Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Robust Text Line Detection in Equipment Nameplate Images*

Jiangyu Lai†1, Lanqing Guo†23, Yu Qiao23, Xiaolong Chen1, Zhengfu Zhang23
Canping Liu1, Ying Li23 and Bin Fu‡23

Abstract— **Scene text detection for equipment nameplates in the wild is important for equipment inspection robot since it enables inspection robot to take specific actions for different equipment's. Although text detection in images has achieved great progress in recent years, the detection for equipment nameplates faces several challenges such as extreme illumination and distortion which significantly decrease the detection performance. In this paper, we propose a deep text detection model Robust Text Line Detection (RTLD) for locating word level text instances in equipment cards. Specifically, the proposed model first employs a corner detection module to determine the four corner points of each nameplate, and then a carefully designed image transformed module transforms the irregular nameplate region into a rectangular region. Finally, text detection module is introduced to locate every word level text instance in the transformed images. We conduct extensive experiments to examine our proposed methods on real equipment nameplate images. Our model achieves 91.2% precision and 92.6% recall on Equipment Nameplate Dataset. The experimental results demonstrate the effectiveness of our models.**

Index Terms— **Artificial Intelligence, Robotic Vision, Deep Learning in Robotics, Scene Text Detection.**

I. Introduction

Inspection robots have became a hot topic in recent years and people can employ them to perform safety management in factory. Since there are various equipments in factory, recognizing them from equipment nameplate is a promising approach and has been received more and more attentions. However, detecting text instances in equipment nameplates in the wild is a challenging task due to extreme illumination and the distortion caused by shooting slope. In this paper, we pay attention on solving this problem and put forward a text detection method to detect word level text instances on equipment nameplates.

Scene text detection is one of the fundamental tasks in computer vision. The goal of text detection task is to find out text region in natural images which can be widely used in various applications such as scene optical character recognition (OCR), image retrieval and multilingual translation. In recent years, scene text detection performance has been significantly improved due to the development of Deep Convolutional Neural Networks (DCNNs) and various state-of-the-art methods have been put forward such as CTPN [1] and EAST [2]. Although these DCNNs based detection methods have made great progress, the text detection for equipment nameplate in the wild is still a challenging task since extreme conditions will cause equipment nameplate become obscure and damaged which may significant decrease detection performance. Moreover, compared with some famous scene text datasets [3], [4], [5], the word level text instances in equipment nameplates have some special characteristics such as multi-line text and extremely long text. In this paper, we propose a Robust Text Line Detection (RTLD) model to detect text region in equipment nameplates. Although equipment nameplates usually have well-defined shape, it will be distorted by image acquisition process. To restore origin shape of equipment nameplates and get rid of complex background information, we employ a corner detection module to locate nameplate region in natural image. The detected irregular nameplate region will be transformed into a rectangular region with projection transformation and the unrelated region will be removed after cropping text region from the transformed image. Finally, a text detection module will be employed to locate every word level instance in equipment nameplates and give the corresponding coordinates as the output of the model.

This paper is organized as following. We first introduce several related works in object detection and text detection tasks. In Section III, we will propose our text detection model for equipment nameplates. Several extensive experiments will be conducted and presented in Section IV to demonstrate the effectiveness for our model. A briefly summary will be given at the end of this paper.

II. Related Works

A. Object Detection

Sliding window approach has been widely used in tradition object detection methods. This approach employ a set of predefined sliding windows with different scales to extract predefined feature vector (such as Haar feature [6] for face

*This work is supported by Guangzhou Power Supply Bureau Co. Ltd (Key Technology Development for Scene Text Detection and Recognition of Equipment Nameplates Based on Deep Learning Method, 080032KK52180002)

† Jiangyu Lai and Lanqing Guo contribute equally to this paper.

‡ Bin Fu is the corresponding author in this paper.

1 Jiangyu Lai, Xiaolong Chen and Canping Liu are with Guangzhou Power Supply Bureau Co. Ltd., Guangzhou, China. {laijy, chenxl1116, liucp}@guangzhou.csg.cn.

2 Lanqing Guo, Yu Qiao, Zhengfu Zhang, Ying Li and Bin Fu are with ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. {lq.guo, yu.qiao, zf.zhang, li.ying, bin.fu}@siat.ac.cn

3 Lanqing Guo, Yu Qiao, Zhengfu Zhang, Ying Li and Bin Fu are with SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

detection and HOG feature [7] for pedestrian detection) for each region proposal. The predefined feature vectors are feed to a classifier (such as SVM) to obtain final predictions. In recent years, Deep Convolutional Neural Networks (D-CNNs) have achieved great progress on object detection task which outperformed tradition models with a large margin. Compared with tradition methods, the DCNNs based method employ a set of cascade convolution filters to automatically extract feature vectors. The common strategy is to generate a set of region proposals (region of interest, ROI) by using a shallow network and then employs a strong classifier network to obtain final detection results. For example, R-CNN [8] model employ Selective Search method to generate a set of region proposals and then use a deep classification network to obtain bounding box for each objects. Faster R-CNN [9] further develops this approach by replacing selective search method with the Region Proposal Network (RPN) which directly extracted region proposals from a deep subnetwork. Inspired by this work, several state-of-the-art anchor based detection frameworks have been put forward in recent years which significantly improve the performance on object detection task.

B. Text Detection

1) Tradition Method: Tradition text detection usually relies on extracting hand-designed feature vectors from input images. It is difficult to detect and locate text region due to various text structure, complex background, low resolution and shape distortion. The Connected Component Analysis Labelling [10], [11] or Sliding Window [12] approaches are widely used to deal with this task. Connected Component Analysis Labelling first extracts region proposals by using several different approaches and then filter non-text regions under predefined rules. In Sliding Window approach, several predefined sliding windows with different scales are employed to classify each regions into text or non-text region.

2) Deep Learning Based Approach: In recent years, various computer vision tasks have achieved significant improvement due to the development of deep leaning technique. With the carefully designed CNNs architecture, deep learning technique employs back propagation method to automatically learn how to extract task-specific feature vectors from data which significant improve model performance. The existing text detection models can be roughly sort into two different categories: regression based detection models and segmentation based models. The regression based models directly regress coordinates of text regions and they are usually modified from some object detection frameworks such as Faster R-CNN [9] and SSD [13]. Since the length and scale between object and text region are different, several methods modify the convolution kernel and anchor scales to detect text regions. Moreover, several approaches further modify object detection model by carefully selecting proper presentation to regress text instance such as CTPN [1] and RRPN [14]. RRD [15] employs two separate branches to extract feature map with classification and regression information from different branches to improve detection performance. Segmentation based models are inspired by semantic or instance segmentation frameworks such as FCN [16] and DeepLab [17]. These methods describe text regions with several predefined properties and then modify a segmentation framework to generate the corresponding heat maps for text regions. For example, Lyu et al.[18] predicts the heat map for the position of text corner and PixelLink [19] generate text/non-text mask and pixel-wise connections to detect text instance.

Since most word level text instances are regular shape and the text regions are densely on equipment nameplates, we adopt regression based method to design our detection model. In the following section, we will give a detailed description about our text detection model.

III. Method

To detect text regions in equipment nameplates, we first employ the Corner Point Detection module to detect four corner points for each equipment nameplate region. The irregular nameplate region will be transformed to a rectangular region by employing projection transformation. Finally, a text detection model will be used to detect near-horizontal text regions in transformed nameplate region. The pipeline has been shown in Fig. 1.

A. Data Augmentation

The parameters of deep learning models are optimized during the training stage. The optimization target is the careful designed loss function according to different computer vision tasks and the optimization process employ back propagation method to gradually find the minimum point of loss function. Since there are large numbers of parameters in DCNNs model, a large numbers of training data are required to train DCNNs model. In this work, we employ two similar datasets to offer more training samples for our detection model. We first pre-train our model on these two datasets and then fine tune our model on equipment nameplates.

B. Corner Point Detection

The shape of equipment nameplates is usually rectangle. However, due to lens distortion and the shooting angle in image acquisition process, the shape of equipment nameplate is irregular. Therefore, image transformed is an important step to transform irregular shape to rectangle shape. In this work, we employ HRNet [21] to detect corners of equipment nameplates from input image.

C. Image Transformed Module

Once we obtain the corner points of equipment nameplate in natural image, we can employ projection transformation to reshape irregular nameplate region into rectangular region.

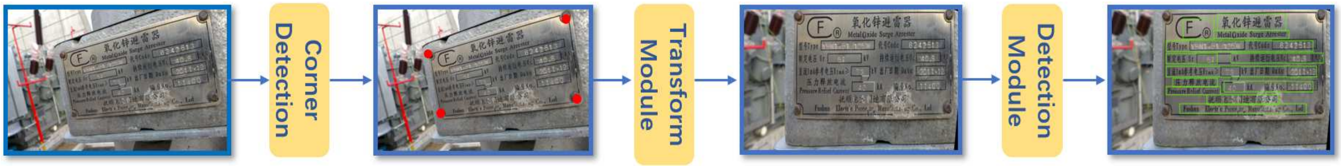

Fig. 1. The detection pipeline for our scene text detection model. We first employs a corner deterction to detect four corner points of each equipment nameplate. A transform module is further employed to transform irregular nameplate region to the corresponding rectangular region. Finally, we detect text regions from the transformed rectangular equipment nameplate.

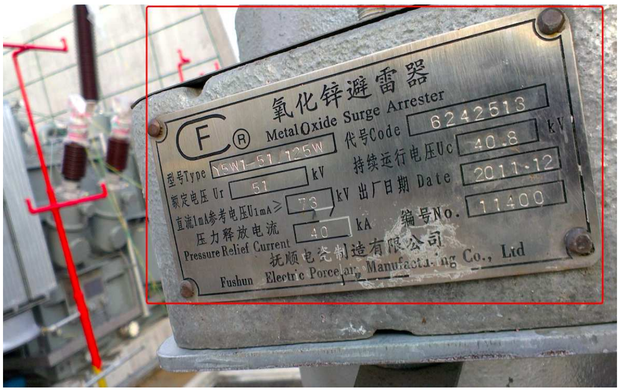

Fig. 2. Extracting minimum horizontal enclosing rectangle from equipment nameplate corner detection. The red rectangular box is the minimum horizontal enclosing rectangle for equipment nameplate.

Moreover, we can further get rid of complex background information by cropping the corresponding nameplate region from the transformed image.

1) Minimum Horizontal Enclosing Rectangle: As shown in Fig. 2, after obtaining the corner coordinates of equipment nameplates, we can calculate the corresponding coordinates of enclosing rectangle from the following relation:

$$R_1(x_{min}, y_{min}) \tag{1}$$

$$R_2(x_{max}, y_{min}) \tag{2}$$

$$R_3(x_{max}, y_{max}) \tag{3}$$

$$R_4(x_{min}, y_{max}) \tag{4}$$

where R_1, R_2, R_3 and R_4 are left-up, right-up, right-down, left-down corner points of the enclosing rectangle.

2) Projection Transformation Matrix: We employ projection transformation to transform the irregular shape into rectangular shape in 2D images. The projection transformation can be written as:

$$[x', y', w'] = [u, v, w] \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{5}$$

where u and v are the coordinates in origin image while coordinates of the transformed image are x and y with the relations $x = x'/w'$ and $y = y'/w'$.

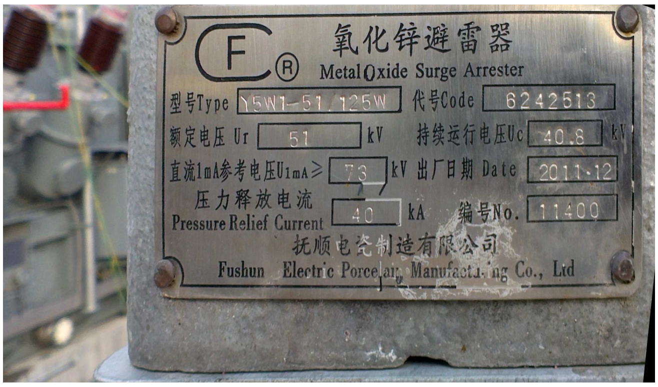
氧化锌避雷器
Metal Oxide Surge Arrester
型号Type Y5W1-51/125W
代号Code 6242513
额定电压 Ur 51 kV
持续运行电压Uc 40.8 kV
直流1mA参考电压U1mA≥ 73 kV
出厂日期 Date 2011-12
压力释放电流 Pressure Relief Current 40 kA
编号No. 11400
抚顺电瓷制造有限公司
Fushun Electric Porcelain Manufacturing Co., Ltd

Fig. 3. The transformed equipment nameplate after projection transformation. Since most text detection models are designed for detecting horizontal text instance, image transformation process will improve detection performance.

In this work, we employ OpenCV built-in function to perform projection transformation by setting the corner points of the nameplate region and the minimum horizontal enclosing rectangle as the reference points before and after transformation. The transformed equipment nameplate is shown in Fig. 3.

D. Text Detection Model for Equipment Nameplate

In this sub-section, we give a detailed description for our text detection module. Our text detection module is based on CTPN model [1] which generates fine-scale proposals for text regions and employs a recurrent neural network (RNN) to connect text proposals into a horizontal text region. In order to detect text regions in equipment nameplate with a high precision, we perform several modifications on this model. (1). The fine-scale proposals are generated by predicting the hight of proposals with fixed width. Since the text regions in equipment nameplates are closed with each other as shown in Fig. 4, the detection network cannot identify the boundary of neighbouring text regions and will fuse them into a single text region if the width of proposal is large. To settle down this problem, we modify CTPN method by selecting a proper width for fine-scale proposals. In this work, we set this width as 24 pixels and the influence of this setting will be discussed in next section.

(2). The CTPN model employs three loss functions L_s^{cl},

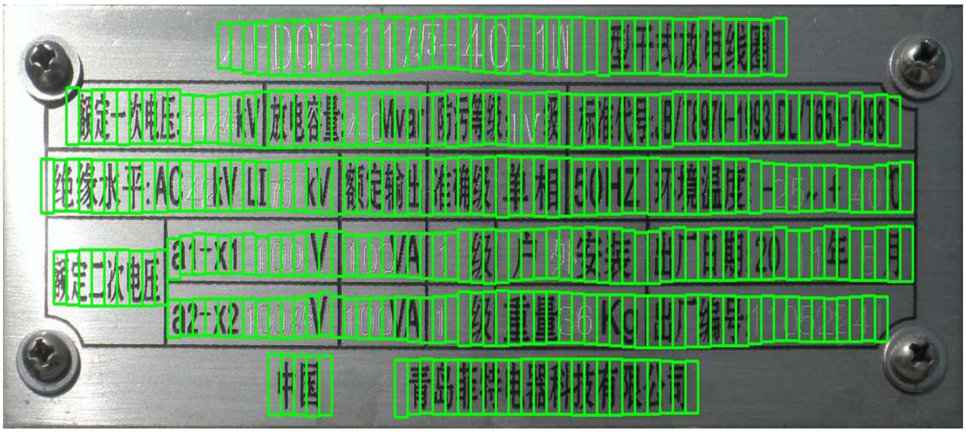

Fig. 4. The text regions in equipment nameplates are closed with each other. Therefore, the fine-scale proposals cannot be separated if the width of fine-scale is large. The green bounding boxes are fine-scale proposals.

L_v^{re} and L_o^{re} to optimize the detection network to predict text/non-text probability (s), vertical coordinate ($v = \{v_c, v_h\}$) and side-refinement offset (o). The multi-task loss function can be written as [1],

$$L = \frac{1}{N_s}\sum L_s^{cl} + \frac{\lambda_1}{N_v}\sum L_v^{re} + \frac{\lambda_2}{N_o}\sum L_o^{re} \tag{6}$$

where N_s, N_v and N_o are the total numbers of training samples for each sub-task in mini-batch, $\lambda_1 = 1$ and $\lambda_2 = 2$ are used to balance different tasks in loss function [1]. The L_s^{cl} is the Softmax Classification Loss while L_v^{re} and L_o^{re} are Smooth L_1 Loss in origin work [1]. However, in our framework, the corner points of text regions in equipment nameplates may contains some noise due to annotation and this will make the training process become unstable if we use Smoother L_1 Loss. Therefore, we change loss function L_v^{re} from the widely used Smooth L_1 Loss to IoU Loss.

IV. Experiments

In this section, we adopt our text detection model and perform several extensive experiments to show effectiveness of our proposed model for detecting text region in equipment nameplates.

A. Dataset

1) ICDAR2013: The ICDAR 2013 dataset [4] is a widely used dataset for detecting horizontal text regions in natural images. It provides 229 images as training set and 233 images as testing set with the word level annotation.

2) HuaWei Scene Text Dataset: The HuaWei Scene Text Dataset provide 3195 natural images with word level text instances for training. We employ this dataset as an external dataset for training our model, since it contains advertising board, banner and product logo in the wild which are similar with equipment nameplates.

3) The Equipment Nameplate Dataset: The Equipment Nameplate Dataset [20] contains 502 images for training and testing. All images are taken in the wild and collected from several kinds of equipment. Text regions are near-horizontal and the word level instances are annotated as rectangle bounding boxes.

B. Evaluation Metric

For equipment nameplates, there is a common situation that several text instances in a same line with small margin between them. In this case, the detection performance are serious decreased since the detection model usually fuses several text regions into a single region and these regions will be regards as false result by the evaluation metric. Therefore, to give an objective judgement of detection model, we employ ”DetEval” method offered by ICDAR 2013 Robust Reading Competition [4] to measure the performance of our detection model.

C. Implement Details

To transform irregular equipment nameplates into origin shape, we first train our corner detection netwok with the same setting as [21]. The ground truth of corner points are generated from Equipment Nameplate Dataset using the annotation of bounding box for equipment nameplates. Our text detection model uses VGG16 [22] as backbone and pre-trained this backbone on ImageNet dataset [23]. Following [1], a set of binary labels are generated to text or non-text anchors which are defined by comparing the overlap with ground truth bounding boxes; 64 positive anchors and 64 negative anchors are selected for each batch. We merge ICDAR 2013 and HuaWei datasets resulting 3423 natural images (228 in ICDAR 2013 training set and 3195 in HUAWEI training set) to pre-train our model. Our detection model train 300K iterations on the Equipment Nameplate Dataset by using a

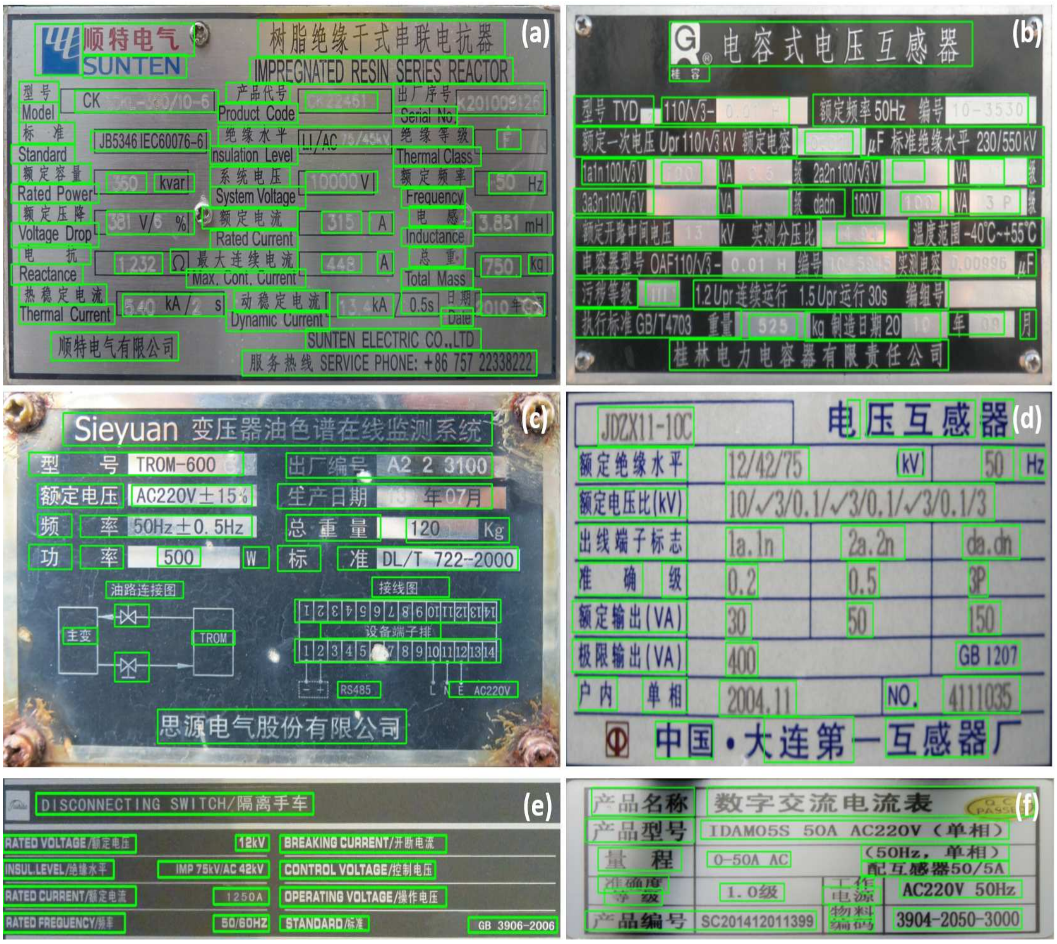

Fig. 5. The text detection result for equipment nameplates. Our model can detect the engraved text region with high precision. The multi-line text instance cannot detect precisely; If two text regions are extremely closed with each other, the detection model may give a wrong prediction.

single Tesla K80 GPU. The initial learning rate $1e^{-5}$ and will decay 0.1 after each 30K iterations.

Since the size of text region and character have large variance after cropping from transformed images, multi-scale input images are employed to solve this problem during training and evaluation process. We resize the cropped image into three different width (600, 800 and 1000 pixels) and adjust length of images to keep the length-width ratio unchanged.

D. Ablation Study

In this sub-section, we conduct several ablation experiments to verify the effectiveness of our proposed text detection model for locating text regions in equipment nameplate.

1) Proposal Width: Since the fine-scale proposal width is an important factor for making different text regions well-separate. In order to verify this statement, we perform several extensive experiments with different proposal width settings. Experimental results are shown in Table 1. From this table, we can find that the proposal width has a significant influence on performance of our detection model and the performance has a variance about 1.5% with different settings. Moreover, too large and too small width both have negative influence on detection performance and the model achieves best performance when fine-scale proposal width is 24 pixels. Therefore, we set fine-scale proposal width as 24 pixels in the following experiments.

TABLE I
DETECTION PERFORMANCE WITH DIFFERENT WIDTH OF FINE-SCALE PROPOSALS

Proposal Width	Recall	Precision	H-mean
8	91.9	89.0	90.5
12	92.7	89.7	91.2
16	92.9	90.1	90.5
20	92.2	91.3	91.7
24	92.6	91.2	91.9
28	90.5	92.3	91.4

2) External Dataset Pre-train: Since the number of training samples plays an important roles in deep learning algorithm, we increase the training samples by employing external datasets to pre-train our scene text detection model. We merge ICDAR 2013 [4] and HuaWei datasets as the pre-trained dataset. We perform ablation experiment to determine the influence of pre-trained dataset on detection performance. With pre-trained dataset, our model achieves 92.6% Recall, 91.2% precision and 91.9% H-mean, while detection performance will decrease to 92.3% Recall, 89.1% precision and 90.7% H-mean without pre-trained dataset. Therefore, employing external training samples will significantly improve the precision (about 2%) of our detection model.

E. Experimental Results and Analysis

We adopt our model to detect text regions on testing set. The Equipment Nameplate Dataset offer 50 images as the test set and each image contains several text regions. As we have discussed, we employ DetEval method [4] to evaluate detection performance of our model. With multi-scale evaluation, our model achieves 91.2% precision, 92.6% recall and 91.9% H-mean.

To further analyse our proposed model, we visualize several detection results in Fig. 5. From this figure, we find that our proposed method can detect word level instances precisely. Moreover, although the engraved text is difficult for text detection model due to the similar appearance with background region, our model can detect them with high precision. There are several drawbacks of this model which need to be further improvement. (1). The proposed model cannot detect multi-line text region precisely. Since most text detection algorithms are design for detecting text region in a single line, it is difficult for our proposed model to detect multi-line text region. To solve this problem, we can design a region fusion mechanism to adaptively fuse text region with high overlap in vertical direction. (2). The model may give a mistake detection result if two text lines are extremely closed with each other. This can be modified to enlarge the size of input images since it makes two text lines separate with each other. This problem can be modified by designing a text region fusion mechanism in horizontal direction.

V. Conclusion

In this paper, we propose a Robust Text Line Detection (RTLD) model to detect and locate word level instances for equipment nameplate in the wild. The proposed model includes three parts, which are corner detection module for equipment nameplate, image regularization module and the text region detection module. With the carefully designed, our model achieves 91.2% precision and 92.6% recall on the Equipment Nameplate Dataset which demonstrates the effectiveness of our proposed model.

Acknowledgment

This work is supported by Guangzhou Power Supply Bureau Co. Ltd (Key Technology Development for Scene Text Detection and Recognition of Equipment Nameplates Based on Deep Learning Method, 080032KK52180002).

References

[1] Z. Tian, et al, "Detecting Text in Natural Image with Connectionist Text Proposal Network," European Conference on Computer Vision, p. 56, 2016.

[2] X. Zhou, et al, "EAST: An Efficient and Accurate Scene Text Detector," Computer Vision and Pattern Recognition, p. 2642, 2017.

[3] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 Robust Reading Competitions," 7th International Conference on Document Analysis and Recognition, p. 682, 2003.

[4] D. Karatzas, et al. "ICDAR 2013 Robust Reading Competition," 12th International Conference on Document Analysis and Recognition, p. 1484, 2013.

[5] D. Karatzas, et al, "ICDAR 2015 Competition on Robust Reading," 13th International Conference on Document Analysis and Recognition, p. 1156, 2015.

[6] R. Lienhart and J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection," International Conference on Image Processing, p. 900, 2002.

[7] Y. Pang, Y. Yuan, X. Li and J. Pan, "Efficient HOG Human Detection," Signal Processing, p. 773, 2011.

[8] R. B. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," Computer Vision and Pattern Recognition, p. 580, 2014.

[9] S. Ren, K. He, R. B. Girshick and J. Sun, "Faster R-cnn: Towards Real-time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1137, 2017.

[10] B. Epshtein, O. Eyal and W. Yonatan, "Detecting Text in Natural Scenes with Stroke Width Transform," Computer Vision and Pattern Recognition, p. 2963, 2010.

[11] L. Neumann and M. Jiri, "A Method for Text Localization and Recognition in Real-world Images," Asian Conference on Computer Vision, p. 770, 2010.

[12] J. J. Lee, P. H. Lee, S. W. Lee, A. L. Yuille, and C. Koch, "AdaBoost for Text Detection in Natural Scene," International Conference on Document Analysis and Recognition, p. 429, 2011.

[13] W. Liu, et al, "SSD: Single Shot Multibox Detector," European Conference on Computer Vision, p. 21, 2016.

[14] J. Ma, et al, "Arbitrary-oriented Scene Text Detection via Rotation Proposals," IEEE Transactions on Multimedia, p. 3111, 2018.

[15] M. Liao, Z. Zhu, B. Shi, G. Xia, and X. Bai, "Rotation-Sensitive Regression for Oriented Scene Text Detection," Computer Vision and Pattern Recognition, p. 5909, 2018.

[16] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," Computer Vision and Pattern Recognition, p. 3431, 2015.

[17] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs," IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 40(4), p. 834, 2018.

[18] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented Scene Text Detection via Corner Localization and Region Segmentation," Computer Vision and Pattern Recognition, p. 7553, 2018.

[19] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting Scene Text via Instance Segmentation," Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[20] Xiaolong Chen, et al, "The Equipment Nameplate Dataset for Scene Text Detection and Recognition," unpublished.

[21] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," Computer Vision and Pattern Recognition, p. 5693, 2018.

[22] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representations, 2014.

[23] J. Deng, et al, "Imagenet: A Large-scale Hierarchical Image Database," Computer Vision and Pattern Recognition, p. 248, 2009.