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

The Equipment Nameplate Dataset for Scene Text Detection and Recognition*

Xiaolong Chen[†1], Zhengfu Zhang[†23], Yu Qiao[23], Pu Zhang[1], Lanqing Guo[23]
Wenrui Chen[1], Chen Chen[23] and Bin Fu[‡23]

***Abstract*— In this paper, we introduce the Equipment Nameplate Dataset, a large dataset for scene text detection and recognition. Natural images in this dataset are taken in the wild and thus this dataset includes various intra-class inconsistency such as ill illumination conditions and partly occluded, which makes our dataset more challenging than other datasets. In order to make people train detection and recognition model separately, we annotate our dataset not only word instance, but also text region by using rectangle bounding boxes. Some detailed statistics information about our dataset will be given so that people can use them to analyse and develop their own models. Moreover, we use our dataset to test some famous detection and recognition models and present the corresponding results in order to make researcher compare them with their own models. Dataset will be publicly available on the website.**

***Index Terms*— Artificial Intelligence, Computer Vision, Scene Text Detection, Scene Text Recognition**

I. Introduction

Recognizing text words in the natural images is one of most significant tasks in computer vision, since various high level applications have strongly rely on this task such as transcribed text from manuscripts to electronic document. Roughly speaking, this task can be divided into two kinds of problems according to the input image: the well-organized images and the natural images. The well-organized images such as printed documents have been widely studied and lots of applications have emerged in past few years. However, the text detection and recognition in natural images are challenging tasks in recent years, since lots of factors will give negative effects on detection and recognize precisions such as distorted and illumination differences as shown in Fig. 1. The text in printed document has less intra-class inconsistent and large inter-class difference, while the text in natural image has large intra-class inconsistent and less inter-class difference which decrease the detection and recognition precision significantly.

The deep learning technique has widely used in many computer vision fields in recent years. The carefully-designed deep network together with some simple non-linear activate functions can achieve high performance and beat the tradition method with a large margin in most computer vision tasks. However, there are millions of parameters which have to be determined in this method and thus a large dataset is required to train deep network in order to obtain a high performance. Moreover, people have shown that, the size of the training dataset is the most important factor in order to obtain a high performance. Several large datasets (such as ImageNet dataset [1] and Microsoft COCO dataset [2]) have been published in recent year which make a significant progress in the corresponding tasks.

In order to improve the performance of scene text detection and recognition for equipment nameplates, we present a large equipment nameplate dataset in natural images. There are totally 502 equipment nameplates images and 870 different words (include Chinese and English) in this dataset. Equipment nameplates are collected with different illumination conditions and backgrounds. The dataset includes regular rectangle nameplates, inclined nameplates, annular nameplates and some irregular nameplates. We have carefully annotated all equipment nameplates in this dataset: Firstly, since the salience detection can remove background information and improve the precision of text detection, we use bounding box to annotate this information of equipment nameplates. People can use this information to get rid of the unrelated background informations before the text detection. Secondly, we annotate the bounding box of every text region in the equipment nameplate, therefore people can use it to train their scene text detection model. Finally, we annotate every word instance in the corresponding bounding box in order to make people train their text recognition model.

Moreover, we have trained some popular state of art models to show the quantity of this dataset, both for text detection and text recognition. The results of these models can be regarded as the baseline of our dataset which are helpful for researchers to evaluate and improve the performance of

*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, Pu Zhang and Wenrui Chen are with Guangzhou Power Supply Bureau Co. Ltd., Guangzhou, China. {chenxl1116, zhangp, chenwr}@guangzhou.csg.cn.

[2] Zhengfu Zhang, Yu Qiao, Lanqing Guo, Chen Chen 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, lq.guo, chen.chen1, bin.fu}@siat.ac.cn

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

978-1-7281-6321-5/19//$31.00

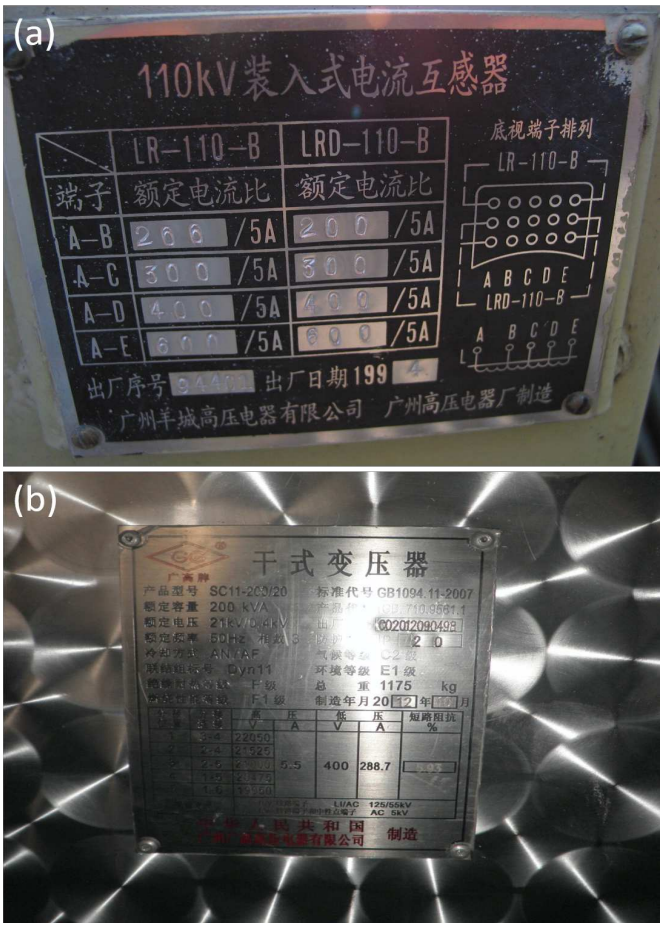

Fig. 1. Some factors will give negative effects on detection and recognize precisions such as distorted (a) and illumination differences (b).

their own model. We will release this equipment nameplate dataset with the source codes and results of baseline models.

The rest of paper is organized as follows. In the Section II, the related works will be summarized and some important works in text detection, text recognition and text dataset will be outlined. In the Section III, the details of equipment nameplate dataset will be discussed. The performance of baseline models in text detection and recognition tasks will be given in the Section IV. Finally, we will give briefly summary and discussion of our dataset.

II. Related work

In this section, we give a briefly review about the development of the text datasets, text detection and text recognition model in recent years.

A. Datasets of Text in Natural Images

The ICDAR datasets are a series of text datasets published in different years and play a significant role in the development of text detection and recognition tasks. The dataset released from ICDAR 2003 Robust Reading Competition [3] is the first text detection and recognition dataset. This dataset was modified and extended in ICDAR 2013 Robust Reading Competition [4] and ICDAR 2015 competition [5] by adding small, occluded, blurred and multi-oriented images in order to simulate real world environment. The Chinese Text in the Wild (CTW) dataset [6] is the largest text dataset. It include 32285 high resolution natural images and 1018402 character instance in total. Although the text information have been carefully made in annotation process, this dataset only annotates Chinese character instance and ignore other character instance such as English and French which causes this dataset is not widely used in non-Chinese community. Since the non-straight line text is a more challenging subtask in text detection and recognition, Total Text dataset [7] has been published to solve this problem. There are 1555 images in this dataset include horizontal text, multi-oriented text and curved text. People can use this dataset to make their model more robust in real world application.

B. Scene Text Detection and Recognition

Text detection and recognition can be viewed as two separate tasks in many text recognition algorithms. Text detection algorithm is designed for finding text from the input image and giving the text region as the output, while the recognition algorithm will use this text region to recognize the word in it. Therefore, these two steps are both important for text recognition and a lot of significant models have been put forward to improve the performance of them. In the following, we will review them separately.

Following the famous objective detection platform SSD [8], several scene text detection models [9], [10] have been put forward. These methods predict the text region and refine the location of detected region with respect to the pre-defined regular grid of the input natural image. Since these methods only operate at the pre-defined locations, the detection models can achieve extremely high speed. The Faster R-CNN model [11] is also a famous object detection platform, and several text detection methods [12], [13] are based on it. These models extract lots of region proposals on feature maps and then use a classifier to judge whether the selected proposals are text region. Text regions are selected by the probability of text/non-text and a post-process is needed to regression the position of selected regions.

After obtaining text regions from the natural image, recognition model will use these regions to recognize character instance. There are two approaches to finish this task, namely, connectionist temporal classification models [14], [15], [17], [18] and attention mechanism models [19], [20], [21], [22], [23]. The first approach uses recursive neural network to calculate the conditional probability of label sequence given per-frame prediction. The pipeline of this model typically follows Shi's work [17], which can be roughly divided into

Index	transcript	invalid	indistinct
1	LC	false	true
2	创立电气	false	false
3	ChuangLi	true	false
4	配电箱	false	false
5	型号	false	false
6	XM	false	false
7	电压	false	false
8	380V	false	false
9	编号	false	false
10	年月	false	false
11	201389	false	false
12	中国广东创立电气设备有限公司	false	false

Fig. 2. The annotated equipment nameplate in our dataset. We use bounding box to annotate nameplate region and every text region in the image. Moreover, transcript and two additional attributes for each region have been offered.

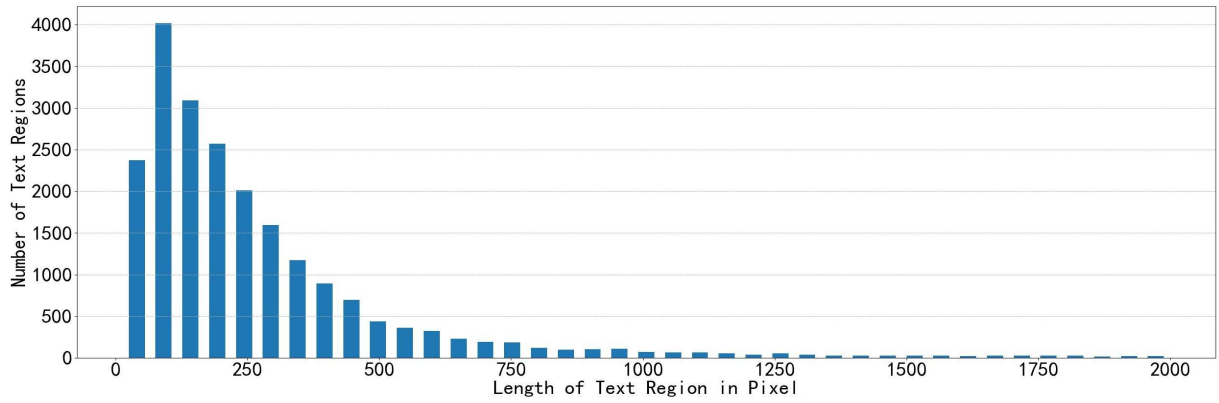

Fig. 3. The length distribution of text region. The distribution of most bounding boxes is in the range [50, 200] pixels, and some bounding boxes have extremely long width.

three steps: the feature vectors are first extracted from text regions by a convolutional network and then a recursive neural network is used to predict the label distribution for each frame, finally, a post-process translates per-frame prediction into the final label sequence. The attention model was first introduced in [24]. This approach uses a encoding structure to extract feature vectors of text regions and a decoder structure to output character instance of these text regions. Based on this pipeline, some modifications have been made in recent years.

III. Equipment Nameplate Dataset

In this section, we give a detailed description about our Equipment Nameplate Dataset, which is a large nameplate dataset in the natural image with Chinese and English text. An overview of this dataset will be discussed firstly, then the annotation method will be presented, and the statistics of this dataset will be given at the end of this section.

A. Overview

In order to create this dataset, we totally collect about 10000 equipment nameplate images in natural scene. All images are taken from different equipments in different places at different time. The resolution of the image is around

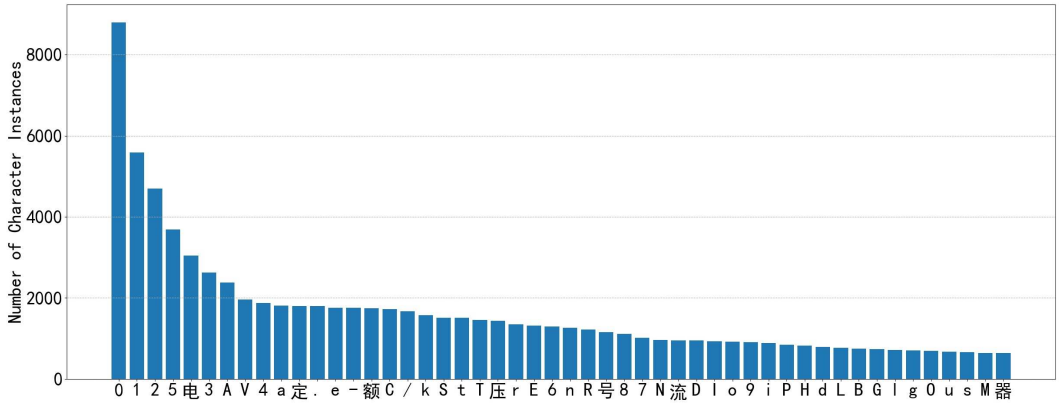

Fig. 4. The frequency of top 50 categories of our dataset. The distribution of categories in our dataset is balanced, therefore, the training process will not suffer data imbalanced problems when people train their model.

$3,000$ pixel in width and $2,000$ pixel in height, therefore the resolution is large enough for any pre- or post-progresses such as crop and resize. We select 502 images to create our dataset, the selected images include different illuminate conditions, planar text, raised text, inclined text, annular text, etc, thus own Equipment Nameplate Dataset offers enough images to handle intra-class inconsistence and inter-class difference problems.

In order to evaluate the performance of different models, we split our dataset into text detection dataset and text recognition dataset. The text detection dataset includes 502 images with 454 images in training set and 48 images in testing set. The text recognition dataset includes 21365 word instances with 17416 word instances in training set and 3949 word instances in testing set.

B. Annotation

In this Section, we give a detailed description about the annotation process of our dataset. The annotated nameplate has been shown in Fig. 2 as an example for our dataset. In order to keep high quantity of annotation process, a professional annotation company was employed and the annotated images have been double checked. Therefore, the annotation in our dataset is accurate and reliable. For each natural image in our dataset, it includes the equipment nameplate and background which will decrease the detection performance since it may offer unrelated information. In order to get rid of the unrelated background information, we annotate the position of equipment nameplate in each natural image with the coordinate of four corner points. Therefore, people can use these points to train a deep network to extract nameplate region and discard unrelated background region. For the text detection, we annotate all text regions in each equipment nameplate with rectangle bounding boxes. The corresponding word instances in each region are given and people can retrieve them with respect to the text region label.

For each text region, apart from the position of bounding box and the word instance of it, there are two additional attributes in our annotation. The first attribute illustrates whether annotated region is valid in text detection and recognition tasks. Equipment nameplates offer some other useful non-text information about the corresponding equipment such as the circuit diagram. In the training process, we can ignore the invalid information rather than setting these region as the false region. Experiment shows that the detection performance will be significant improved if we ignore the invalid region in training process. The second attributes illustrates whether the text region is indistinct. Since the equipment in the wild will be damaged in some case, the some parts of text region may be indistinguishable by people. We set these text regions as the invalid regions and they can be ignored in the training process.

C. Statistics

The equipment nameplate dataset has 502 images with 21365 word instances. In the following, we will give a detailed statistics information of our dataset and people can use these information to analyse and design their models.

The length distribution of text region bounding boxes has been shown in Fig. 3. From this figure, we find that the length of most bounding boxes is in the range [50, 200] pixels, and some bounding boxes have extremely long width which may cause a low performance of detection if the receptive field of detection model is not enough to cover this region. Therefore, large receptive field in horizontal direction is important for detection model.

The frequency of top 50 categories of our dataset has been plotted with respect to training set and test set in Fig. 4. The figure shows that the distribution of top 50 categories in our dataset is balanced. Therefore, our equipment nameplate dataset can measure scene text recognition performance without the problem of imbalanced data samples.

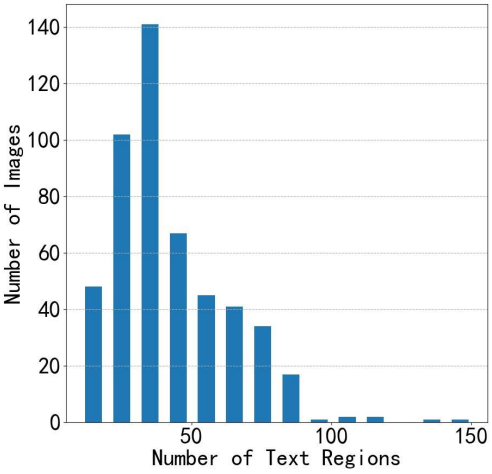

Fig. 5. The number of text regions in each image. Our dataset is diversity and thus our dataset can be used to train a robust scene text recognition model.

Finally, Fig. 5 indicates the number of text regions in each image, we find that our dataset is diversity in this aspect and this diversity is useful to train a robust model and avoid over-fitting in training process.

IV. Baseline Algorithms

In this section, we use some popular detection and recognition algorithms as our baseline models and report the performance of them. These baseline models can be viewed as the benchmark of our dataset in order to evaluate and analyse other models. All source codes and results will be released soon.

In the following, we will discuss text detection algorithm firstly.

A. Detection

As we have discussed in introduction part, a lot of detection algorithms have been put forward in recent years. Among them, several algorithms have achieved great success and have been widely used in industrial community. In this section, we select CTPN [12] as our baseline models and employ our Equipment Nameplate Dataset to test the detection performance of it. Since most existing scene text detection models are designed for detecting horizontal text region, we remove unrelated background information and rectify equipment nameplate images into near-horizontal by employing perspective transformation before text detection.

The detailed implement and setting has been include in [24] and we just give a brief introduction of this model. The CTPN model uses a sliding window in feature maps and produces a sequence of proposals by predicting the vertical location from a series predefined vertical anchors (10 anchors has been used in our setting) with fixed horizontal location. The non-maximum suppression (NMS) has been used to generate text proposals with the text/non-text score > 0.7. Finally, the text region will be generated from these proposals by using side-refinement process to connect them into a line.

We adopt CTPN model on Tensorflow platform and train this model on our equipment nameplate dataset. CTPN model achieves 91.2% precision, 92.6% recall and 91.9% H-mean which demonstrate our dataset can efficiently evaluate the performance of scene text detection model. The detection results have been shown in Fig. 6. From this figure, we find that our detection dataset is a challenging dataset since it offers some hard text region samples such as engraved text regions and disturbed regions which will significant decrease detection performance in natural scene.

B. Recognition

After obtaining text region of the image, we need to extract the text content in the corresponding text region and obtain the final recognition result. We adopt the state-of-the-art text recognition model CRNN [17] for the text recognition task. Since the variance of scales and width/height ratio is huge between different text regions, zero-padding have been employed to normalize text region into a predefined input size. The renormalized text regions are feed to ResNet [26] network to extractor context feature for text recognition. We pass the resulting feature map to a two-layers BiLST to extract sequence information and then employ CTC [27] method to give the final predictions. For our Equipment Nameplate Dataset, the recognition accuracy of CRNN is 89.9%.

V. Conclusions

In this paper, we introduce a large text detection and recognition dataset, named Equipment Nameplate Dataset. It includes 502 natural images taken from the wild with 21365 carefully annotated text regions. The statistics of this dataset are given and this will give useful information to analyse and design detection and recognition models. Moreover, scene text detection and recognition models have been tested on this dataset and the results can be served as the benchmark of this dataset. The dataset will become public available after this paper publish.

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] J. Deng, et al, "Imagenet: A Large-scale Hierarchical Image Database." In IEEE Conference on Computer Vision and Pattern Recognition pp. 248-255, 2009.

Fig. 6. The result of text detection algorithm. The detection results show that our dataset is a challenging dataset since it includes some difficult samples such as engraved text regions and disturbed regions.

[2] T. Y. Lin, et al, "Microsoft COCO: Common Objects in Context," In European Conference on Computer Vision, pp. 740-755, 2014.

[3] S. M. Lucas, et al, "ICDAR 2003 Robust Reading Competitions," In Seventh International Conference on Document Analysis and Recognition, pp. 682-687, 2003.

[4] D. Karatzas, et al, "ICDAR 2013 Robust Reading Competition," In 2013 12th International Conference on Document Analysis and Recognition, pp. 1484-1493, 2013.

[5] D. Karatzas, "ICDAR 2015 Competition on Robust Reading," In 2015 13th International Conference on Document Analysis and Recognition, pp. 1156-1160, 2015.

[6] T. Yuan, et al, "A Large Chinese Text Dataset in the Wild," Journal of Computer Science and Technology, vol. 34, pp. 509-521, 2019.

[7] CK. Chng and CS. Chan, "Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition," In 2017 14th International Conference on Document Analysis and Recognition, pp. 936-942, 2017.

[8] W. Liu, et al, "SSD: Single Shot Multibox Detector," In European Conference on Computer Vision, pp. 21-37, 2016.

[9] B. Shi, X. Bai and S. Belongie, "Detecting Oriented Text in Natural Images by Linking Segments," In IEEE Conference on Computer Vision and Pattern Recognition, pp. 2550-2558, 2017.

[10] M. Liao, B. Shi, X. Bai, X. Wang and W. Liu, "Textboxes: A Fast Text Detector With a Single Deep Neural Network," In Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[11] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-cnn: Towards Real-time Object Detection with Region Proposal Networks," In Advances in Neural Information Processing Systems, pp. 91-99, 2015.

[12] Z. Tian, W. Huang, T. He, P. He and Y. Qiao, "Detecting Text in Natural Image with Connectionist Text Proposal Network," In European Conference on Computer Vision, pp. 56-72, 2016.

[13] J. Ma, et al, "Arbitrary-oriented Scene Text Detection via Rotation Proposals," IEEE Transactions on Multimedia, vol. 20, pp. 3111-3122, 2018.

[14] G. Alex, M. Liwicki, H. Bunke, J. Schmidhuber and S Fernndez, "Unconstrained on-line handwriting recognition with recurrent neural networks," In Advances in neural information processing systems, pp. 577-584, 2008.

[15] B. Su and S. Lu, "Accurate Scene Text Recognition Based on Recurrent Neural Network," In Asian Conference on Computer Vision, pp. 35-48, 2014.

[16] B. Shi, X. Wang, P. Lyu, C. Yao and X. Bai, "Robust Scene Text Recognition with Automatic Rectification," In IEEE Conference on Computer Vision and Pattern Recognition, pp. 4168-4176, 2016.

[17] 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, pp. 2298-2304, 2016.

[18] F. Yin, Y. C. Wu, X. Y. Zhang and C. L. Liu, "Scene Text Recognition with Sliding Convolutional Character Models," arXiv preprint arXiv:1709.01727, 2017.

[19] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu and S. Zhou, Focusing Attention: Towards Accurate Text Recognition in Natural Images," In Proceedings of the IEEE International Conference on Computer Vision, pp. 5076-5084, 2017.

[20] C. Y. Lee and S. Osindero, "Recursive Recurrent Nets with Attention Modeling for OCR in the Wild," Conference on Computer Vision and Pattern Recognition, pp. 2231-2239, 2016.

[21] 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.

[22] Z. C. Liu, Y. X. Li, F. B. Ren, W. L. Goh and H. Yu, "Squeezedtext: A Real-time Scene Text Recognition by Binary Convolutional Encoder-decoder Network," In Thirty-Second AAAI Conference on Artificial Intelligence, 2018

[23] F. N. Zhan and S. J. Lu, "Esir: End-to-end scene text recognition via iterative image rectification," Computer Vision and Pattern Recognition, pp. 2059-2068, 2019.

[24] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv preprint arXiv:1409.0473, 2014.

[25] J. Lai, et al, "Robust Text Line Detection in Equipment Nameplate Images," unpublished.

[26] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[27] A. Graves, S. Fernndez, F. Gomez and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," In Proceedings of the 23rd International Conference on Machine Learning, pp. 369-376, 2006.