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

Orientation Robust Scene Text Recognition in Natural Scene[*]

Xiaolong Chen[†1], Zhengfu Zhang[†23], Yu Qiao[23], Jiangyu Lai[1], Jian Jiang[1], Zeyu Zhang[2]and Bin Fu[‡23]

***Abstract*— In recent years, scene text recognition has achieved significant improvement and various state-of-the-art recognition approaches have been proposed. This paper focused on recognizing text in natural photos of equipment nameplates, which has wide applications in industrial automations. This task only receives little attentions in previous works. The challenge of this problem comes from multi-orientation, curved, noisy and blurry text patches in equipment nameplates. To address this problem, we propose a deep model for text recognition in multi-oriented nameplates, namely, Orientation Robust Scene Text Recognition (ORSTR). Specifically, our model employs a rectification module to transform curved, distorted or multi-orientation text to near-horizontal text with a carefully designed rectification module. Once the near-horizontal text has been generated, recognition network will output the predictions of text patches. Our scene text recognition model achieves 90.8% recognition accuracy on equipment nameplate dataset which outperforms previous scene text recognition model (CRNN) about 0.8%. Several extensive experiments have been conducted to verify the effectiveness of our model.**

***Index Terms*— Artificial Intelligence, Robotic Vision, Scene Text Recognition**

I. Introduction

Scene text recognition is one of most important artificial intelligence tasks which has attracted great attentions in recent years due to its widely applications such as image retrieval and multilingual translation. Recant advances in Deep Convolutional Neural Networks (DCNNs) has significant improved scene text recognition performance. Various DCNNs based recognition models have been put forward, such as CRNN [1] and ASTER [2]. CRNN combines Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to extract sequence and context information from input text patches. The resulting feature vectors are passed to transcription layer to obtain final predictions. ASTER adds a Text Rectification Network before recognition network to transform curved text region into horizontal one which significant improve recognition performance for multi-orientation and curved text.

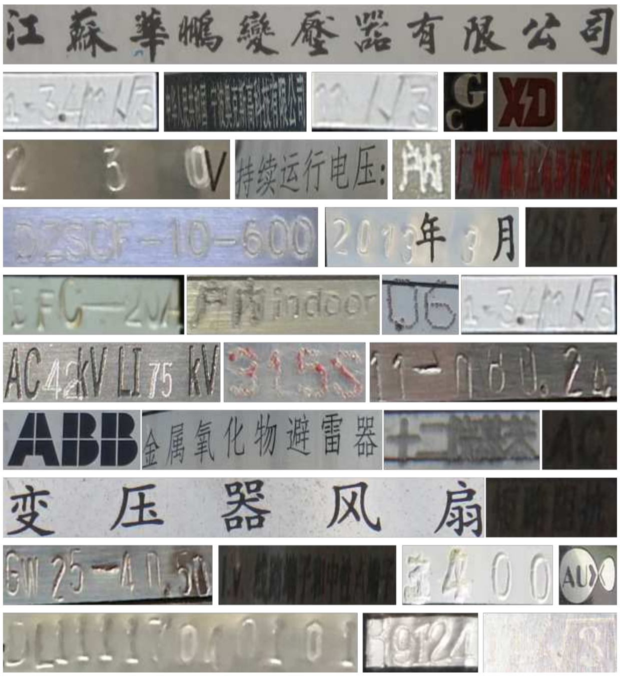

Fig. 1. Reading text instance in natural scene is a challenging task since diverse text appearance and some extreme conditions will give negative effects on recognition performance such as multi-orientation, curved, engraved and blurry.

Although CNN based scene text recognition approaches [3], [4], [5], [6] have achieved significant development, it is still a difficult task for recognizing text content from equipment nameplates in natural scene due to the diverse text appearance and some extreme conditions as shown in Fig. 1. Moreover, since most recognition models are designed to recognize horizontal text patches, the recognition precision will significant decrease for multi-orientation and curved text. To solve these problems, we propose Orientation Robust Scene Text Recognition (ORSTR) model for equipment nameplates in this paper which consists of a rectification module and a recognition network as shown in Fig. 2. Specifically, to handle curved and multi-orientation text, our model first employ a carefully designed rectification module to transform various text regions to near-horizontal regions. The near-horizontal regions will be passed into a deep convolutional neural network to project input text information into a high-

[*] 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)

[†] Xiaolong Chen and Zhengfu Zhang contribute equally to this paper.

[‡] Bin Fu is the corresponding author in this paper.

[1] Xiaolong Chen, Jiangyu Lai and Jian Jiang are with Guangzhou Power Supply Bureau Co. Ltd., Guangzhou, China. {chenxl1116, laijy, jiangj}@guangzhou.csg.cn.

[2] Zhengfu Zhang, Yu Qiao, Zeyu Zhang 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. {zf.zhang, yu.qiao, zy.zhang3, bin.fu}@siat.ac.cn

[3] Zhengfu Zhang, Yu Qiao, Zeyu Zhang and Bin Fu are with SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society.

dimensional feature space. Since the ratio of length and width has large variance, convolutional neural network cannot capture enough sequence information. We employ a shadow recurrent neural network to further extract sequence information to improve recognition performance. Finally, an attention based recognition prediction module will give the prediction for text patches as the output of our text recognition model.

In this paper, we first give a briefly discussion about several related works on scene text recognition in next section. Our proposed recognition method will be presented in Section III, and several extensive experiments will be conducted in Section IV to verify the effectiveness of our recognition model. Finally, a briefly summary will be given for this work.

II. Related Work

Text recognition in natural scene is a challenge task and has achieved significant progress in recent years due to the development of Convolutional Neural Network. A comprehensive study for the development of scene text recognition has been given in [7]. In this section, we only offer a briefly introduction of several state-of-the-art recognition model in recent years. Since existing recognition approaches can be roughly classify into two categories, bottom-up approach and top-down approach, we will introduce these two different approaches separately.

A. Bottom-Up Approach

Early works mainly focus on bottom-up approaches which recognize every individual character and then connect each character into a series of different words. For example, in traditional method, feature vectors are extracted by hand crafted feature extraction module such as strokelet generation [8] and semi-markov conditional random field [9], then the resulting feature vectors are recognized from a character classifier. Since the hand crafted feature cannot capture rich recognition information, the recognition performance is limited. These methods are significant improved by replacing hand crafted feature with the feature extracted from neural networks. For example, Jaderberg et al. [10] proposed a unconstrained text recognition model by employing CNN to extract recognition feature.

B. Top-Down Approach

The other methods employ a top-down fashion to recognize text patches from natural scene images. The top-down recognition approaches directly recognizing an entire text or word from origin text images without any predictions of individual characters. Jaderberg et al. [11] converts text recognition problem into image classification problem which employs a 90k-class classification network to recognize 90k English words. Due to huge categories in classification network, the recognition performance is not satisfactory. Moreover, out-of-vocabulary words cannot be recognized and thus this model cannot be widely used in general case. To settle down this weakness, people employ sequence method to recognize text line with arbitrary length. CRNN [1] employs a CNN to extract recognition information from the input text patches and then uses a RNN to further capture long-range sequence information. In order to model conditional probability for arbitrary-length text, a CTC [12] Loss has been employed in this work. In recent years, attention mechanism has been widely used in recognition models which generates a focusing map for each character position in text regions to improve recognition performance. However, since sequence model is designed for predicting probability in 1D sequence, these models cannot recognition characters and words in multi-orientation and curved text. To solve this problem, inspired by Spatial Transformer Networks [13], ASTER [2] employs a Text Rectification Network to transform multi-orientation or curved text into horizontal text and then employs sequence model to recognize text content in transformed horizontal text region. Various methods have been put forward to recognize multi-orientation and curved texts. For example, [14] design a Iterative Rectification Network to perform thin-plate-spline transformation to refine near-horizontal rectified text regions. Moreover, Symmetry-constrained Rectification Network (ScRN) [15] predicts several local attributes of text instances to obtain more accurate control points for thin-plate-spline transformation.

In this paper, we put forward a text recognition model to recognize characters and words from equipment nameplates in natural images. Since equipment nameplates contain both horizontal, multi-orientation and curved texts as shown in Fig. 1, our model first transforms non-horizontal text to near-horizontal text by a carefully designed rectification model. Then the transformed text is passed to CNN and RNN modules to extract rich recognition information. Finally, we employs an attention based recognition module to produce the prediction for input text regions.

III. The Proposed Method

In this section, we will give a detailed description of our proposed text recognition model for equipment nameplates in natural scene. The pipeline for our proposed recognition method has been shown in Fig. 2.

A. Rectification Module

Since equipment nameplates include some multi-orientation and curved text patches, a rectification module has been adopted to transform irregular text patches into near-horizontal text patches. Our rectification module consists of three parts, text shape regression network, least square refinement module and thin-plate-spine transformation module. The pipeline of our Rectification Module has been shown in Fig. 3.

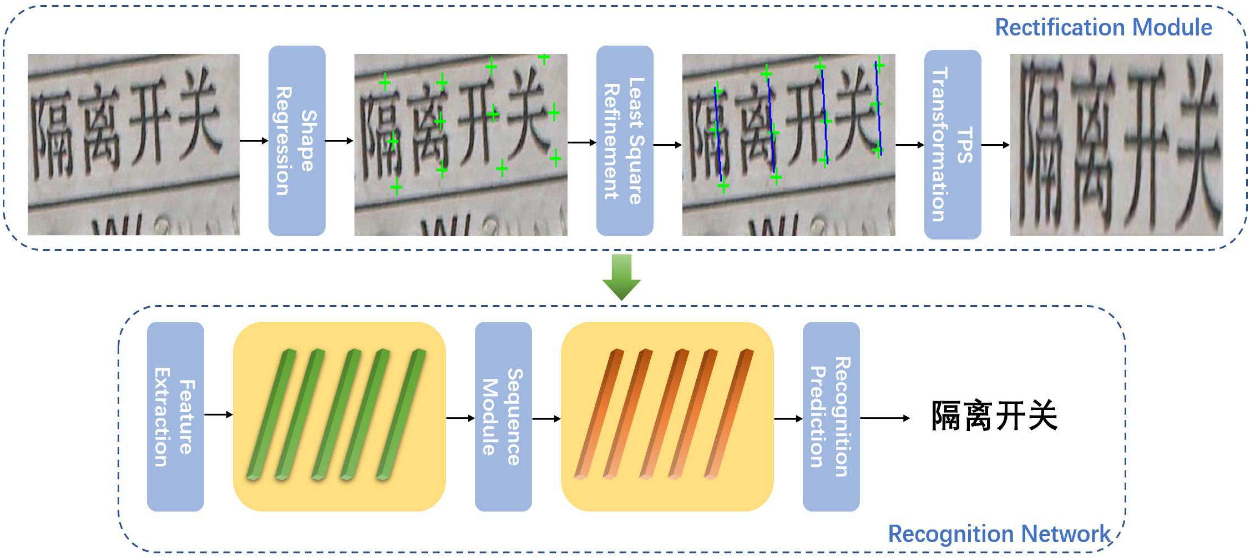

Fig. 2. The schematic diagram for our Orientation Robust Scene Text Recognition (ORSTR) model. The proposed recognition model for equipment nameplates consists of a rectification module and a recognition network. To handle curved and multi-orientation text, our model first employ a carefully designed rectification module to transform various text regions to near-horizontal regions. The near-horizontal regions will be passed into a series of deep neural networks to produce final prediction.

1) Text Shape Regression Network: In order to perform rectification for text patches, we first generate a set of sampling groups by fitting the shape of characters in each text patch. In this work, we employ a convolutional neural network to predict n equal spacing sampling groups $S = [s_1, s_2, \cdots, s_n] \in R^6$ where each sampling group $s_i = (u_i, m_i, l_i)$ consists three points which two points at upper u_i and lower l_i text edges and one center point m_i in the middle of two edge points as shown in Fig. 3(a). Since the prediction of upper and lower edge points maybe not accurate, middle point will be employed to refine the position of them.

2) Least Square Refinement Module: In this work, the middle point has been predicted in order to refine the position of upper and lower edge points. In ideal case, three points (upper, middle and lower) in each sampling group will form a single straight line. However, convolutional neural network cannot regression position coordinate with high precision. Since the size of input text patch is much smaller than object detection, the regression variance will become a serious problem when we perform thin-plate-spline transformation. To settle down this problem, we employ a Least Square Refinement (LSR) Module to produce a stable prediction for upper and lower edge points. Given a specific sampling group $s_i = (u_i, m_i, l_i)$, our LSR module employs Least Square Method (LSM) to obtain a linear function f_i as shown in Fig. 3(b).

Assuming we use Least Square method to fit t points with the equation,

$$f_i = \beta_0 + \beta_1 x \tag{1}$$

From Least Square method, the coefficient of linear function can be determined by

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{2}$$

$$\beta_1 = \frac{Cor(y, x)}{Var(x)} \tag{3}$$

where $\bar{y}$ and $\bar{x}$ are the average of y and x coordinate of t points. Cor(y, x) is the covariance of y and x. Var(x) is the variance of x coordinate. Once we obtain the linear function f_i for each sampling groups, we can employ this function to obtain a refined upper and lower edge points with the following equations,

$$\bar{y}(u_i) = A_i \; x(u_i) + B_i \tag{4}$$

$$y(u_i) = A_i \; \bar{x}(u_i) + B_i \tag{5}$$

$$\bar{y}(l_i) = A_i \; x(l_i) + B_i \tag{6}$$

$$y(u_i) = A_i \; \bar{x}(l_i) + B_i \tag{7}$$

As shown in Fig. 3(c), the refined edge points u_i and l_i can be calculated by:

$$y_r(u_i) = \frac{\bar{y}(u_i) + y(u_i)}{2} \tag{8}$$

$$x_r(u_i) = \frac{\bar{x}(u_i) + x(u_i)}{2} \tag{9}$$

$$y_r(l_i) = \frac{\bar{y}(l_i) + y(l_i)}{2} \tag{10}$$

$$x_r(l_i) = \frac{\bar{x}(l_i) + x(l_i)}{2} \tag{11}$$

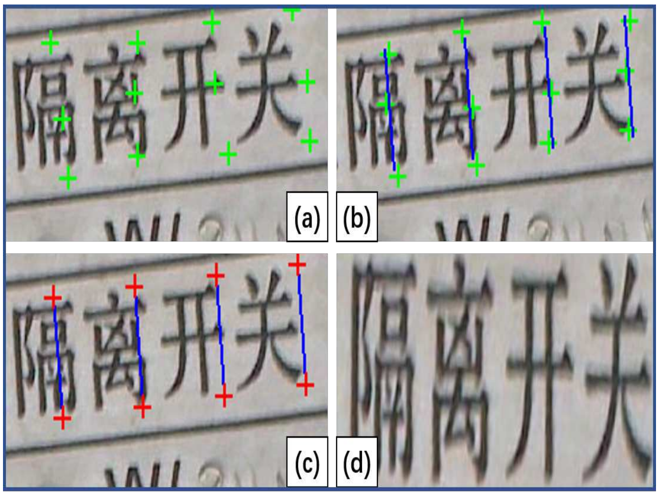

Fig. 3. The schematic diagram for our Rectification Module. (a). A set of sampling groups generated by fitting the shape of characters. (b). The corresponding linear function produced by Least Square Refinement Module. (c). The refined upper and lower edge points for each sampling group. (d). The rectified image after thin-plate-spline transformation module.

3) Thin-Plate-Spline Transformation Module: Multi-orientation or curved text can be transformed to near-horizontal text by thin-plate-spline (TPS) transformation. We separate upper and lower boundaries of text region with equal spacing and employ TPS to transform estimated edge points to these predefined points. Assuming the estimated edge points are S_r and the predefined grid points are S_g, the TPS transformation parameters can be calculated by the following equation [2]

$$T = \begin{pmatrix} H & 1_n & S_r \\ 1_n^T & 0 & 0 \\ S_r^T & 0 & 0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} S_g \\ 0 \\ 0 \end{pmatrix} \tag{12}$$

with

$$H = [U(s - s_1), U(s - s_2), \cdots, U(s - s_n)]^T \tag{13}$$

and

$$U(s) = s^2 \log s^2. \tag{14}$$

Once we obtain TPS transformation parameters, we can perform TPS to reshape irregular text into near-horizontal text as shown in Fig. 3(d).

B. Recognition Network

In this work, our recognition network include three modules. The first module is feature extraction module which takes the rectified RGB image as input and encode RGB image information into a visual feature map. We employ the widely used ResNet [16] as the backbone of our feature extraction module. Since the most text directions are horizontal, the feature map is reshaped into a series of column feature maps which each column feature map encode the information of a fixed-width region in rectified text according to the receptive field of backbone. The second module is sequence module. Each column in feature map is regard as a frame of the sequence and then passed into sequence model. In this work, we follow the common setting [1], [17] to employ Bidirectional LSTM (BiLSTM) to extract sequence feature vector with rich contextual information. Finally, we employ recognition prediction module with attention mechanism [17], [3] to output the prediction for text line.

C. Relation to Other Approaches

Our scene text recognition can read multi-orientation and curved text instance by a carefully designed rectification module. Compared with state-of-the-art recognition model ASTER [2], we propose a Least Square Refinement Module to obtain more accurate control points for performing TPS transformation. The Least Square Refinement Module make our recognition more robust for recognizing multi-orientation and curved text instances.

IV. Experiments

In this section, our scene text recognition approach has been adopted for equipment nameplates. Moreover, several extensive experiments have been conducted to demonstrate the effectiveness of our recognition model.

A. Dataset

1) Synth90K: is a synthetic text dataset [18] with 9 million images and has been widely used for pre-training recognition model for different datasets. All synthetic images in this dataset are regarded as training set.

2) SynthText: is also a synthetic dataset introduced by [20]. This dataset mainly focuses on text detection task, thus it only offers the synthetic images and the corresponding annotated bounding boxes. People modify this dataset for recognition task by cropping text patches using the corresponding annotated bounding box.

3) The Equipment Nameplate Dataset: contains 21365 word instances with 17416 word instances in training set and 3949 word instances in testing set [19]. All word instances are cropped from 502 images which are taken in the wild and collected from several kinds of equipments. The word level instances are carefully annotated with rectangle bounding boxes.

B. Implement Details

We employ a four-convolution-layers shallow neural network which each convolution layer followed with batch normalization, ReLU and Max Pooling (except the last convolution layer) as our Text Shape Regression Network to generate several sampling groups for Least Square Refinement Module. Following [21], we adopt our recognition network. We employ ResNet101 [16] pre-trained on ImageNet dataset [22] as Feature Extraction Module to encode RGB image

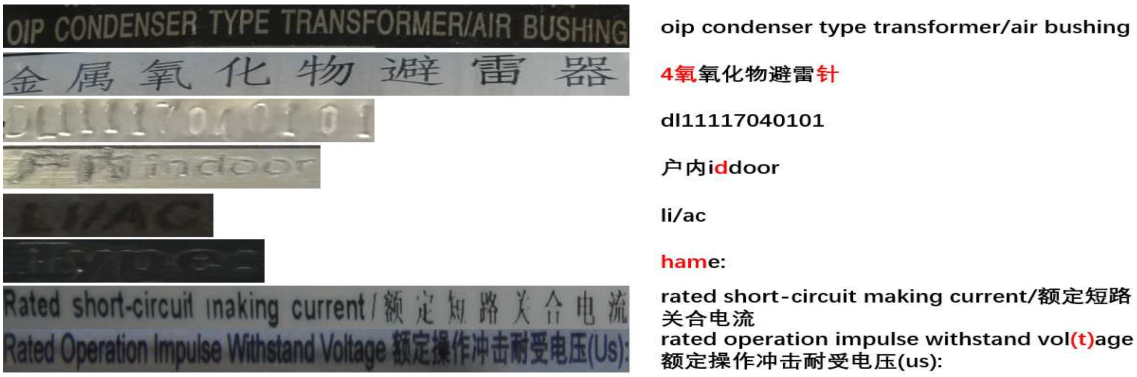
oip condenser type transformer/air bushing
4氧氧化物避雷针
dl11117040101
户内iddoor
li/ac
hame:
rated short-circuit making current/额定短路关合电流
rated operation impulse withstand vol(t)age 额定操作冲击耐受电压(us):

Fig. 4. The scene text recognition results of our ORSTR model. We select four pairs of recognition results to further analyse recognition performance, include curved, engraved, extreme illumination and extremely long patches. For each pair, the upper patch is positive sample while the lower patch is negative sample. The character in red color is the false prediction in our model.

information into a visual feature map. A CRNN-liked [1] two-layers Bidirectional LSTM are used to extract sequence information from feature maps. Finally, one layer attention decoder [17], [3] has been adopted to output predictions for text patches.

We merge Synth90K and SynthText datasets resulting 14.4 million synthetic images to pre-train our model by using AdaDelta [23] optimizer with the parameter $\rho = 0.95$. The Equipment Nameplate Dataset are training 300K iterations on a single Tesla K40 GPU. All experiments are performed on PyTorch platform.

C. Ablation Study

In this sub-section, we perform several extensive experiments to verify the effectiveness of our model for recognizing text patches in equipment nameplates.

TABLE I
ABLATION STUDY FOR THE INFLUENCE ON RECOGNITION PERFORMANCE WITH/WITHOUT RECTIFICATION MODULE

	30K	60K	90K	120K	300K
Att without TPS	88.09	88.57	88.65	88.85	89.15
Att with 4 groups	88.39	88.72	88.82	89.15	89.31

1) Recognition Performance with/without Rectification Module: As we have discussed before, there are several multi-orientation and curved text patches in Equipment Nameplate dataset. In this paper, we have employed a Rectification Module to settle down this problem. In order to verify this design, we train our model with and without Rectification Module and present experimental results in TABLE I. From this table, we find that Rectification Module significant improve recognition performance from 89.15% to 89.31% in terms of recognition accuracy which demonstrate the effectiveness of our Rectification Module.

TABLE II
ABLATION STUDY FOR THE INFLUENCE ON RECOGNITION ACCURACY OF THE NUMBER OF SAMPLING GROUPS IN RECTIFICATION MODULE

	30K	60K	90K	120K	300K
Att with 4 groups	88.39	88.72	88.82	89.15	89.31
Att with 8 groups	88.82	89.00	89.15	89.15	89.53
Att with 12 groups	89.10	89.41	89.61	89.98	90.08
Att with 16 groups	89.23	89.53	89.86	90.06	90.06
Att with 20 groups	89.43	89.86	89.86	89.93	90.08
Att with 24 groups	89.87	90.19	90.19	90.37	90.75

2) The Number of Sampling Groups in Rectification Module: In our model, the number of sampling groups is an important hyper-parameter since more sampling points will give a better fitting for irregular shape and will offer more useful information for TPS transformation. Therefore, we conduct several extensive experiments with different numbers of sampling groups in rectification module to verify this statement. Experimental results with a set of different iterations have been shown in TABLE. II. From this table, we can draw the following conclusions:

(1). The recognition performance will significant improved when we employ more sampling groups to perform TPS transformation since more sampling groups will give a better fitting for irregular text line in horizontal directions.

(2). TPS module can efficient improve recognition performance and only four sampling point will significant increase recognition accuracy.

(3). The number of sampling groups is an important hyper-parameter since the recognition accuracy will improve about 1.5% by different settings.

As we have pointed out in (1), the more sampling groups, the better recognition performance. Therefore, we set the number of sampling groups as 24 for all experiments in the following.

D. Experimental Results and Analysis

We employ our scene text recognition model to recognize text patches in Equipment Nameplate test set [19]. Our model achieves 90.8% accuracy with 24 sampling groups and 300K iterations which outperform baseline model [19] about 0.8%. This result illustrates that our model can recognize text patches on equipment nameplates with high precision.

We further analyse our recognition model by selecting several correct and false examples as shown in Fig. 4. Several conclusion can be drawn from this figure. (1). Thanks to the proposed rectification module, our model can recognize multi-orientation and curved text lines. (2). Although engraved characters are difficult to recognize due to the same appearance with background, our model can correctly identify engraved characters on equipment nameplate. (3). Moreover, for extreme illumination conditions, our model can give satisfactory results in some case which demonstrate that our recognition model is robust in various environment. (4). Our model will give false prediction on extremely long text in some case. Since most text patches in Equipment Nameplate Dataset are shot text such as name and value of parameter, the dataset cannot offer enough training examples for long text. Moreover, it is a inherent limit for CNN and RNN model to model extremely long sequence due to the limited receptive field for CNN model and fast forgotten for RNN model.

V. Conclusion

In this paper, to recognize text instance in equipment nameplates, we put forward Orientation Robust Scene Text Recognition (ORSTR) model which consist a Rectification Module and a Recognition Network. The Rectification Module is employed to transforming irregular text such as multi-orientation and curved texts to near-horizontal text. The Recognition Network include three parts, which are a CNN based Feature Extraction Module, a RNN based Sequence Module and an attention based Recognition Prediction Module. Our proposed model achieves 90.8% accuracy on test set of Equipment Nameplate Dataset which outperforms baseline model about 0.8%.

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] B. Shi, X. Bai and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 39(11), p. 2298, 2018.

[2] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao and X. Bai, "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41(9), p. 2035, 2019.

[3] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu and S. Zhou, "Focusing Attention: Towards Accurate Text Recognition in Natural Images," International Conference on Computer Vision, p. 5086, 2017.

[4] F. Bai, Z. Cheng, Y. Niu, S. Pu, and S. Zhou, "Edit Probability for Scene Text Recognition," In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1508-1516, 2018.

[5] X. Yang, D. He, Z. Zhou, D. Kifer and C. L. Giles, "Learning to read irregular text with attention mechanisms," International Joint Conference on Artificial Intelligence, p. 3280, 2017.

[6] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu and S. Zhou, "AON: Towards Arbitrarily-Oriented Text Recognition," Computer Vision and Pattern Recognition, p. 5571, 2018.

[7] S. Long, X. He and C. Yao, "Scene Text Detection and Recognition: The Deep Learning Era," arXiv: Computer Vision and Pattern Recognition, 2018.

[8] M. Jaderberg, A. Vedaldi and A. Zisserman, "Deep Features for Text Spotting," European Conference on Computer Vision, p. 512, 2014.

[9] J. Seok and J. H. Kim, "Scene text recognition using a Hough forest implicit shape model and semi-Markov conditional random fields," Pattern Recognition, vol. 48(11), p. 3584, 2015.

[10] M. Jaderberg, K. Simonyan, A. Vedaldi and A. Zisserman, "Deep Structured Output Learning for Unconstrained Text Recognition," International Conference on Learning Representations, 2018.

[11] M. Jaderberg, K. Simonyan, A. Vedaldi and A. Zisserman, "Reading Text in the Wild with Convolutional Neural Networks," International Journal of Computer Vision, vol. 116(1), p. 1, 2016.

[12] A. Graves, S. Fernandez, F. J. Gomez and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," International Conference on Machine Learning, p. 369, 2006.

[13] M. Jaderberg, K. Simonyan, A. Zisserman and K. Kavukcuoglu, "Spatial transformer networks," Neural Information Processing Systems, p. 2017, 2015.

[14] F. N. Zhan and S. J. Lu, "Esir: End-to-end Scene Text Recognition via Iterative Image Rectification," Conference on Computer Vision and Pattern Recognition, pp. 2059-2068, 2019.

[15] M. K. Yang, Y. S. Guan, M. H. Liao, X. He, K. G. Bian, S. Bai, C. Yao and X. Bai, "Symmetry-constrained Rectification Network for Scene Text Recognition," arXiv preprint arXiv:1908.01957, 2019.

[16] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Computer Vision and Pattern Recognition, p. 770, 2016.

[17] B. Shi, X. Wang, P. Lyu, C. Yao and X. Bai, "Robust Scene Text Recognition with Automatic Rectification," Computer Vision and Pattern Recognition, p. 4168, 2016.

[18] M. Jaderberg, K. Simonyan, A. Vedaldi and A. Zisserman, "Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition," arXiv: Computer Vision and Pattern Recognition, 2014.

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

[20] A. Gupta, V. Andrea and Z. Andrew, "Synthetic Data for Text Localisation in Natural Images," Computer Vision and Pattern Recognition, p. 2315, 2016.

[21] J. Baek, et al, "What is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis," arXiv: Computer Vision and Pattern Recognition, 2019.

[22] O. Russakovsky, et al, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115(3), p. 211, 2015.

[23] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," arXiv: Learning, 2012.