ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20211027110610
Original scientific paper

Text Detection of Transformer Based on Deep Learning Algorithm

Yu CHENG, Yiru WAN, Yingjie SIMA, Yinmei ZHANG, Sanying HU, Shu WU*

Abstract: Transformers are important equipment in the power system. At present, the text information collection of transformer nameplates is through manual, which is inefficient. Therefore, it is necessary to find a high-precision automatic detection method of transformer text information. However, the current text detection algorithms have limited ability to detect special characters on the transformer. And they will also have the problem of incomplete detection in detecting the dense text and long text on the transformer nameplate. We propose a text detection network based on segmentation to automatically calibrate the text box of transformer nameplates. Our network is based on DB (differential binarization) network. It has a new feature fusion structure, which refers to the feature fusion structure of the u-net network. The proposed network has achieved better performance than the advanced scene text detection algorithms (DB, East) on the English scene text dataset icdar2015 and the Chinese-English mixed scene text dataset icdar2017. And it also has good performance in GPU occupancy, reasoning speed, and other indicators. The text detection results of actual transformer pictures show that the proposed algorithm solves the problem of poor detection performance of existing deep learning networks in dense text and long text of transformer pictures.

Keywords: deep learning; feature fusion; text detection network based on classification; transformer text detection

1 INTRODUCTION

Nowadays, transformers have been widely used in various fields of industrial engineering. We use transformers with different rated parameters in industrial projects. These transformers are replaced frequently and massively. Using manual methods to record transformer nameplate information will have problems such as low efficiency and wrong judgment. In the process of using the transformer, people's wrong judgment of the transformer information may lead to circuit failure or excessive power consumption. That may threaten the personal safety of workers and the property safety of the factory. Therefore, the automatic acquisition method of transformer information is of great significance in power equipment management. At present, deep learning algorithms are widely used [16-18], which can collect transformer images in real-time, extract equipment nameplate information automatically, and improve equipment management efficiency and statistical accuracy.

However, there is no deep learning algorithm for transformer text detection. There are various backgrounds on the transformer nameplate, and the text types on the transformer include dense text and long text, which are difficult to detect complete text. Therefore, the text detection of the transformer is a new challenge. There are some similarities between device nameplate text detection and scene text detection and transformer text detection. However, the accuracy of these methods for transformer text detection is not ideal, which limits the application of these methods in this field.

Transformer text detection is a special case of equipment nameplate text detection. At present, most of the text detection methods used in this field are scene text detection methods. There are two kinds of scene text detection methods: regression-based methods and segmentation-based methods.

The regression-based text detection methods [1- 5, 13, 14] obtain the text instance bounding box in the picture by regression. There are several regression methods in recent years: Textboxes [1] is based on SSD network [2] and modifies the size of anchor frame and convolution kernel in the network as a text detection network; East [3] and DeepReg [4] adopt the method of no anchor box, and use pixel-level regression to detect multi-directional text instances; DeRPN [5] adopts the dimension decomposition region to complete the negative feedback of output, and it can better solve the scale problem in scene text detection. Regression-based text detection methods usually use simple methods in the post-processing part like non-maximum suppression to remove redundancy and adjust text boxes. However, the current text detection methods which are based on regression have not achieved satisfactory results in the calibration of irregular text boxes.

Segmentation-based text detection methods [6-9, 15] usually detect the text position in the image by combining pixel-level prediction and post-processing algorithms. In recent years, there are several segmentation-based networks architectures: Mask Textspotter [6] is an instance segmentation method based on Mask R-CNN, which is adjusted to detect text instances with arbitrary shapes; PSENet [7] detects text instances with different scale cores by using the extension method of piecewise progressive scale; DB [8] integrates binarization into the processing of feature maps to improve the text detection results without reducing the reasoning speed. In this paper, we use ResNet-50 as the backbone [9], use u-net for feature fusion, and finally, use DB network for post-processing. Our network ensures the reasoning speed and accuracy of text box calibration. And it also improves the redundancy of the backbone in the DB method and reduces the model size.

Fig. 1 is a transformer image. From this figure, we can find that the main reasons which affect the text detection effect on transformer pictures are as follows:

1) The background is complex. The transformer image of text detection has more than ten different background colors. After long-term use, many transformers will have problems such as wear, fading, blurred words, and so on. Moreover, acquisition methods will affect images quality. If the illumination in the shooting scene is uneven, it will reduce the definition of the image. The background of the transformer is complex for many reasons, which increases the difficulty of accurately calibrating the text in the picture.

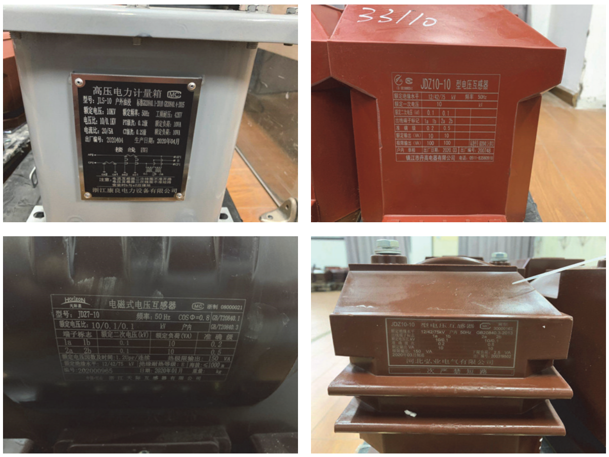

Figure 1 Different kinds of transformers. Transformers' background is complex and the text distribution includes dense texts and long texts

2) Text distribution. The ability of a text detection network to detect dense text and long text depends on the size of the network receptive field. Different text distribution types have higher requirements for the generalization of text detection networks. If the text is small and densely distributed, the network needs a smaller receptive field. If the length of the text is long and the text's font is large, the larger is the feeling field required by the network. The transformer nameplate carries the transformer's important information. The features of the transformer image are complex, so the feature extraction ability and feature resolution ability of the algorithm are required to be high.

Because of the interference by the above reasons, the text detection of the transformer is a challenging work. The features of the transformer image are complex, so the feature extraction ability and receptive field range of the algorithm are required to be high. We propose a text detection depth neural network based on segmentation to solve the text detection problem of the transformer. DBU (differential binarization with U-net) network solves the problem that the DB network has a poor effect on the dense text of transformer pictures. And it also solves the problem that the East network has a poor effect on long text detection. At the same time, compared with the feature fusion module in the DB network, a new feature fusion module used in the proposed network has faster gradient transmission and reasoning speed. Compared with the East network, the training needs fewer iterations.

2 ALGORITHM

2.1 General Framework of Text Detection Network

We proposed a deep learning text detection network based on segmentation. And it can realize the text detection of transformer nameplates. The network framework is shown in Fig. 2:

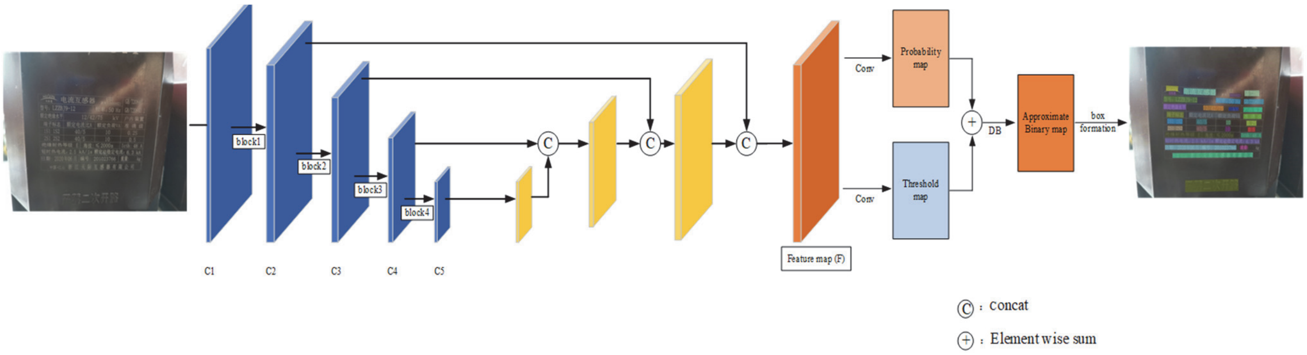

Figure 2 Network structure of DBU

The input image size of the network is 640 × 640, and the output is the text box coordinates $[(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)]$. We use the ResNet-50 network as the proposed network's backbone to extract features. The feature fusion part uses the U-net structure to fuse feature maps of different scales to enlarge the receptive field. Then the text box is extracted from the fused feature graph by the DB network.

2.2 Feature Fusion

The structure of U-net [10] feature fusion is as follows. After the ResNet-50 backbone, we obtain four feature graphs C_1, C_2, C_3, and C_4. They have 64, 128, 256, and 512 channels. After twice up-sampling, the C_4 feature image fuses with the C_3 feature image, and then 3 × 3 convolution and 1 × 1 convolution are performed to get the merged feature image "C_3'". Then we fuse C_3' and C_2 according to a similar method to get C_2'. And C_2' merges C_1 to get the final feature image F. The structure of the U-net is shown in Fig. 3.

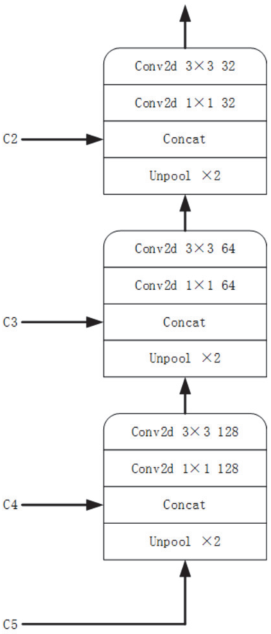

Figure 3 Network structure of U-net

Compared with FCN, U-net retains more features through channel number stitching. When we use the ResNet-50 for feature extraction, the deeper the network layer is, the larger the receptive field is. Shallow convolution pays more attention to the texture features of the image, while deep network pays more attention to the semantic features. Our new feature fusion structure saves the feature maps of different receptive fields in different channels of feature map *F*. This will enable the final feature map of the network to retain as many features as possible. The feature fusion structure of the DBU network can retain the features of different dimensions at the same time.

2.3 DB Network Architecture

Differential binarization network [8]. The DB classification network architecture is shown in Fig. 4 below:

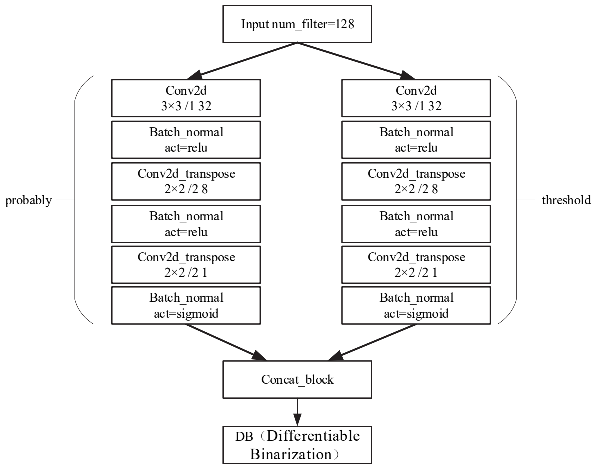

Figure 4 Network structure of DB

We send the feature map *F* to DB classification network. The feature map *F* obtains the probability feature *P* and the threshold feature *T* through two channels with the same structure but different training labels (composed of a 3 × 3 convolution kernel and two 2 × 2 convolutions). The label of the probability graph is the confidence that the pixels in the image belong to the text. And the label of the threshold graph is the text box in the figures. After training, we get a feature graph representing the probability that the pixels in the image belong to the text and a threshold graph marking the text box in the image. And then we send them into the binary network for approximate calculation to obtain the binary map *B*.

In general, binarization is calculated as follows:

$$B_{i,j} = \begin{cases} 1, P_{i,j} \geq t \\ 0, \text{Otherwise} \end{cases} \tag{1}$$

where *t* is the threshold we set to 0.3, and $P_{i,j}$ represents the pixels on the feature map.

The standard binarization formula described in the above formula is non-differentiable. And we cannot optimize the parameters with the network in the process of training. To solve this problem, we use an approximate continuous function in the DB network. This function is an alternative to the standard binarization formula for binarization. The approximate binarization formula is as follows:

$$\widehat{B}_{i,j} = \frac{1}{1+\mathrm{e}^{-k(P_{i,j}-T_{i,j})}} \tag{2}$$

In the above formula, is approximately a binary mapping, *T* is the adaptive threshold mapping learned from the network, and *K* is the amplification factor. We set the *K*'s value to 50 in the training process. The mapping result of this function is similar to that of the standard binary function. But its differentiability enables its constituent parameters to be continuously optimized with the network in the training process.

The probability map and the binary map use the same label. Reduce each dimension box by a certain offset. The definition of the offset is as follows:

$$D = \frac{A(1-r^2)}{L} \tag{3}$$

where *L* is the textbox's perimeter, a is the area of the box, and *R* is the preset scaling factor. We define it as 0.4. The resulting label graph is called G_p, and the original label text box is *G*.

The threshold map uses different labels from the probability map and binary map and expands the original box *G* by *D* offset to get the box G_d. Calculate the distance from the point in the calibration frame G_d to the nearest edge of *G*, divide it by the offset *D* for normalization, and obtain the normalized distance from the point between the G_d frame and the G_p frame to the *G* frame. The label of the threshold feature map is composed of the normalized distance of each point. The value is reduced between 0 and 1, but the label's setting may cause gradient disappearance if the label's point value is 0 or 1. We set it between 0.3 and 0.7 to ensure the smooth iteration of parameters in the training process.

2.3 Loss Function

In the model training process, the loss function includes the confidence loss (L_p) 2-valued graph loss (L_b) and the threshold graph loss (L_t) [8], the confidence loss (L_p). And the binary graph loss (L_b) is the binary cross-entropy loss:

$$L_p = L_b = \sum_{i=1}^{n} (y_i \log \widehat{y_i} + (1-y_i)\log(1-\widehat{y_i})) \tag{4}$$

where y_i is the predicted value of the *i*-th sample, $\widehat{y_i}$ is the true value of the ith sample.

The threshold map loss (L_t) for:

$$L_t = \sum_{i=1}^{R_d} \left| y_i^* - x_i^* \right| \tag{5}$$

where R_d is the number of pixels in the predicted text box, y_i^* is the true value of the threshold prediction module, x_i^* is the predicted value of the threshold prediction module.

3 EXPERIMENT AND ANALYSIS

3.1 Datasets

At present, there is no transformer image data set specially used for transformer text detection, so the data sets used in the training of detection part are icdar2015 data set and icdar2017rctw data set similar to transformer pictures:

Icdar2015 scene detection dataset [11]: There are 1200 images in this dataset, including 1000 training images and 200 test images. Most of these images are English street view images, and the resolution of Google glasses used to collect the pictures is 720 × 1280. Text instances in the dataset are word-level labels. We train the network on the ic15 dataset and evaluate the performance index of the training results. After that, we use the trained model as the pre-training model of the network to train on the ic17 dataset.

Icdar 2017 Reading Chinese Text in the Wild [12]: There are 12263 images in this dataset, including 8034 training images and 4229 test images. Most of these images are Chinese street scene pictures that were taken by a camera, and a few are screenshots, including scene types like indoor scenes and outdoor streets. The text dimensions are quadrilateral in this dataset. The text of the transformer picture is similar to Chinese street scene text in terms of background color. The Chinese text on the transformer accounts for the vast majority of the required detection characters. Therefore, training on this dataset can effectively improve the text detection effect of our network on the transformer pictures.

3.2 Details

We have trained all models in 1200 epochs on the ic15 dataset, 734 iterations per epochs, a total of 880800 iterations to confirm the text detection effect of the DBU network. And the training batch was set to 8. We record the model with the best results of the test set in training and use the best model for performance evaluation. We set the initial learning rate to 0.001, and the first stage estimation's exponential decay rate is used β_1 as 0.9, the second stage estimation's exponential decay rate is used β_2 as 0.999. The training images are enhanced by anti-color, slight angle rotation, clipping, Gaussian noise interference, and merging, etc. All processed images are readjusted to 640 × 640 and then input into the network for training to improve the training speed and reduce the GPU memory.

3.3 Ablation Experiment

We use the icdar2015 dataset to demonstrate the performance improvement of our proposed deep learning network. It can be seen from the table that the proposed R50_DBU (Resnet-50 as backbone network) is compared with other networks. The DBU network achieves the highest accuracy of 85% and h-means of 81% on the ic15 dataset. At the same time, it has good performance in recall rate of 77% and training speed of 0.12 s/step, and it also has the least GPU occupation. The performance of the DBU network is better than that of the DB and East network. Compared with the R50_DB network, our network improves the accuracy index by 4%, H-means by 1%. And our network's reasoning speed is nearly doubled. Compared with R50_East the reasoning time of the East network is increased, the accuracy and other indicators are greatly improved, and the number of iterations required for training is far less than that of the East network.

Ablation experiments on the icdar2017 dataset demonstrated the performance of the R50_DBU network in Chinese English mixed text detection. In the comparison of Darknet53_DB, R50_East, and R50_DB, our network achieved the highest recall rate of 48% and the highest H-mean of 55%. At the same time, it also had the least GPU occupation and faster training speed. Compared to the R50_DB network, the accuracy of R50_DBU decreased slightly, but the recall rate increased by 6%, the H-mean increased by 2%, and the reasoning speed nearly doubled. Compared with the R50_East network, although the time required for reasoning has increased, other indicators such as accuracy have been greatly improved, including accuracy 36%, recall 9%, H-mean 22%, and the number of iterations required for training is far less than R50_East network.

Our network has an optimized feature fusion structure, which can fuse more scale feature maps. That is why our network is better than DB and East networks in precision and h-mean. At the same time, the feature fusion structure of our network reduces the redundancy, which makes it occupy less GPU and faster training speed than the DB network.

Table 1 The ic15 dataset training results

Network	Precision	Recall	H-mean	GPU use	Training speed time/step
Mv3_DB	78%	59%	67%	~ 6 GB	0.13 s
R50_DB	81%	**78%**	8%	~ 9 GB	0.23 s
Mv3_East	77%	78%	77%	~ 6 GB	**0.03 s**
Darknet53_DB+swish	82%	75%	78.6%	~ 5 GB	0.14 s
R50_East	70%	77%	73%	~ 6 GB	0.08 s
R50_DBU	**85%**	77%	**81%**	**~ 5 GB**	0.12 s

Dataset training parameter: batch size = 4; epoch =1200; single GPU; 2080ti; training images = 1000

Table 2 The ic17 dataset training results

Network	Precision	Recall	H-mean	GPU use	Training speed time/step
Darknet53_DB+swish	62%	39%	47%	~ 5 GB	0.14 s
R50_East	29%	39%	33%	~ 6 GB	**0.08 s**
R50_DB	**71%**	42%	53%	~ 9 GB	0.23 s
R50_DBU	65%	**48%**	**55%**	**~ 5 GB**	0.12 s

Dataset training parameter: batch size = 4; epoch =600; single GPU; 2080ti; training images = 8046

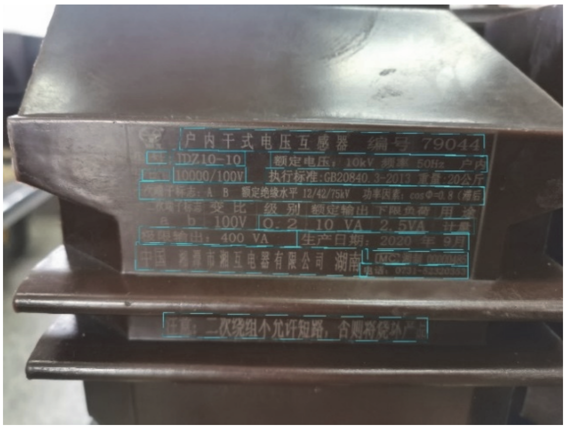
户内干式电压互感器 编号 79044
JDZ10-10
10000/100V
执行标准:GB20840.3-2013
400 VA
注意:二次绕组不允许短路,否则将烧坏产品

Figure 5 Picture of DB network detect transformer text. DB network omitted some dense text in detection

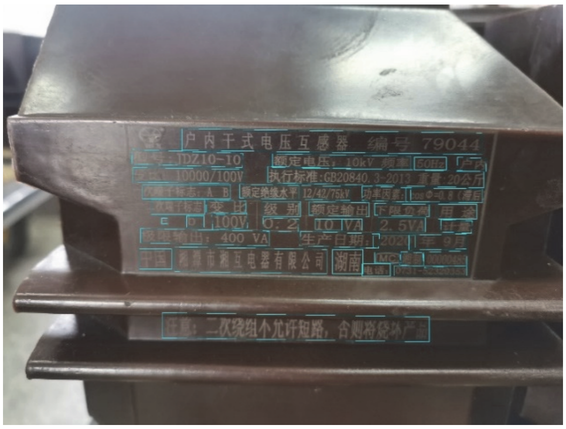
户内干式电压互感器 编号 79044
JDZ10-10
10000/100V
400 VA
注意:二次绕组不允许短路,否则将烧坏产品

Figure 6 Picture of DBU network detect transformer text. DBU network detected most of the texts on the transformer, including dense texts and long texts

Fig. 5 and Fig. 6 show the comparison effect of the DBU network and the DB network on transformer pictures. The detection effect of our proposed DBU network after training on the same dataset is better than that of the DB network. We can see from the two text detection images that the DB network failed to detect the dense part of the text on this transformer. But the DBU network detects all the text on the transformer image.

3.4 Detection Result of Transformer

The text detection results of transformer pictures are shown in Fig. 7 and Fig. 8. The areas marked with different colors in the pictures are the text boxes detected by our network.

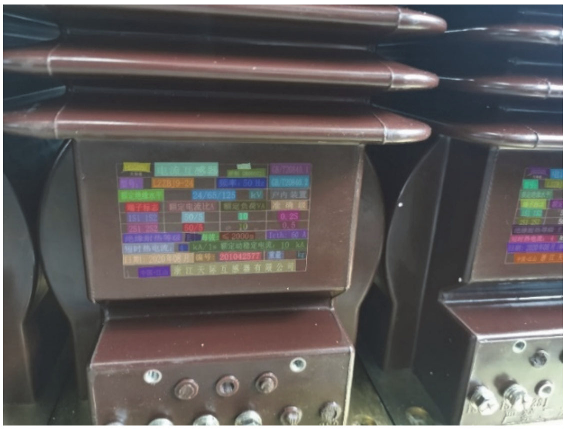

Figure 7 Picture 1 of DBU network transformer detection

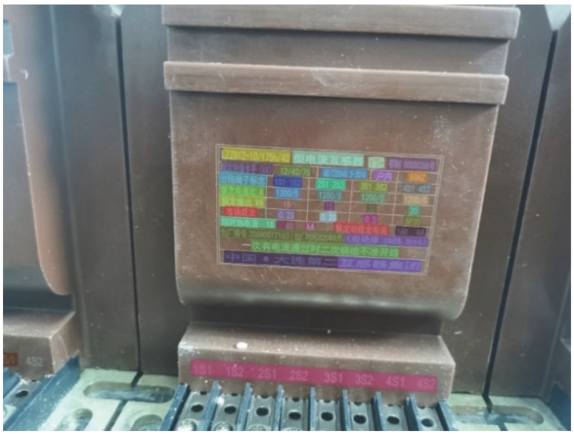

Figure 8 Picture 2 of DBU network transformer detection

From these two pictures, we can see that our proposed DBU text detection network can detect most of the text on transformers, including small texts, dense texts, and long texts. Our network not only ensures the accuracy of detection but also makes the detected text have semantic continuity.

4 CONCLUSION

Transformers are widely used in industrial engineering, the automatic detection method of transformers' nameplate information is of great significance. We proposed a deep neural network based on segmentation to detect the text of the transformer pictures. The network determines the characters, numbers, and symbols in the image from the original transformer images to facilitate the recording and managing of the transformer nameplate information.

Compared with the DB network, the DBU text detection network proposed has a certain improvement in the accuracy and h-means of the ic15 dataset and has a significant improvement in the network performance of the ic17 dataset. At the same time, the reasoning speed of the DBU network is better than that of the DB network, and the number of iterations required is much less than that of the East network.

There are still some technical problems in our research. For example, the accuracy of the network trained in public datasets such as ic15 can be further improved, the network structure can be further simplified, and the scale of the model can be further reduced. In future research, we will try to introduce the transformer image dataset and simplify the model to improve the text detection network. The application of text detection on the nameplate of transformers and other equipment is still in its infancy. In the future, we will further improve the evaluation index to make it more reasonably reflect the accuracy of detection and identification.

5 REFERENCES

[1] Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. (2017). Textboxes: A fast text detector with a single deep neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*. http://doi.org/10.13336/j.1003-6520.hve.20191344

[2] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., & Reed, S. (2016). SSD: single shot multibox detector. *Proceedings of ECCV*. https://doi.org/10.1007/978-3-319-46448-0_2

[3] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., &

Liang, J. (2017). EAST: an efficient and accurate scene text detector. *Proceedings of CVPR*. https://doi.org/10.1109/CVPR.2017.283

[4] He, W., Zhang, X., Yin, F., & Liu, C. (2017b). Deep direct regression for multi-oriented scene text detection. *Proceedings of International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2017.87

[5] Xie, L., Liu, Y., Jin, L., & Xie, Z. (2019b). Derpn: Taking a further step toward more general object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 9046-9053. https://doi.org/10.1609/aaai.v33i01.33019046

[6] Lyu, P., Liao, M., Yao, C., Wu, W., & Bai, X. (2018b). Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *Proceedings of ECCV*, 67-83. https://doi.org/10.1007/978-3-030-01264-9_5

[7] Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., & Shao, S. (2019b). Shape robust text detection with progressive scale expansion network. *Proceedings of International Conference on Computer Vision*, 2019a, 9336-9345. https://doi.org/10.1109/CVPR.2019.00956

[8] Minghui, L., Zhaoyi, W., Cong, Y., Kai, C., & Xiang, B. (2020). Real-time Scene Text Detection with Differentiable Binarization. *Proceedings of the AAAI Conference on Artificial Intelligence.* https://doi.org/10.1609/aaai.v34i07.6812

[9] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition.

[10] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical.

[11] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S. K., Bagdanov, A. D., Iwamura, M., Matas, J., Neumann, L., Chandrasehar, V. R., Lu, S., Shafait, F., Uchida, S., & Valveny, E. (2015). ICDAR 2015 competition on robust reading. *Proceedings of ICDAR.* https://doi.org/10.1109/ICDAR.2015.7333942

[12] Baoguang, S. et al. (2017). ICDAR 2017 competition on robust reading. *Proceedings of ICDAR.*

[13] Liao, M., Shi, B., & Bai, X. (2018). Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, *27*(8), 3676-3690. https://doi.org/10.1109/TIP.2018.2825107

[14] Liu, Y. & Jin, L. (2017). Deep matching prior network: Toward tighter multi-oriented text detection. *Proceedings of CVPR.* https://doi.org/10.1109/CVPR.2017.368

[15] Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X., & Jia, J. (2019). Learning shape-aware embedding for scene text detection. *Proceedings of CVPR*, 4234-4243. https://doi.org/10.1109/CVPR.2019.00436

[16] Zhang, Q., Liu, S., Gong, D., & Tu, Q. (2019). A Latent-Dirichlet-Allocation Based Extension for Domain Ontology of Enterprise's Technological Innovation. *International Journal of Computers Communications & Control*, *14*(1), 107-123. https://doi.org/10.15837/ijccc.2019.1.3366

[17] Ma, Y., Zhang, Z., Ihler, A., & Pan, B. (2018). Estimating Warehouse Rental Price using Machine Learning Techniques. *International Journal of Computers Communications & Control*, *13*(2), 235-250. https://doi.org/10.15837/ijccc.2018.2.3034

[18] Ma, Y., Zhang, Z., & Ihler, A., 2020. A Deep Choice Model for Hiring Outcome Prediction in Online Labor Markets. *International Journal of Computers Communications & Control*, *15*(2). https://doi.org/10.15837/ijccc.2020.2.3760

Contact information:

Yu CHENG, Senior Engineer
Hangzhou power supply company of State Grid Zhejiang Electric Power Co. Ltd.,
Marketing technology center, Zhejiang Province, Hangzhou, China
E-mail: 653970631@qq.com

Yiru WAN, Engineer
Hangzhou power supply company of State Grid Zhejiang Electric Power Co. Ltd.,
Marketing technology center, Zhejiang Province, Hangzhou, China
E-mail: 15652264@qq.com

Yingjie SIMA, Assistant Engineer
State grid Zhejiang jiande power supply Co. Ltd,
No. 288 Xin'an Road, Xinanjiang Street, Jiande city, Zhejiang Province, China
E-mail: s_myj@163.com

Yinmei ZHANG, Engineer
Hangzhou power supply company of State Grid Zhejiang Electric Power Co. Ltd,
Marketing technology center, Zhejiang Province, Hangzhou, China
E-mail: mmmzum5@163.com

Sanying HU, Engineer
Hangzhou power supply company of State Grid Zhejiang Electric Power Co. Ltd,
Marketing technology center, Zhejiang Province, Hangzhou, China
E-mail: 527275879@qq.com

Shu WU, Postgraduate
(Corresponding author)
Beijing University of Posts and Telecommunications,
10 Xitucheng Road, Haidian District, Beijing, China
E-mail: wushu@bupt.edu.cn