ICIC Express Letters
Part B: Applications
Volume **13**, Number **4**, April **2022**

ICIC International ©2022 ISSN 2185-2766
pp. **363–370**

RESEARCH AND ANALYSIS OF SCENE TEXT DETECTION AND RECOGNITION TECHNOLOGY BASED ON DEEP LEARNING

Yanju Liu[1,2,*], Xinhai Yi[2], Yange Li[2], Huiyu Zhang[2] and Yanzhong Liu[2]

[1]School of Mathematics and Information Science
Nanjing Normal University of Special Education
No. 1, Shennong Road, Nanjing 210038, P. R. China

[2]School of Computer and Control Engineering
Qiqihar University
No. 42, Wenhua Avenue, Jianhua District, Qiqihar 161000, P. R. China
{ yixinhai000; 15146692464 }@163.com; 1196744431@qq.com; yanzhongliu@qqhru.edu.cn
*Corresponding author: yanjuliu@qqhru.edu.cn

Received August 2021; accepted November 2021

Abstract. *With the development of deep learning technology in the field of computer vision, there are breakthroughs in scene text detection and text recognition technology. Affected by extreme lighting, occlusion, blur, multi-direction and multi-scale in natural scenes, there are still huge challenges facing unconstrained scene text detection and recognition. In this paper, the basic concept of the problem is introduced, the scene text detection and text recognition technology are deeply studied from the perspective of deep learning, and the method and regression based on segmentation in the text detection technology are summarized. The combination of the advantages of the method can solve the problem of low recall rate of small text areas, while adapting to multi-scale text. Through the combination of the CTC mechanism and the attention mechanism in the text recognition method, mutual supervision can be achieved, the recognition performance is improved, and the error rate of long text recognition is reduced.*
Keywords: Deep learning, Computer vision, Natural scene, Text detection, Text recognition

1. **Introduction.** Scene Text Recognition (STR), a sub-problem of Optical Character Recognition (OCR) [1], is to extract text from natural scenes and convert it into character. As the demand for text recognition has become more and more complex, the demand for text recognition in natural scenes has also become stronger. However, there are still great difficulties to achieve high recognition accuracy in natural scenes, such as complex background and lighting, multi-direction and multi-scale, multi-language, long text lines, curved text, and noise interference [2]. Scene text recognition technology based on deep learning solves the above problems effectively, making the recognition method more flexible and robust.

In recent years, the introduction of some new test databases and some new test results, as well as the application of some new in-depth learning methods in the field of text detection and recognition in natural scenes, have greatly promoted the development of related technologies. This paper summarizes the relevant literature in recent years, and divides the text detection and recognition of natural scenes based on in-depth learning into text detection methods and text recognition methods according to the model functions. Each type of method can be divided into several categories according to the technical characteristics of the implementation. These model methods will be introduced and analyzed in detail in the future, in order to provide some reference and help for researchers.

DOI: 10.24507/icicelb.13.04.363

2. **Natural Scene Text Detection Method Based on Deep Learning.** Good text detection results directly affect the performance of text recognition, so it is necessary to introduce a scene text detection method based on in-depth learning. In the method based on deep learning, the effective feature is to be learned directly from the training data and the above bottlenecks are broken and the detection method is more flexible.

2.1. **Detection method based on regression.** The default detection border parameters are initialized in the method, and the regression method is used to continuously learn the parameter values to fit the text sample area. Generally, the detection method based on Convolutional Neural Networks (CNN) is to input several predicted candidate regions into the CNN for feature extraction, and then to classify the candidate regions, and to determine whether target instances are contained or not.

Liao et al. [4] designed multiple default boxes with different scales, and set a vertical offset for each default box to avoid poor detection performance for the sparse vertical directions between the boxes. However, it only detects horizontal text, and the proportionate single rectangle is no longer used to handle text that is more curved or rotated. For this reason, Liao et al. [5] detected multi-directional text by regressing the endpoint coordinates of the text polygon. Rotated text can be detected by the method effectively. In the test phase, non-maximum suppression is used to merge the results of all text box layers. However, the whole network steps are complicated and the training time is long. Zhou et al. [6] proposed the Efficient and Accuracy Scene Text (EAST) model with two steps only to optimize the detection process. Different levels of feature maps are extracted from the input image in the fully convolutional networks stage, and the features are merged from top to bottom in the non-maximum suppression stage. EAST model can detect rotated text, and reduce the intermediate steps and components such as candidate box suggestion and word partition. It improves the processing efficiency and performance effectively. However, the receptive field of EAST model is limited by the size of network receiving domain, and its performance is poor in detecting long text.

The performance of the bottom-up method is much better than the top-down method in detecting curved text and long text, but its effect is poor in detecting dense text in scenes, and the post-processing is complex. Shi et al. [7] proposed a bottom-up detection method named Seglink. In this method, segment is used to cover the multi direction border of a word or a part of a text line, and the link links two segments to indicate that they belong to the same text content. The relevant segments are fused into the final text line according to the confidence score of segment and link. The limitation of the artificially set default frame ratio on the results is removed effectively in this model. However, the curved text and text lines with large character spacing are still not detected in the model. Distortion or curve text cannot be detected because segments combining algorithm uses line fitting when merging. Here, the merging algorithm can be modified to detect distortion or curve text. Tang et al. [8] solved the two problems above effectively by Instance-aware Component Grouping (ICG). In ICG, the relationship between segments is defined as attractive link or repulsive link, which is used to assist the separation of dense text. The instance aware loss function is used to give more loss weight for the detection area and link with poor detection, so as to realize the post-processing. In other bottom-up models, the detection performance of dense text can also be improved by using ICG. However, the model is lack of semantic information, and the training model's dataset has only a small amount of vertical text, so it is easy to make mistakes in the judgment of text line direction and the failure of detection.

Previous studies were based on anchor box, a two-dimension box [4,5]. If the two-dimensional problem is decomposed into one-dimensional problems, the problem is much simpler and the parameter space is much less.

2.2. **Detection method based on segmentation.** In scene text detection method based on segmentation, a classification problem of text/background is usually regarded the text detection problem based on idea of semantic segmentation. Long et al. [9] proposed TextSnake model based on the circle, which is shown in Figure 1. The text center line, text region and the representation attributes of these circles are predicted continuously by fully convolutional networks. Multiple circles are stacked to form a sequence to represent the geometric features of text lines. The shape and process of text instances can be accurately predicted by this method, mainly because of text center line mechanism. Text center lines can be seen as the skeleton supporting text instances, on which geometric attributes provide more detail. However, this method is tested under the premise of the "snake-shaped" text line instance, that is, there is a unique start point and end point in the text line. There may be detection error in the condition of connections or overlaps between multiple text lines.

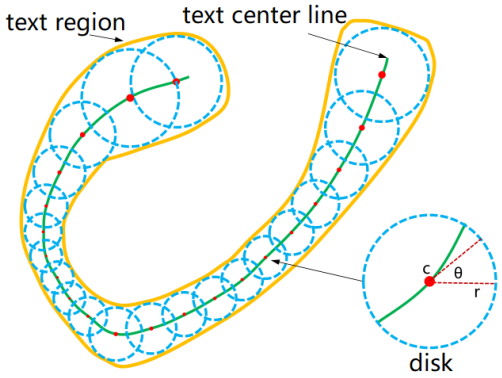

FIGURE 1. Illustration of the TextSnake representation

Wang et al. [10] proposed Progressive Scale Expansion Network (PSENet) that cannot be expanded gradually from small to large until the size of the kernel is equal to that of the original text by the incremental scaling algorithm based on breadth first search according to the scale of the kernel. The dense of text can be separated by smallest kernel and the final bounding box can be obtained by corresponding methods according to the shape of the text. PSENet is more accurate for text line segmentation and inter character segmentation of dense text and it is robust for any shape text. However, the network structure and post-processing process of the model are complex, which is needed a long processing time. For this reason, Wang et al. [11] reduced the computational cost and the complexity of post-processing using lightweight CNN for feature extraction. However, the feature sensing field extracted by lightweight network is smaller and the expression ability is weaker. Then the segmentation head is used to refine the features to solve the above problems. The segmentation head consists of a feature pyramid enhancement module and a feature fusion module. Feature pyramid enhancement module is a U-shaped cascadable module that feature depth and expression of different scales can be enhanced by fusing different levels of information. The feature fusion module fuses the features of different depths into the final segmentation features. Finally, a learning pixel aggregation method based on clustering idea is proposed to aggregate the text instance pixels to the correct core to reconstruct the complete text instance and suppress the adhesion and overlap of adjacent text. The detection efficiency and the detection performance of long text and dense text are greatly improved by this model.

Liao et al. [12] proposed a Differentiable Binarization (DB) model, which is inserted into the partition network to joint optimization using the idea of border learning. Whether ResNet-50 or lightweight ResNet-18, excellent detection performance is obtained by the backbone network of the model, and result of real-time reasoning is achieved in the case

of ResNet-18. However, the result of this model is mistaken when a text in the center of another text instance. It is a common limitation of segmentation-based scene text detectors.

2.3. **Detection method based on regression and segmentation.** The operations of multi-information segmentation and fusion are usually needed in detection method based on segmentation. The kind of detection method is flexible and more suitable for irregular text detection. The method based on regression is not robust enough to detect multiple text sizes, but it can detect small text. Therefore, the advantages of the two methods can be combined to detect irregular text. Liu et al. [13] proposed Pyramid Mask Text Detector (PMTD) model in which the shape and position information is encoded into the monitor instead of binary classification at pixel level and a soft text mask is predicted for each text instance in the model. In PMTD model, 2D soft mask is reconstructed into 3D space and the optimal pyramid is regressed from these 3D points by planar clustering algorithm. However, quadrilateral detection box is still used to represent the text instance area in this model and it is not effective in detecting irregular text.

According to the problem of detecting curvilinear text, Xie et al. [14] suppressed false positives effectively by context semantic information and rerating mechanism for all predicted text instances. For similar problems, Wang et al. [15] proposed a ContourNet model that is a scale insensitive adaptive region proposal network.

Because the scene text detection and recognition technology will be applied in the actual scene, the unconstrained scene usually has the problems of strong illumination, blur, occlusion and so on. Many models based on deep learning may have been fitted on the training set and test set, so some antagonistic samples should be added to the training model to make the model more robust.

3. **Natural Scene Text Recognition Method Based on Deep Learning.** Scene character recognition technology has been attracting more and more attention in the field of computer vision and machine intelligence. Text recognition scene technology is important for scene understanding. It is still the most challenging problem for text recognition in an unconstrained environment. There are many text recognition models proposed based on deep learning by researchers.

3.1. **Recognition method based on CTC.** Sequential text alignment in text recognition technology can be solved by Connectionist Temporal Classification (CTC) [16] mechanism. He et al. [17] proposed a Deep-Text Recurrent Network (DTRN) recognition model in which CNN and RNN are put in a network for end-to-end joint training. The scene text recognition problem is regarded as a sequence marking problem. The sequence output of Long Short-Term Memory (LSTM) [18] is mapped into its target string and the result is adjusted by CTC.

Shi et al. [19] proposed Convolution Recurrent Neural Network (CRNN) that can stack multiple Bidirectional Long Short-Term Memory (Bi-LSTM) [20] in deep Bi-LSTM and end-to-end character recognition is realized after connected the end of bidirectional LSTM network with CTC model. CRNN model consists of convolution layer, circulation layer and transcription layer. Here the convolution layer is based on VGG-VeryDeep, the recurrent layer is composed of a deep Bi-LSTM and is designed on top of the convolutional layer to predict the label distribution of the feature sequence, and each frame prediction of RNN is converted into tag sequence in the transcription layer. CRNN network can be used in dictionary free or dictionary-based tasks, its scale is smaller, it does not involve level normalization or length restriction, and the recognition performance is improved greatly. However, because the CTC loss function is quoted and the character position is uncertain, as well as the correlation between features is not considered by the whole network, the features learned by CNN are unsatisfactory.

3.2. **Recognition method based on attention.** Scene text recognition models based on the encoder-decoder framework is mainstream currently, while the input sequence can only be encoded as a fixed length vector in the traditional encoder decoder framework. Outputs of the encoder based on attention mechanism are the sequence with uncertain length vectors and the target data and related data are given more larger weight. Vector representation of longer input sequences can be learned more reseanable.

It is still a challenging task to recognize irregular text. Shi et al. [21] proposed Robust text recognizer with Automatic Rectification (RARE), in which the input image with irregular characters is corrected by using Thin-Plate Splines (TPS) [22] transform in Spatial Transformer Network (STN) [23], the corrected characters are arranged along the horizontal line, which is more suitable for SRN recognition. The structure of RARE is shown in Figure 2. In Sequence Recognition Network (SRN), the corrected result is used as input, and the recognition problem is modeled as a sequence recognition problem based on attention mechanism, so that the whole model can be identified end-to-end. The recognition performance of irregular text in natural scenes is greatly improved through the application of this model. However, the nonlinear function tanh is used as the activation function of the final full connection layer in the rare model, which ensures that the sampling point is within the picture, but slows down the convergence speed. The picture of the control point obtained by the positioning network is slightly larger, which leads to the need for more parameters in the prediction. For this reason, Shi et al. [24] improved the RARE model and proposed Attentional Scene Text Recognizer with Flexible Rectification (ASTER), whose structure is shown in Figure 3. The model uses Bi-LSTM and attention to do end-to-end joint training. Different from RARE, tanh is not used as the activation function in the last fully connected layer, and the values in the sampler are trimmed, which not only reduces the retention of gradient in the process of back propagation, but also ensures effective sampling. At the same time, different sizes of images are used in STN for positioning network and sampling network. The positioning network will obtain control points from smaller images, which reduces the parameters needed for prediction.

FIGURE 2. Overview of the RARE model

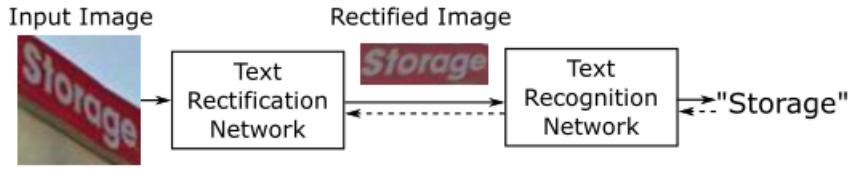

FIGURE 3. Overview of the ASTER model

The recognition network based on attention mechanism usually depends on the incoming decoding information. Due to the coupling relationship between attention mechanism and decoding information, the error of decoding information is bound to accumulate and spread, so the traditional attention mechanism always has serious alignment problems. Wang et al. [25] effectively solved the problem of error accumulation and propagation in decoding information through the idea of "decoupling". Cheng et al. [26] found that when

processing low pixel/complex images, the method based on attention mechanism does not perform well. It detects and corrects the center of attention through focusing network, which effectively solves the "attention shift problem".

TPS transformation is used to find some control points [23,24] and the text position is determined after the control points are regressed. Some scholars have solved this problem from a character-level perspective, such that the character-level rotation is limited and relatively easy to recognize, no matter how the line is distorted. If we can design a good detector and classifier at the character level, we can solve the problem of irregular text recognition very well.

3.3. **Recognition method based on CTC and attention.** The probability of alignment between predicted text sequence and actual text sequence generated by attention mechanism may increase the alignment error, while attention mechanism can be guided to get better alignment by Connectionist Temporal Classification (CTC) [27] mechanism. Therefore, the combination of the two mechanisms can promote each other and improve the recognition performance. Litman et al. [28] proposed Selective Context Attentional Text Recognizer (SCATTER) model by fusing the CTC and attention mechanism. The input image is corrected by TPS transform firstly. Then the visual features are extracted in CNN layer. Multiple Bi-LSTM encoders are stacked. And a cascaded selective attention decoder is designed by CTC aided training. A two-step 1D attention mechanism is used to decode the visual features of CNN layer and the context features of Bi-LSTM layer. Deeper Bi-LSTM is successfully trained in SCATTER by the idea of stacking, which has improved for the recognition performance greatly. The whole network is more stable and robust by the intermediate supervision and selective decoder. Similarly, Hu et al. [29] also used the hybrid idea and attention is used to supervise CTC for better recognition. In this model, attntion and CTC are combined to supervise and guide the alignment of CTC. Meanwhile, graph convolution network is added to the CTC branch to improve the model expression ability.

4. **Common Datasets.** At present, it is the more commonly data set for scene text detection and recognition, such as horizontal datasets ICDAR2013 [30], ICDAR2015 [31], SVT [32], and KAIST [33], for arbitrary shape, such as MSRA-TD500 [34], COCO-Text [35], and SCUT-CTW1500 [36], for multilingual data, such as ICDAR2017-MLT [37], and ICDAR2019-MLT [38], and for Chinese data, such as ICDAR2019-ReCTS [39], and CTW [40].

The datasets published by ICDAR in recent years have a multi-language version such as 2017-MLT and ICDAR2019-MLT. This also indicates that the status of multilingual text detection and recognition is increasing, and the demand for multilingual text detection and recognition technology is also increasing. Half of the widely used benchmark datasets have incomplete annotations, such as ignoring case sensitivity and punctuation, which should provide new, more comprehensive annotations for these datasets.

5. **Conclusions.** Scene text detection technology and character recognition technology are analyzed in this paper. In text detection technology, text detection technique can be more flexible for detecting an irregular text segmentation based method. However, the method based on regression can detect small text, but it cannot adapt to the multi-scale of text. Performance can be improved by combining the two mechanisms in a hybrid approach. In the text recognition method, the alignment problem between the text sequence and the actual text sequence is solved effectively by CTC mechanism. Due to the uncertainty of character position, and because the relationship between features is not considered by CTC, the recognition of long text may be mistaken. The attention mechanism enables the decoder to pay more attention to the information related to the target text, but the attention mechanism generates alignment probability, which may lead

to alignment error. The combination of CTC and attention can supervise each other to improve the recognition performance.

Acknowledgment. This work is partially supported by Heilongjiang Provincial Department of Education (grant no. 135309466) and Qiqihar University (grant no. YJSCX20210 79).

REFERENCES

[1] M. A. Radwan, M. I. Khalil and H. M. Abbas, Neural networks pipeline for offline machine printed Arabic OCR, *Neural Processing Letters*, vol.48, no.2, pp.769-787, 2018.

[2] R.-M. Wang, N. M. Sang, D. Ding, J. Chen et al., Text detection in natural scene image: A survey, *Acta Autom. Sin.*, vol.44, no.12, pp.2113-2141, 2018.

[3] M. H. Nguyen, A label-oriented approach for text classification, *International Journal of Innovative Computing, Information and Control*, vol.16, no.5, pp.1593-1609, 2020.

[4] M. Liao, B. Shi, X. Bai, X. Wang and W. Liu, Textboxes: A fast text detector with a single deep neural network, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.31, no.1, 2017.

[5] M. Liao, B. Shi and X. Bai, TextBoxes++: A single-shot oriented scene text detector, *IEEE Trans. Image Processing*, vol.27, no.8, pp.3676-3690, 2018.

[6] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He and J. Liang, East: An efficient and accurate scene text detector, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5551-5560, 2017.

[7] B. Shi, X. Bai and S. Belongie, Detecting oriented text in natural images by linking segments, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2550-2558, 2017.

[8] J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu and X. Bai, Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping, *Pattern Recognition*, vol.96, DOI: 10.1016/j.patcog.2019.06.020, 2019.

[9] S. Long, J. Ruan, W. Zhang, X. He, W. Wu and C. Yao, TextSnake: A flexible representation for detecting text of arbitrary shapes, *Proc. of the European Conference on Computer Vision (ECCV)*, pp.20-36, 2018.

[10] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu and S. Shao, Shape robust text detection with progressive scale expansion network, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9336-9345, 2019.

[11] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu and C. Shen, Efficient and accurate arbitrary-shaped text detection with pixel aggregation network, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.8440-8449, 2019.

[12] M. Liao, Z. Wan, C. Yao, K. Chen and X. Bai, Real-time scene text detection with differentiable binarization, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.34, no.7, pp.11474-11481, 2020.

[13] J. Liu, X. Liu, J. Sheng, D. Liang, X. Li and Q. Liu, Pyramid mask text detector, *arXiv Preprint*, arXiv: 1903.11800, 2019.

[14] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao and G. Li, Scene text detection with supervised pyramid context network, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.33, no.1, pp.9038-9045, 2019.

[15] Y. Wang, H. Xie, Z. J. Zha, M. Xing, Z. Fu and Y. Zhang, ContourNet: Taking a further step toward accurate arbitrary-shaped scene text detection, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11753-11762, 2020.

[16] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, *Proc. of the 23rd International Conference on Machine Learning*, pp.369-376, 2006.

[17] P. He, W. Huang, Y. Qiao, C. Loy and X. Tang, Reading scene text in deep convolutional sequences, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.30, no.1, 2016.

[18] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation*, vol.9, no.8, pp.1735-1780, 1997.

[19] 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 Trans. Pattern Analysis and Machine Intelligence*, vol.39, no.11, pp.2298-2304, 2016.

[20] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.31, no.5, pp.855-868, 2008.

[21] B. Shi, X. Wang, P. Lyu, C. Yao and X. Bai, Robust scene text recognition with automatic rectification, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4168-4176, 2016.

[22] F. L. Bookstein, Principal warps: Thin-plate splines and the decomposition of deformations, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.11, no.6, pp.567-585, 1989.

[23] K. Choi, G. Fazekas, M. Sandler and K. Cho, Convolutional recurrent neural networks for music classification, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.2392-2396, 2007.

[24] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao and X. Bai, ASTER: An attentional scene text recognizer with flexible rectification, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.41, no.9, pp.2035-2048, 2018.

[25] T. Wang, Y. Zhu, L. Jin, C. Luo, X. Chen, Y. Wu, Q. Y. Wang and M. Cai, Decoupled attention network for text recognition, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.34, no.7, pp.12216-12224, 2020.

[26] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu and S. Zhou, Focusing attention: Towards accurate text recognition in natural images, *Proc. of the IEEE International Conference on Computer Vision*, pp.5076-5084, 2017.

[27] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, *Proc. of the 23rd International Conference on Machine Learning*, pp.369-376, 2006.

[28] R. Litman, O. Anschel, S. Tsiper, R. Litman, S. Mazor and R. Manmatha, SCATTER: Selective context attentional scene text recognizer, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11962-11972, 2020.

[29] W. Hu, X. Cai, J. Hou, S. Yi and Z. Lin, GTC: Guided training of CTC towards efficient and accurate scene text recognition, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.34, no.7, pp.11005-11012, 2020.

[30] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn and L. P. De Las Heras, ICDAR 2013 robust reading competition, *2013 12th International Conference on Document Analysis and Recognition*, pp.1484-1493, 2013.

[31] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, L. Matas, L. Neunmann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida and E. Valveny, ICDAR 2015 competition on robust reading, *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp.1156-1160, 2015.

[32] K. Wang, B. Babenko and S. Belongie, End-to-end scene text recognition, *2011 International Conference on Computer Vision*, pp.1457-1464, 2011.

[33] S. Lee, M. S. Cho, K. Jung and J. H. Kim, Scene text extraction with edge constraint and text collinearity, *2010 20th International Conference on Pattern Recognition*, pp.3983-3986, 2010.

[34] C. Yao, X. Bai, W. Liu, Y. Ma and Z. Tu, Detecting texts of arbitrary orientations in natural images, *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1083-1090, 2012.

[35] A. Veit, T. Matera, L. Neumann, J. Matas and S. Belongie, COCO-Text: Dataset and benchmark for text detection and recognition in natural images, *arXiv Preprint*, arXiv: 1601.07140, 2016.

[36] Y. Liu, L. Jin, S. Zhang and S. Zhang, Detecting curve text in the wild: New dataset and new solution, *arXiv Preprint*, arXiv: 1712.02170, 2017.

[37] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, W. Khlif, M. M. Luqman, J. Burie, C. Liu and J. M. Ogier, ICDAR 2017 robust reading challenge on multi-lingual scene text detection and script identification – RRC-MLT, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol.1, pp.1454-1459, 2017.

[38] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J. Burie, C. Liu and J. M. Ogier, ICDAR 2019 robust reading challenge on multi-lingual scene text detection and recognition – RRC-MLT-2019, *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp.1582-1587, 2019.

[39] R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang, X. Bai, B. Shi, D. Karatzas, S. Liu and C. V. Jawahar, ICDAR 2019 robust reading challenge on reading Chinese text on signboard, *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp.1577-1581, 2019.

[40] T. L. Yuan, Z. Zhu, K. Xu, C. J. Li, T. J. Mu and S. M. Hu, A large Chinese text dataset in the wild, *Journal of Computer Science and Technology*, vol.34, no.3, pp.509-521, 2019.