Port container number detection based on improved EAST algorithm

Port container number detection based on improved EAST algorithm

XingQi Feng[1*], QingLiu[2] and ZhiWei Wang[3]

[1]School of Automation, Wuhan University of Technology, Wuhan, Hubei, China

[2]School of Automation, Wuhan University of Technology, Wuhan, Hubei, China

[3]School of Automation, Wuhan University of Technology, Wuhan, Hubei, China

[*]Corresponding author's e-mail: fxq1067544213@163.com

Abstract: The methods of intelligent port container number detection tend to be diversified. However, traditional container number positioning methods and current deep learning algorithms are mostly multi-stages, which need to be optimized during difficultly training, resulting in bad model effect and time-consuming. In view of the above problems, the main contributions of this paper are drawing on the EAST text detection method to obtain an improved EAST algorithm that directly predicts the box number area. This new algorithm not only eliminates multiple intermediate stages, but also improves the regression method and loss function of the box number area, optimizes the boundary, balances positive and negative samples to adapt to the detection area of the box number based on the original algorithm. At the same time, it adopts the lightweight design idea, and uses ShuffleNet's channel shuffling and depth separable convolution for reference to optimize the original model, which reduces the complexity of time and space, compresses the model and reduces the detection time. The final experimental results show that the accuracy of the algorithm reaches 97.5% while keeping the FPS index no less than 14.

1.Introdution

At present, the detection of container quantity is an important link in the construction of intelligent port. Among them, as the premise of follow-up work, container number positioning plays an important role in the whole process of container number text information extraction and recognition. Previous text detection algorithms are used for target detection. Good results have been achieved at different levels. Its core is to design features that can distinguish text from background.

For example, on the basis of traditional image processing algorithms, features are designed manually to capture scene text features. Early scholars such as Zhuo Junfei proposed using the edge features of the target area and vertical projection method for positioning[1]. However, due to the detection of regional pollution and environmental shadow, the target area is not accurate; later, Wan Yan et al. Another detection method based on color features is proposed[2], which uses rough texture information to detect target candidate regions, and then uses three channels with different RGB values to detect. Clustering the same color of different colors, combining with connectivity analysis to locate the target area accurately. However, the gray value of color is greatly affected by weather and light, so it is not reliable in practical application. The traditional methods have their own advantages and disadvantages, but in the application of box number detection, due to the influence of environmental

shadow, body color, detection area pollution, irregular box number printing and other factors, the traditional method is not ideal.

Algorithms based on deep learning learn effective features directly from training data. Although there are many deep learning algorithms in the field of text detection[3-6], they have some disadvantages over traditional algorithms, such as RCNN, Fast RCNN, SSD and other mainstream algorithms[7-10], including candidate box extraction and candidate box extraction which may be a suboptimal and time-consuming stage and step resulting poor effect of text detection. Therefore, the accuracy and efficiency of these methods can not meet the current situation. Therefore, the model based on EAST framework solves these shortcomings[11].EAST does not generate any candidate boxes, only including the convolution network and NMS merging stage. NMS (non maximum suppression), that is to remove the redundant bounding box, and only keep the highest overlapping bounding box with the ground truth. The network returns to the polygon positioning box, which can locate the inclined text at any angle. This method is used in container number detection and provides a new application for intelligent identification of container number in the future.

The procedure of the whole box number detection algorithm is to roughly locate the whole box number area, then fine locate the rough detection area, and finally complete the combination of character lines to lay a solid foundation for the subsequent recognition. Sometimes it is necessary to perform a padding process with a rough positioning box number picture to fit the network, as shown in Figure 1.

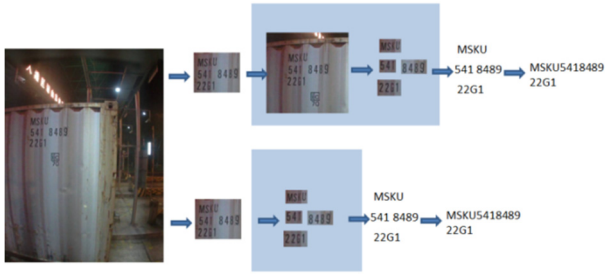

Figure 1. Container localization process

2.EAST model structure and optimization

2.1.Original EAST model structure

The EAST network[12] is divided into three parts: feature extraction layer, feature fusion layer and output layer, as shown in Figure 2:

Feature extraction layer: The backbone uses VGG-16 for feature extraction, and extracts feature maps(expressed as f) at different levels to obtain feature maps at different scales. The purpose is to solve the problem of sharp changes in the scale of text lines. Large-size layers can be used to predict small text lines. Small size layers can be used to predict large text lines.

● Feature merging layer, which merges the extracted features. The merge rule adopts the U-net method, and the top features of the network are extracted from the features downward according to the corresponding rules. Merging method:

$$g_i = \begin{cases} unpool(h_i), \mathrm{i} \le 3 \\ conv_{3\times3}(h_i), \mathrm{i} = 4 \end{cases} \tag{1}$$

$$h_i = \begin{cases} f_i, \mathrm{i} = 1 \\ conv_{3\times3}(conv_{1\times1}([g_{i-1}; f_i])), otherwise \end{cases} \tag{2}$$

● Network output layer: The final output of the network has 3 major parts, which are: score map: a parameter that represents the confidence of this prediction box; text boxes: 4 parameters, (x, y, w, h),

which are the same as ordinary bounding box parameter of the target detection task; text rotation angle: 1 parameter, which indicates the rotation angle of the text boxes.

● Loss aspect: L_s and L_g respectively indicate whether there is text (score in the pixel map) and the loss of IOU and geometry map, λ_g represents the importance between the two losses. In the original experiment, λ_g is set to 1. The specific calculation method of this value will be given later.

$$L = L_S + \lambda L_g \tag{3}$$

$$L_g = L_{AABB} + \lambda_\theta L_\theta \tag{4}$$

Among them, L_{AABB} represents the loss value between the four vertex coordinates and the true value coordinates of the predicted text box. Similarly, L_θ is the loss value of the text rotation angle and the true value rotation angle of the predicted text box.

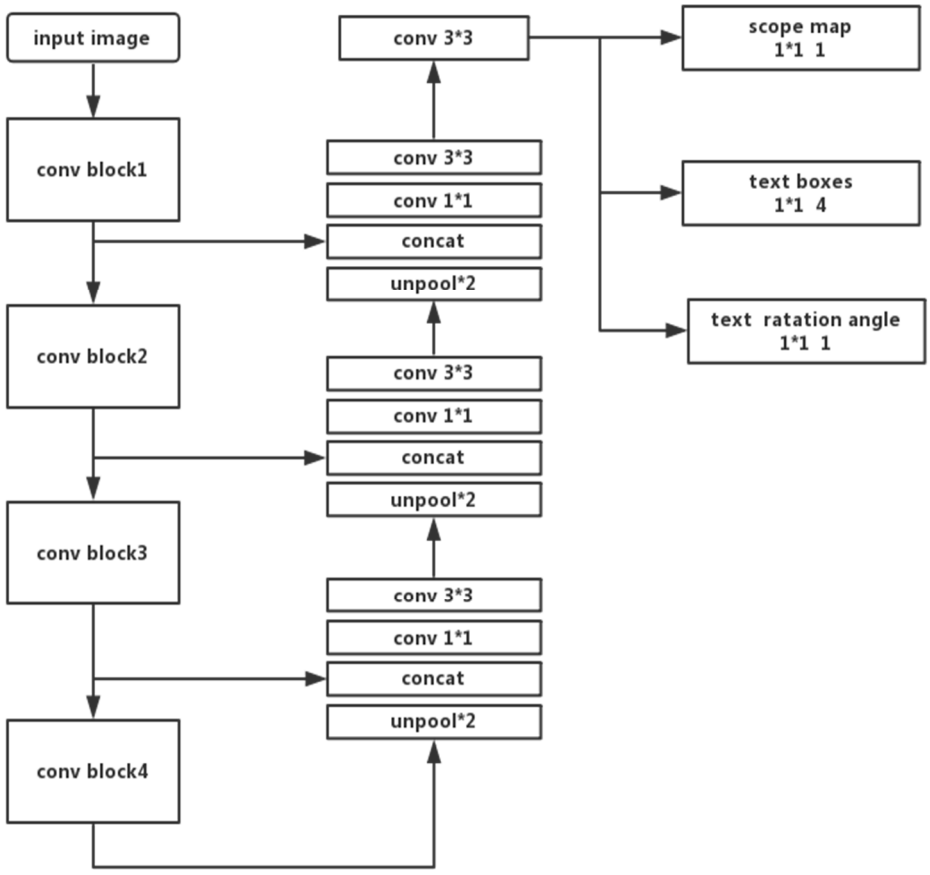

Figure 2. Original EAST model structure

2.2.Optimization of output layer regression

After rough positioning of the sample, four different types of box number areas can be obtained, as shown in Figure 3. It can be seen from the Figure that the text line of the horizontal box number is shorter, but the text line of the vertical box number is long. After verification, it is found that the original EAST algorithm has a poor detection effect on long texts[13-15], so the optimization algorithm improves the ability to predict long texts. Therefore, the output layer structure of EAST is modified, and one channel predicts whether the pixel is in the text box; Two channels predict whether the pixel is the head pixel or the tail pixel of the text box; four channels predict the offset of the two vertex coordinates corresponding to the head pixel or the tail pixel, as shown in Figure 4(a).

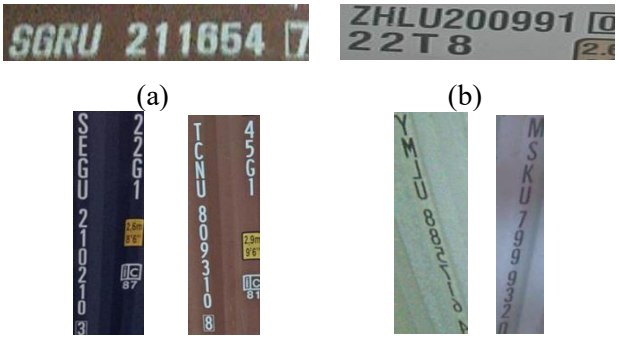

Figure 3. Different arrangement of container number:
(a) Horizontal Single-line characters (b) Horizontal double-line Character (c) Vertical double-line characters (d) Vertical treble-line characters

After optimization, the output layer is a 1-bit score map, indicating whether it is in a text box; 2-digit vertex code, whether it belongs to the box pixel box border pixels and whether it is the head or tail; 4-bit geo is a border pixel can predict the coordinates of 2 vertices. All pixels form the box number box shape, and then use only the weighted average of all boundary pixel predictions to predict the two vertices at the ends of the short side of the head or tail.4 vertex coordinates. The polygon box drawn directly based on these 4 point coordinates is the positioning box. This regression method solves the problem that the receptive field of the original algorithm is not large to a certain extent, and makes the effect of long box number detection better.

Specifically, after feature extraction and feature fusion, the output of the feature map is 7 channels, and the output value of the first three channels is the probability value. Whether the conditions meet the requirements depends on the threshold value set by yourself. The last four channel values are the coordinates of the predicted boundary vertices. If they are head pixels, the coordinates of the two vertices of the head boundary are predicted; if they are tail pixels, the coordinates of the two vertices of the tail boundary are predicted.

2.3.Lightweight design of backbone network

In order to improve the accuracy of East detection, the residual network of RESNET is used to replace the VGg network in order to deepen the depth of the network. However, the original residual structure has high time and space complexity in the network. Even if the general lightweight network is introduced, the 1×1 volume in the depth separable convolution takes up a lot of calculation, and makes the channels full of constraints, to a certain extent, reduces The accuracy of the model and the overall network are also inefficient due to the intensive 1×1 convolution, which cannot fully integrate the characteristic channels. Therefore, this paper uses the channel shuffle strategy of ShuffleNet[16] for reference, and then combines the depth separable convolution to build a lightweight East network. The overall structure is shown in Figure 4.

It can be seen from the figure that the channel shuffle structure is mainly designed in the feature extraction layer, and the deep separable convolution is used in the feature fusion stage, but one thing is more important: Although the deep separable convolution can effectively reduce the calculation amount, its storage and access efficiency is poor, so the first convolution does not use the basic unit of ShuffleNet, but uses the conventional convolution, and only in the subsequent stage can it be used The lightweight strategy of channel shuffle + depth separable convolution is used.

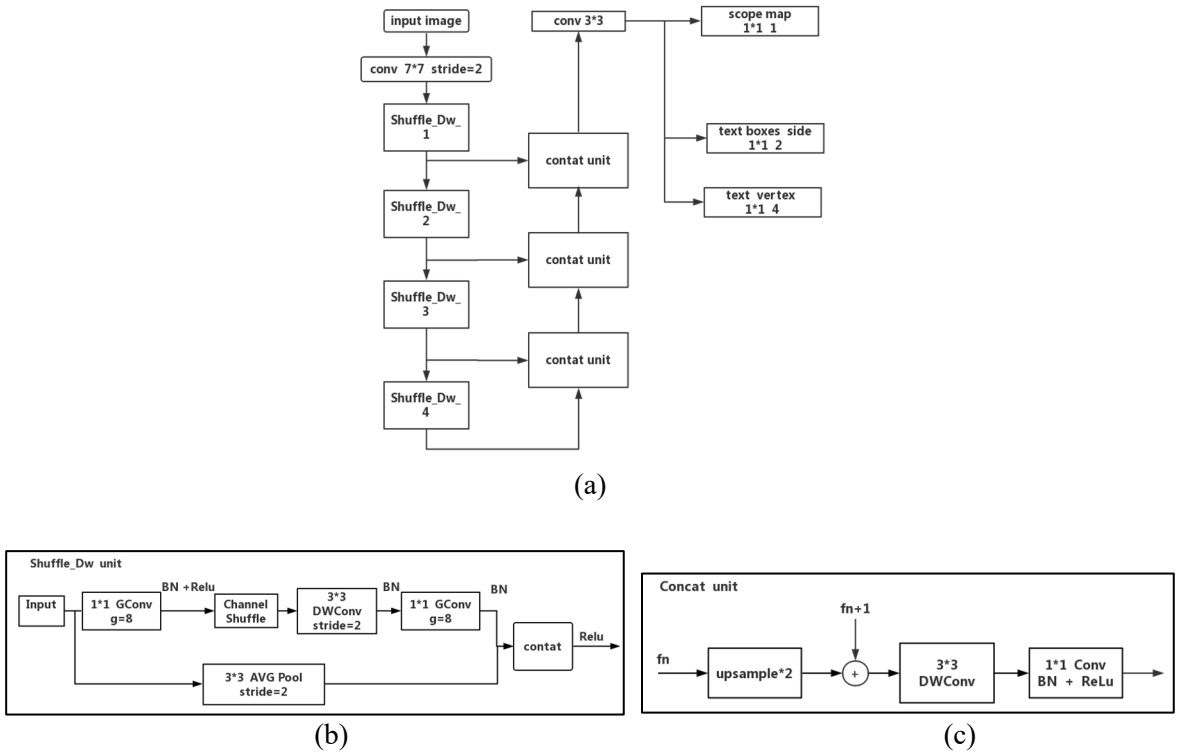

Figure 4. (a) The overall structure of the improved EAST (b)The shuffle_Dw unit structure.(c)Concat unit structure

2.4.The improvement of loss

The improved EAST network uses dice loss, but in the case number sample picture, the proportion of the case number text to the whole picture is very different (as shown in Figure 3, some text proportion is larger than the sum of other real samples), resulting in uneven distribution of positive and negative samples. Although background rate is introduced in the original EAST to limit the proportion of positive and negative samples, the value is still set by human, which has limited effect on samples with large box number difference and cannot take all samples into account.

$$L_{AABB} = \text{dice}\_loss = \frac{(2 \times y_{ture} \times y_{pre})}{y_{ture} + y_{pre}} \tag{5}$$

$$L_g = L_{AABB} + L_\theta \tag{6}$$

$$K_i = \frac{S_i}{S} \tag{7}$$

$$L_{AABB} = \frac{\sum_{x \in S} IOU(y_{true}, y_{pre}) \times y_{trainmask}}{K_i} \tag{8}$$

Therefore, this paper refers to the instance balance method of pixelLink[17-19] to balance the weight of small area and large area samples. The method is: set the same weight K_i for all case number areas of batch size samples.

K_i and K are respectively the number of positive samples and the total number of pixels in a batch. IOU(intersection over Union) in the formula is a concept used in target detection, which is the overlapping rate of the generated candidate box and the original marker box, that is, the ratio of their intersection and union.

From formula (5-8), it can be seen that when the number of samples is large, the weight will be suppressed. When the number of samples is small, the weight of small sample area will be relatively larger, while the weight of large text area will be relatively smaller, which is more conducive to the detection and detection of samples. The function of $Y_{trainmask}$[20-21] is to remove the frame whose length or width is less than the threshold value as a difficult sample, and also to remove the interference of negative samples and improve the positioning effect significantly.

2.5.Optimization of Box Number Boundaries

In order to improve the detection effect of the box number boundary, another threshold L_{lim} is set after when the dice loss is less than the threshold value. The purpose is to add the distance from the point to the border for learning after the initial accuracy is achieved, so that the loss of the center is relatively reduced, and the loss of the edge is relatively increased, so as to optimize the boundary. The overall formula is (10).

$$L_{\mathrm{AABB}} = -(\lambda(1-\mathrm{y}_{true})\times\log(1-sigmoid(y_{pre}))+ \frac{(1-\lambda)\times\mathrm{y}_{true}\times\log(sigmoid(y_{pre}))}{m\_dis\tan ce}) \tag{9}$$

$$L_{\mathrm{AABB}} = \begin{cases} \mathrm{dice}\_loss = \dfrac{(2\times y_{ture}\times y_{pre})}{y_{ture}+y_{pre}}, & L > L_{\lim} \\ -(\lambda(1-\mathrm{y}_{true})\times\log(1-sigmoid(y_{pre}))+ \dfrac{(1-\lambda)\times\mathrm{y}_{true}\times\log(sigmoid(y_{pre}))}{m\_dis\tan ce}), & L \le L_{\lim} \end{cases} \tag{10}$$

λ is the average value of y_{true} as the coefficient of cross entropy function, and m_distance is the minimum distance between pixels and four sides of the frame. By introducing this parameter, the center of gravity of loss can be distributed to the boundary. Through continuous learning, the boundary can be effectively optimized. In this paper, the value of L_{lim} is not more than 1e-3.

After the improvement in the above three aspects, the flowchart of the training and testing of the entire box number positioning algorithm is shown in Figure 5.

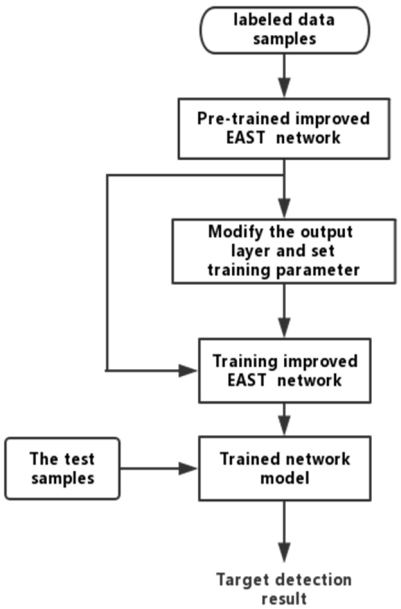

Figure 5. EAST Training process flow diagram

3. Experimental results and analysis

The model of this box number positioning was built under the framework of Tensorflow 1.14.0, and was trained using Window 10 operating system with AMD Ryzen 7 3750H and 1660ti graphics card. 26,000 labeled image samples were produced, of which 20,000 were the training data set. 6000 test sets. Trained between the original EAST algorithm and the improved algorithm in this article. The test results are shown in Table I. It can be seen that based on the improved EAST network model, while maintaining a similar iteration time, the same labeling method is used. By modifying the output layer of the network, improving the coordinate regression method, using the channel shuffling and depth separable convolution method of ShuffeNet to import the network and optimize the loss function, the accuracy of the improved model can reach 97.5%. The positioning effect of the whole box number is shown in Figure 6 and Figure 7.

mAP (mean average precision) is the average precision sum of all categories divided by all categories, that is, the average precision of all categories in the data set.

FPS refers to the number of images (frames) recognized in one second.

Table 1 Comparison results of the same test set

Optimization Method	Samples numble	Iterations	mAP	FPS
Original algorithm	6000	50000	85	14
Only regression	6000	50000	89	14.5
Original algorithm+lightweight	6000	50000	84	20.5
regression method + Instance balance	6000	50000	92	14.2
regression + Instance balance + number boundary	6000	50000	97.5	14
regression + Instance balance+ number boundary+lightweight	6000	50000	96	19

Figure 6. Vertical treble-line/double-line characters positioning effect

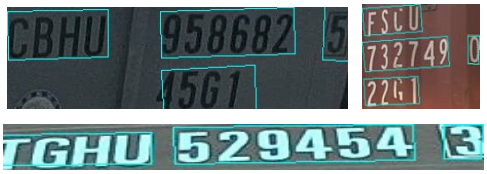

Figure 7. Horizontal Single-line/double-line characters positioning effect

In order to reflect the superiority of the model in this paper, Table II shows the accuracy and time consumption of other algorithms in detecting the box number area. After comparison, it can be seen that the SVM detection algorithm based on Hog features takes time Seriously, because the box numbers are densely distributed, and there are interferences of other characters around the same character block, when these interference character regions are trained as negative samples, the accuracy of the container number region will be greatly reduced, which leads to the robustness of the algorithm. Poor. The accuracy of the box number positioning based on YOLO in the early days was only 71%, because the edge characters of the box number are not in the detection box predicted by the model. This is because the YOLO algorithm often performs poorly when detecting objects close to each other.

Although the SSD-based algorithm has been greatly improved compared to the previous one, the positioning effect has a great relationship with the way of labeling the box number data set, the anchor value parameter settings of the network framework, the training parameters, etc. Moreover, the coordinates returned by the algorithm It usually forms a rectangular frame, which has a poor positioning effect on inclined box numbers.

Table 2 Comparison results with other algorithms

Methods	mAP	FPS
Hog+SVM	68	2
YOLO	71	10
SSD	78	11
Ours: regression + Instance balance+ number boundary	97.5	14
Ours: regression + Instance balance+ number boundary+lightweight	96	19

From Table 1, it can be seen that the speed of the classic lightweight EAST algorithm is the fastest, but the box number accuracy is not high. After changing the output layer regression method and making improvements for loss, the accuracy of detection is greatly improved, especially after combining the sample balance and the optimized boundary scheme, the accuracy once reached 97.5%.In the final lightweight algorithm, while the accuracy is only slightly reduced (from 97.5% to 96%), the time complexity is greatly reduced, even the detection speed is greatly improved, from 14 frames / second to 19 frames / second, and perfect results are achieved in both accuracy and speed. Comparing with Table 2, it can be seen that the improved EAST algorithm in this paper is much faster and more accurate than the traditional image processing method and the traditional deep learning method.

4.CONCLUSION

Based on the EAST algorithm, this algorithm modifies the output layer regression mode, improves the accuracy to a certain extent, and adds instance balance makes the algorithm more reasonable in dealing with sample weight, and the boundary optimization scheme is designed to further improve the positioning edge effect. Especially, this algorithm uses shuffle and deep separable convolution to optimize the original model, which reduces the space-time complexity, compresses the model and

reduces the detection time. Compared with the EAST algorithm, the accuracy and recall rate are improved, and the comprehensive performance is improved compared with other excellent methods of box number detection. However, there are still some shortcomings. The following areas may be optimized:

(1) The receptive field of the network is not large enough, so we need to further improve the network structure and try to introduce void convolution into the network layer.

(2) Try more scale training, and study the optimization of EAST internal parameters and loss weight in order to improve the positioning accuracy.

References

[1] Zhuo junfei, hu yu. Research on license plate detection algorithm based on edge detection and projection method [J]. Chinese science and technology bulletin,2010,26(03):438-441.
[2] Wan yan, xu qinyan, huang mengmeng. Research on license plate detection based on texture and color in complex background [J]. Computer application and software, 2013, 30(10): 259-262+316.
[3] Ren S, He K, Girshick R, et al.Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. 2015
[4] He K,Zhang X, Ren S, et al.Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016
[5] He W, Zhang X Y, Yin F, et al.Deep direct regression for multi-oriented scene text detection[C]//2017 IEEE International Conference on Computer Vision, 2017.
[6] Yao C, Bai X, Liu W, et al.Detecting texts of arbitrary orientations in natural images[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[7] Zhang Z, Zhang C, Shen W, et al.Multi-oriented text detection with fully convolutional networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[8] Neumann L, Matas J.Real- time scene text localization and recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[9] Neumann L, Matas J.A method for text localization and recognition in real- world images[C]//Asian Conference on Computer Vision, 2010.
[10] Gupta A, Vedaldi A, Zisserman A.Synthetic data for text localisation in natural images[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2315-2324.
[11] Bai X, Yao C, Liu W.Strokelets: a learned multi- scale representation for scene text recognition[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[12] Zhou X, Yao C, Wen H, et al.EAST: an efficient and accurate scene text detector[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[13] Yin X C, Yin X, Huang K, et al.Robust text detection in natural scene images[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36 (5): 970-983.
[14] Zhou xiaoyan, wang ke, li lingyan. Overview of target detection algorithm based on deep learning [J]. Electronic measurement technology,2017,40(11):89-93.
[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg. EAST: Single Shot MultiBox Detector. arXiv:1512.02325v5 [cs.CV] 29 Dec 2016
[16] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, Jian Sun.ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices.arXiv:1707.01083 [cs.CV]Tue, 4 Jul 2017.
[17] Deng D, Liu H, Li X, et al.PixelLink: detecting scene text via instance segmentation[C]//AAAI Conference on Artificial Intelligence, 2018.
[18] Liu W, Anguelov D, Erhan D, et al.SSD: single shot multibox detector[C]//European Conference on Computer Vision, 2016.

[19] Tang X, Zheng L, Ma J, et al.PMA: pixel- based multianchor algorithm for image recognition on multi-core systems[C]//International Workshop on Programming Models & Applications for Multicores & Manycores, 2012.
[20] He W, Zhang X Y, Yin F, et al.Deep direct regression for multi- oriented scene text detection[C]//2017 IEEE International Conference on Computer Vision, 2017.
[21] Epshtein B, Ofek E, Wexler Y.Detecting text in natural scenes with stroke width transform[C]//2010 IEEE Conference on Computer Vision and Pattern Recognition, 2010.