FMN: Two-stage web image fine-grained text detection method based on full convolutional network

Jiajie Fan¹, Yongli Wang, Shurong Sun²

¹School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
²Zhenjiang Nalan Sisi Information Technology Co., Ltd. Zhenjiang 212100)
yongliwang@njust.edu.cn

Abstract. In the Internet, there are large-scale web images with different image sizes and different resolutions. In order to accurately and efficiently identify fine-grained texts in web images, this paper uses FCN to support the ability to semantically segment images, and treats text and background as different detection targets. A two-stage web image fine-grained text detection method FMN is proposed. For the problem that the output result of the whole convolutional network is not fine enough, the MSER+NMS algorithm is used to extract the fine-grained features of the FCN output, and the ellipse fitting method is introduced to detect the text. Features such as character tilt and morphological differences ultimately result in text detection of the image. Compared with the existing methods, the FMN algorithm proposed in this paper has achieved certain improvement in detail resolution and precision.

1. Introduction
In today's Internet era, web composite images are an important medium for transmitting information, and each image either contains complex typography, or contains intensive small text or multiple languages, or contains a watermark background. These problems pose new challenges to the text detection and recognition of images, mainly reflected in the following aspects:

1) The universality of the model, the model can be highly robust in complex layout, multi-word direction and size, watermark text, etc.

2) Fine-grained, for the text detection method, as far as possible, separate a single character independent area instead of the entire text rectangular frame, which is convenient for segmenting a single character for subsequent text recognition.

This paper proposes a FMN text detection method based on full convolutional network. The main contributions are as follows:

1) The Oval_MSER and Oval_NMS algorithm is used to detect the text area, and more image edge detail information is obtained, thereby improving the detection accuracy.

2) the ellipse fitting method is used to obtain not only the size information of the region, but also the tilt angle of the text, so that the detection result is more accurate and the segmentation is better. Character area.

2. Related work
At present, in the text detection method based on deep learning, the method based on text region suggestion and image semantic segmentation is the most widely used.
The text detection method based on image segmentation \cite{1, 2, 3} regards text detection as a generalized "segmentation problem". This type of method usually use the full convolutional network (FCN) to perform pixel-level text/background annotation. For the first time, Zhang Z et al\cite{1} used a Fully Convolutional Network (FCN) to process images from the pixel layer. Zhou X et al\cite{2} proposed a simple and efficient text detection framework based on Fully Convolutional Network and Non-maximum suppression (NMS), which firstly passes full convolution.

The semantic segmentation method can better avoid the text arrangement direction and the text area aspect ratio, however, the subsequent processing of this kind of method is more complicated, which leads to the low robustness of the detection method. Therefore, improving the robustness of the post-processing stage is the difficulty based on the semantic segmentation method.

3. FMN two-stage image text detection model

The FMN text detection model is divided into two stages to detect the position of the text in the image. The first stage: the original image to be detected is input into the FCN network for forward calculation to obtain the output image of the network; the second stage: the original image and the output image are further detected by the Oval_MSER and Oval_NMS algorithms to obtain the final text position.

Figure 1. .FCN-MSER-NMS algorithm flow chart

3.1. FCN text prediction network introduction

The FCN network replaces the fully connected layer of the convolutional neural network with a deconvolution layer, and finally generates a predicted image by upsampling. Its network structure is shown in Figure 2:

Figure 2. FCN network structure diagram.

Figure 3. Original image a, marker image b, text prediction image c output by FCN text network.
3.2. FMN algorithm introduction

Algorithm 1: FCN-MSER-NMS text detection algorithm

| Input | text image to be detected img, FCN network output image Pre_image |
|-------|---------------------------------------------------------------|
| Output| Image class text area collection Q                           |

1) 1) Using the MSER to obtain a set of predicted text regions \( Q \{ (x_{up}, y_{up}, x_{dn}, y_{dn}) \} \) from Pre_image
2) \( K = \emptyset \)
3) For \( q_m \) in \( Q \):
4) \( Q = Q - q_m \)
5) Perform Oval\_MSER detection on the \( q_m \) area of the original image to obtain Oval\_Set
6) Use Oval\_NMS to merge the extra detection frames in Oval\_Set, and eliminate them to get set \( K \).
7) \( K = K + \text{Oval\_NMS(Oval\_Set)} \)
8) End for
9) Return \( K \)

The nouns appearing in the algorithm give definitions and theoretical basis:

Definition 1: Aspect Ratio
For the pixel point coordinates \((i, j)\) in the extreme value stable region \( V_i \), \( w = \max(i) - \min(i) \) denote the width of the extreme stable region, \( h = \max(j) - \min(j) \) represents the length of the extreme stable region, and the aspect ratio can be expressed as

\[ e = \frac{w}{h} \quad (1) \]

Definition 2: Boundary Density of Text Images
Define the boundary density to represent the boundary pixel intensity of the extreme stable region

\[ \rho_{area\_edge}(x, y) = \frac{1}{\text{width} \times \text{height}} \sum_{i=0}^{\text{width}} \sum_{j=0}^{\text{height}} f(x, y) \quad (2) \]

Where \( f(x, y) \) represents the boundary image of the region, and height, width respectively represents the height and width of the stable region.

\[ f(x, y) = \begin{cases} 1 & (x, y) \in area\_edge \\ 0 & \text{else} \end{cases} \quad (3) \]

Definition 3: Over\_Fit (ellipse fit)
Take the pixel point set \( \{(x_i, y_i)\}_{i=1,2,3...N} \) of the fitted region; let the elliptic equation be

\[ Ax^2 + Bxy + Cy^2 + Dx + Ey + 1 = 0 \quad (4) \]

The optimal elliptic equation can be found by least squares:

\[ F(x, y) = Ax^2 + Bxy + Cy^2 + Dx + Ey + 1 \]

\[ \text{min:} D(x, y) = \sum_{i=1}^{N} F(x_i, y_i)^2 \quad (5) \]

Solve the above formula using the least squares method, Calculate the ellipse parameters:

\[ x_0 = \frac{B + D - 2 + AE}{4 + C + E - D^2}, \quad y_0 = \frac{A + D - 2 + BE}{4 + C + E - D^2}, \quad \theta = \frac{1}{2} \tan^{-1} \left( \frac{D}{C - E} \right), \quad a = \frac{2(C + x_0^2 E + y_0^2 D + x_0 y_0 - 1)^2}{C + E \sqrt{(C - E)^2 + D^2}}, \quad b = \frac{2(C + x_0^2 E + y_0^2 D + x_0 y_0 - 1)}{C + E \sqrt{(C - E)^2 + D^2}} \]
Algorithm 2: Oval_MSER algorithm in FMN.

Input : image img to be detected and class of text area S
Output : Elliptical text box collection P

1) For $S_i$ in $S$
2) MSER detection on $S_i$ to obtain the set of extreme value stable regions Set$_{S_i}$:
3) for oval in Set$_{S_i}$:
4) /* Use Definition 1 and Definition to filter out background regions */
   if: $0.4 < \text{oval}[i].e < 2.5.$ and $\rho(\text{oval}[i]) > 0.2$:
   5) continue
   6) else
   7) Set$_{S_i}$ = Set$_{S_i}$-oval
   8) end if
   9) end For
10) End for

Algorithm 3: Oval_NMS algorithm in FMN.

Input : the set of elliptical text boxes to be detected $P\{(x_m, y_m, a_m, b_m, \theta_m)\}$
Output : final text box collection $K$

1) Convert the elliptical box set $P$ into a rectangular box set $B$
2) while $B! = \emptyset$: /* Converting the area box in $P$ to a rectangular moment text box set $B$ */
3) The detection frame $p_m$ closest to the origin $(0,0)$ of the ellipse center point
4) $B = B - p_m$ /* Remove the current detection frame $p_m$ from the collection */
5) for $p_n\{(x_n, y_n, a_n, b_n, \theta_n)\}$ in $B$:
6) Overlap = IOU((area($p_m$),area($p_n$))
7) if overlap > overlapThresh and $\|\theta_m - \theta_q\| < \pi/8$:
8) $P = P - p$ /* Remove the current detection frame $p$ from the set $P$ */
9) $K = K + p_m$
10) End if
11) End for
12) End while
13) Return $K$

4. Experiment

4.1 Data Set
1) MTWI-2018\cite{7}
   The experimental data set is the Chinese network picture data set MTWI-2018 under the advertising scene, which is the official data set of Ali Tianchi image text detection competition.
2) MSRA-TD500\cite{8}
   The image text area in the MSRA-TD500 data set has 500 pictures., and most of the text is distributed on the roadside guide cards. and the text includes Chinese, English, and numbers.
3) IDCAR-2015\cite{9}
   IDCAR-2015 is a new set of scene-based databases published by the International Document Analysis and Recognition Conference in 2015.
4.2. Experimental measures
Ground-truth is calculated by the method shown in Fig. 4, where G and D are respectively represented as Ground-truth and test results. Considering that it is not convenient to calculate the coverage between G and D, rotate G and D according to their center points $C_G$, $C_D$ to the position shown by $G'$, $D'$ (as shown in Figure 4).

\[ m(G, D) = \frac{A(G' \cap D')}{A(G' \cup D')} \]  

$A(G' \cap D')$ and $A(G' \cup D')$ respectively represent the intersection and union regions between $G'$ and $D'$, and draw on the PASCAL target detection performance evaluation method [8].

Accuracy (P) and recall rate (R) of text detection results and the F-score:

\[ P = \frac{|T_P|}{|E|}, \quad R = \frac{|T_P|}{|T|}, \quad F = \frac{2 \times P \times R}{P + R} \]  

$T_P$, $E$, $T$ respectively represent the correct set of test results, the set of test results and the Ground-truth set.

4.3. Experiment results

### Table 1. Experiment Results on Data Sets

| Method          | MTWI 2018 Precision | MSRA-TD500 Precision | ICDAR R2015 Precision | MTWI 2018 Recall | MSRA-TD500 Recall | ICDAR 2015 Recall | MTWI 2018 F-score | MSRA-TD500 F-score | ICDAR 2015 F-score |
|-----------------|---------------------|-----------------------|-----------------------|-----------------|------------------|------------------|------------------|------------------|------------------|
| FCN             | 0.58                | 0.83                  | 0.71                  | 0.43            | 0.67             | 0.43             | 0.49             | 0.74             | 0.54             |
| FCN-NMS         | 0.32                | 0.71                  | -                     | 0.37            | 0.61             | -                | 0.39             | 0.66             | -                |
| MSER-NMS        | 0.65                | 0.88                  | 0.85                  | 0.58            | 0.79             | 0.80             | 0.61             | 0.83             | 0.82             |
| SSD             | 0.58                | 0.87                  | 0.83                  | 0.54            | 0.67             | 0.78             | 0.56             | 0.76             | 0.80             |
| RRD             | 0.63                | 0.87                  | 0.85                  | 0.55            | 0.73             | 0.79             | 0.58             | 0.74             | 0.82             |
| MCN             | 0.61                | 0.88                  | 0.72                  | 0.57            | 0.79             | 0.80             | 0.59             | 0.83             | 0.76             |

4.4. Experiment Results on Data Set

![Image of Oval MSER and Oval NMS results for class text regions](image)

5. Conclusion
The FMN algorithm is used for scene text detection. The MSER ellipse fitting algorithm is introduced in the post-processing stage of the full convolutional network output to solve the non-vertical level of the image text area; and the Oval_NMS algorithm based on the fitting result is used to eliminate the...
extra bounding box. The result of fine-grained segmentation is achieved.

ACKNOWLEDGMENT
This article has been awarded by the Fundamental Research Fund for the Central Universities (30918015103, 30918012204), Nanjing Science and Technology Development Plan Project (201805036), and "13th Five-Year" equipment field fund (61403120501), China Academy of Engineering Consulting Research Project(2019-ZD-1-02-02).

References
[1] Zhang Z, Zhang C Q, Shen W, Yao C, Liu W Y, Bai X (2016). Multi-oriented text detection with fully convolutional networks. Conference on Computer Vision and Pattern Recognition. 4159-4167
[2] Zhou X Y, Yao C, Wen H, Wang Y Z, Zhou S C, He W R et al. (2017) East: an efficient and accurate scene text detector. Conference on Computer Vision and Pattern Recognition. 2642-2651
[3] He W H, Zhang X Y, Yin F, Liu C L. (2017) Deep direct regression for multi-oriented scene text detection. In: The IEEE International Conference on Computer Vision
[4] Deng D, Liu H F, Li X L, Cai D.(2018) PixelLink: detecting scene text via instance segmentation. In: The IEEE International Conference on Computer Vision
[5] He T, Huang W L, Qiao Y, Yao J.(2016) Accurate text localization in natural image with cascaded convolutional text network. In: The IEEE International Conference on Computer Vision
[6] Shi B G, Bai X, Belongie S.(2017) Detecting oriented text in natural images by linking segments. The 2017 IEEE Conference on Computer Vision and Pattern Recognition. 3482-3490
[7] MTWI-2018(2018) https://tianchi.aliyun.com/competition/entrance/231685/information
[8] Everingham M, van Gool L, Williams C K I, Winn J, Zisserman A.(2010) The pascal visual object classes (VOC) challenge. International Journal of Computer Vision. 88(2): 303-338
[9] Yin X C, Yin X W, Huang K Z, Hao H W.(2014) Robust text detection in natural scene images. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(5): 970-983
[10] Liao M, Zhu Z, Shi B, et al.(2018) Rotation-Sensitive Regression for Oriented Scene Text Detection[J]. Conference on Computer Vision and Pattern Recognition[C] (CVPR).
[11] Liu Z, Lin G, Yang S, et al.(2018) Learning Markov Clustering Networks for Scene Text Detection IEEE/CVF Conference on Computer Vision and Pattern Recognition[C] (CVPR).