2.1. CFPM

Based on the design concept of FPN network structure, this paper improves a feature enhanced network connection mode, CFPM, is showed in figure 3. In order to balance the accuracy and running speed of the model, resnet-18 is adopted as backbone network, which is commonly used lightweight backbone network, can significantly reduce the volume of the model and reduce the inference time. However, compared with other larger models, it has the disadvantage of weak representation ability, so we need to pay attention to the design of feature fusion scheme.

Figure 3. Details of CFPM network, the Deconv and \( \oplus \) operations are collectively referred to as Deconv modules, and Conv and \( \oplus \) operations are collectively referred to as Conv modules.

The original input image size of resnet-18 is 224×224, but when resnet-18 is used as the backbone network, the last FC layer is removed. Only the output of the previous five stages is reserved as the input of feature enhanced network, that is, resnet-18 is taken as the backbone network of FCN, it can fully adapt to the input of different sizes. Figure 3 also roughly reflects the phenomenon that the higher the convolution level is, the more channels are, and the feature diagrams of the same stage have similar sizes. In order to clearly represent the characteristics of different sources, the second, third, fourth and fifth stages of backbone network output are collectively referred to as primary features, or original features; secondary features are derived from primary features; and tertiary features are derived from secondary features, which are identified by dashed frames and different colours.

In addition to the backbone network, CFPM network structure mainly includes two modules, Deconv module and Conv module. Deconv module is used to obtain secondary features. In order to reduce the number of channels in the high-level network output characteristic graph, each Deconv module in the graph contains 1×1 convolution and concatenate operations. The specific settings of Deconv modules are shown in table 1.
### Table 1. Detail setting of Deconv module.

| Conv Kernel  | Output Channel | Stride | Padding |
|--------------|----------------|--------|---------|
| 1\times1 conv | 128            | 1      | 0       |
| 4\times4 deconv | 128          | 2      | 1       |
| concat       | -              | -      | -       |
| 3\times3 conv | 128            | 1      | 1       |
| batchnorm    | 128            | -      | -       |
| relu         | 128            | -      | -       |

The five stages of Resnet-18 output feature map shapes are 1/2, 1/4, 1/8, 1/16 and 1/32 of the original image, corresponding channels are 64, 64, 128, 256 and 512. In this study, deconvolution is utilized for up-sampling. Compared with the traditional interpolation algorithm, deconvolution has more parameters and can continuously learn to optimize during the training process. Concat indicates that feature maps of the same size are stitched together by channels to form a feature map of more channels; to avoid the Checkerboard Artefacts, a convolution layer of 3\times3 is deployed after deconvolution to strengthen the connection between each value in the feature map and the surrounding area. The specific settings of Conv modules are shown in Table 2.

### Table 2. Detail setting of Conv module.

| Conv Kernel  | Output Channel | Stride | Padding |
|--------------|----------------|--------|---------|
| 3\times3 conv | 128            | 2      | 1       |
| concat       | 256            | -      | -       |
| 1\times1 conv | 128            | 1      | 0       |
| batchnorm    | 128            | -      | -       |
| relu         | 128            | -      | -       |

Conv module is simpler than Deconv module, the convolution core of 3\times3 reduces the size of the input signature map, concat operation combines the secondary feature map with the convoluted third feature map, and 1\times1 convolution is used to reduce the number of channels. The formalization output of a CFPM module is:

\[
C = \{f_1; f_2; f_3; f_4\}
\]

\(C\) represents the output of the CFPM, \(\{;;;\}\) is a collection of features with different sizes and \(f_i\) is a feature with different sizes. This feature map is the same size as the one using the FPN network. It ensures that the CFPM network structure can be used as a feature enhancement network to replace the FPN network, seamlessly linking up with subsequent segmentation heads proposed by other researchers, and has wider usage scenarios.

#### 2.2. Improved Differentiable Binarization Module

The IDB is inspired by the original DB module [16]. In the design of a segmentation head network with approximate differentiable binarization, three outputs are generated, the probability map \(P\), the threshold map \(T\) and the approximate binary map \(B\). When different supervisory information is used in training, the loss function of the IDB network structure is composed of three parts, which are recorded as \(L_p\), \(L_t\) and \(L_b\), respectively, so the loss function can be expressed as a formula (2):

\[
L = \alpha \times L_p + \beta \times L_t + \log_{10}(L_b)
\]
\( \alpha, \beta \) are weight factor set to 10 and 10, respectively, \( L_p \) adopted the Dice Loss and \( L_q \) adopted the Active Contour Loss function. \( L_q \) is computed as the sum of L1 distances between the prediction and label inside the dilated text polygon:

\[
L_q = \sum_{i \in R} |y^* - x^*|
\]  

(3)

AC loss is derived from ACWE (Active Contour without Edge) model. The ACWE model does not depend on the gradient of the image, the energy function of ACWE is defined by equation (4):

\[
F_1(C) + F_2(C) = \int_{\text{inside}(C)} |c_1 - I|^2 dxdy + \int_{\text{outside}(C)} |c_2 - I|^2 dxdy
\]

(4)

\( C \) represents the curve to be updated, \( F_1(C) \) and \( F_2(C) \) represents the energy value from inside and outside the curve, \( c_1 \) represents the average value of pixels within the curve, \( c_2 \) represents the average value of pixels outside the curve, and \( I \) represents the image to be segmented. Furthermore, the idea of level set method is introduced into ACWE model to solve the problem of curve evolution:

\[
F(C, c_1, c_2) = \mu \times \int_{\Omega} |\nabla H(\phi)| dxdy + \lambda \int_{\Omega} H(\phi) |c_1 - I|^2 dxdy + \lambda \int_{\Omega \setminus \Omega_C} H(-\phi) |c_2 - I|^2 dxdy
\]

(5)

\( \Omega \) represents the image area within the curve. For image semantic segmentation task, the design goal of AC loss is to find the global minimum energy of active contour model efficiently. The definition of AC loss is shown in equation (6):

\[
Loss_{AC} = \text{Length} + \lambda \times \text{Area}
\]

\[
= \int_C |\nabla u| ds + \lambda \int_{\Omega} ((c_1 - v)^2 - (c_2 - v)^2) u dxdy
\]

(6)

In the above formula, \( v \) and \( u \) respectively represent the mask of the prediction result and the groudtruth. The above equation is discretized as:

\[
Loss_{AC} = \text{Length} + \lambda \times \text{Area}
\]

\[
= \sum_{i=1,j=1}^{i=1,j=1} \sqrt{(\nabla u_{i,j})^2 + (\nabla u_{j,i})^2} + \varepsilon
\]

\[
+ \lambda \times \left( \sum_{i=1,j=1} \left( u_{i,j} - (c_1 - v_{i,j}) \right)^2 \right) + \left( \sum_{i=1,j=1} \left( 1 - u_{i,j} - (c_2 - v_{i,j}) \right)^2 \right)
\]

(7)

In the supervised learning paradigm, \( c_1 \) and \( c_2 \) represent the pixel mean values in the foreground and background respectively. Since the mask is used as the supervision label, the values can be simply taken as \( c_1 = 1 \) and \( c_2 = 0 \).

It can be found that using AC loss does not change the output of the original DB module, so the prediction result of IDB module is the same as that of DB module. In this paper, the CFPM-DB algorithm using the original DB module and the CFPM-IDB algorithm using the IDB module are compared.
3. Results and Discussion

3.1. Datasets
The results of proposed method are evaluated in two public datasets, ICDAR 2015 dataset and Total-text [17]. ICDAR 2015 dataset is released for ICDAR 2015 robust reading competition. All pictures are taken by Google glasses in the street or shopping centre. The text is in English, and the size is fixed at 1280×760. A total of 1500 images are included. The training set contains 1000 images, and the test set contains 500 images. Each picture gives the label of irregular quadrilateral and lists the coordinate position of four points clockwise. There are about 2000 recognizable word level text boxes in the whole data set, ignoring the labels that cannot be recognized or less than three characters. Due to the random pictures, there are some problems such as distortion, illumination, too small area of text area and contrast. This is the most frequently used dataset of text detection task. Total dataset was published on the basis of the traditional horizontal text and multi-directional text, the main feature of this dataset is curve text instance. The text contained in the data set is highly diversified in the direction. More than half of the images are combined in more than two directions. There are 1255 images in the training set and 300 pictures in the test set. Only English examples are included. The label in the dataset is the coordinates of the vertex of each polygon. The number of labels is related to the shape of each text area.

3.2. Implementation Details
For all models, we firstly adopted the Synth-Text dataset to pre-train, set the batch size 32, the optimizer SGD, the initial learning rate is set to 0.001 with the total 50000 epochs, the weight decay is set to 4×10^-4, and nesterov momentum is set to 0.99. In order to compare with the original DB algorithm, CFPM feature extraction module and IDB module are not used in the pre-training on Synth-Text dataset.

The data augmentation for the real-world training data includes: (1) Flip randomly. It is divided into random up and down flip and random left and right flip. (2) Random rotation. With the same parameters used by other algorithms, the associated text regions in the original image are randomly rotated between [-10° and 10°] with the image centre as the origin. The blank part produced by image rotation is filled with 0. (3) Random cutting and size adjustment. In order to make full use of these positive samples, it is a natural way to keep and enlarge the text area after cutting most of the images. In this study, we only used a single size 736×736. (4) Contrast adjustment. From the ICDAR 2015 dataset, it can be found that some text regions are not easily identified due to the interference of illumination. The contrast adjustment can enhance the robustness of the model in different contrast images.

In table 3, EAST, as the representative of regression-based text detection method, received much attention when it was published in 2017. As a simple and classical algorithm, EAST was introduced as text detection algorithm by OpenCV since version 3.4.2. Seglink and PixelLink are the representatives of two bottom-up text area construction algorithms. When facing multiple shape text areas, they have finer expression and more reliable text boundaries. However, these algorithms have the drawbacks of cumbersome and time-consuming post-processing steps. When facing fixed shape text area detection tasks, the accuracy is improved but the speed is reduced, and cannot handle curved text instances. CFPM-IDB performs better than the original two methods, although increasing the model volume, it is more accurate. The results in table 4 are similar with results in table 3.
Table 3. Detection results on the ICDAR 2015 dataset.

| Method      | Precision/% | Recall/% | H-mean/% | FPS |
|-------------|-------------|----------|----------|-----|
| EAST [18]   | 78.3        | 83.3     | 80.7     | 13.2|
| Seglink [15]| 76.8        | 73.1     | 75.0     | -   |
| PixelLink [11]| 82.9       | 81.7     | 82.3     | 7.3 |
| PSENet [12] | 86.9        | 84.5     | 85.7     | 1.6 |
| DB-18       | 86.8        | 78.4     | 82.3     | 48  |
| CFPM-DB     | 88.2        | 80.1     | 83.9     | 32  |
| CFPM-IDB    | 88.6        | 81.4     | 84.8     | 32  |

Table 4. Detection results on the Total-text dataset.

| Method      | Precision/% | Recall/% | H-mean/% | FPS |
|-------------|-------------|----------|----------|-----|
| EAST [18]   | 50.0        | 36.2     | 42.0     | -   |
| PSE Net [12]| 84.02       | 77.96    | 80.87    | -   |
| CSE [19]    | 81.4        | 79.7     | 80.2     | -   |
| LOMO MS [20]| 87.6        | 79.3     | 83.3     | -   |
| DB-18       | 88.3        | 77.9     | 82.8     | 50  |
| CFPM-DB     | 89.6        | 79.4     | 84.0     | 33  |
| CFPM-IDB    | 88.8        | 81.2     | 84.8     | 33  |

4. Conclusions
In this paper, we propose a new feature enhancement network structure suitable for text detection and improve a semantic segmentation network with AC loss. Experiments are carried out on quadrilateral text datasets ICDAR 2015 and curved text datasets Total-text respectively to prove the effectiveness of the proposed method. The CFPM-IDB method achieves a good balance in accuracy and speed. In addition, the proposed method can easily replace the existing methods to reduce the additional adaptation works.

References
[1] Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. 2015 Competition on robust reading 2015 13th International Conference on Document Analysis and Recognition (ICDAR) (IEEE) pp 1156-1160.
[2] Liu W, Anguelov D, Erhan D, et al. 2016 SSD: Single shot multibox detector European Conference on Computer Vision (Cham: Springer) pp 21-37.
[3] Redmon J and Farhadi A 2018 Yolov3: An incremental improvement arXiv preprint arXiv:1804.02767.
[4] Ren S, He K, Girshick R, et al. 2015 Faster R-CNN: Towards real-time object detection with region proposal networks Advances in Neural Information Processing Systems pp 91-99.
[5] Kim K H, Hong S, Roh B, et al. 2016 Pvanet: Deep but lightweight neural networks for real-time object detection arXiv preprint arXiv:1608.08021.
[6] Liao M, Shi B, Bai X, et al. 2017 Textboxes: A fast text detector with a single deep neural network Thirty-First AAAI Conference on Artificial Intelligence.
[7] Liao M, Zhu Z, Shi B, et al. 2018 Rotation-sensitive regression for oriented scene text detection Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 5909-5918.
[8] Zhang Z, Zhang C, Shen W, et al. 2016 Multi-oriented text detection with fully convolutional networks *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 4159-4167.

[9] He K, Gkioxari G, Dollar P, et al. 2017 Mask R-CNN *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1-1.

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

[11] Deng D, Liu H, Li X, et al. 2018 Pixellink: Detecting scene text via instance segmentation *Thirty-Second AAAI Conference on Artificial Intelligence*.

[12] Wang W, Xie E, Li X, et al. 2019 Shape robust text detection with progressive scale expansion network *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 9336-9345.

[13] Tian Z, Huang W, He T, et al. 2016 Detecting text in natural image with connectionist text proposal network *European Conference on Computer Vision* (Cham: Springer) pp 56-72.

[14] Hu H, Zhang C, Luo Y, et al. 2017 Wordsup: Exploiting word annotations for character based text detection *Proceedings of the IEEE International Conference on Computer Vision* pp 4940-4949.

[15] Shi B, Bai X and Belongie S 2017 Detecting oriented text in natural images by linking segments *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 2550-2558.

[16] Liao M, Wan Z, Yao C, et al. 2020 Real-time scene text detection with differentiable binarization *Thirty-four AAAI Conference on Artificial Intelligence* pp 11474-11481.

[17] Ch’ng C K and Chan C S 2017 Total-text: A comprehensive dataset for scene text detection and recognition *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (IEEE) pp 935-942.

[18] Zhou X, Yao C, Wen H, et al. 2017 East: An efficient and accurate scene text detector *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* pp 5551-5560.

[19] Liu Z, Lin G, Yang S, et al. 2019 Towards robust curve text detection with conditional spatial expansion *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 7269-7278.

[20] Zhang C, Liang B, Huang Z, et al. 2019 Look more than once: An accurate detector for text of arbitrary shapes *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 10552-10561.