

Research Article

Rethinking Multilingual Scene Text Spotting: A Novel Benchmark and a Character-Level Feature Based Approach

Siliang Ma¹ , Yong Xu^{1, 2, *}

¹School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

²Pengcheng Laboratory, Shenzhen, China

Abstract

End-to-end multilingual scene text spotting aims to integrate scene text detection and recognition into a unified framework. Actually, the accuracy of text recognition largely depends on the accuracy of text detection. Due to the lackage of benchmarks with adequate and high-quality character-level annotations for multilingual scene text spotting, most of the existing methods train on the benchmarks only with word-level annotations. However, the performance of multilingual scene text spotting are not that satisfied training on the existing benchmarks, especially for those images with special layout or words out of vocabulary. In this paper, we proposed a simple YOLO-like baseline named CMSTR for character-level multilingual scene text spotting simultaneously and efficiently. Technically, for each text instance, we represent the character sequence as ordered points and model them with learnable explicit point queries. After passing a single decoder, the point queries have encoded requisite text semantics and locations, thus can be further decoded to the center line, boundary, script, and confidence of text via very simple prediction heads in parallel. Furthermore, we show the surprisingly good extensibility of our method, in terms of character class, language type, and task. On the one hand, DeepSolo not only performs well in English scenes but also masters the Chinese transcription with complex font structure and a thousand-level character classes. On the other hand, based on the extensibility of DeepSolo, we launch DeepSolo++ for multilingual text spotting, making a further step to let Transformer decoder with explicit points solo for multilingual text detection, recognition, and script identification all at once.

Keywords

Multilingual Scene Text Image, Scene Text Recognition, Character-Level Annotations, Synthetic Benchmark

1. Introduction

Multilingual scene text spotting still remains great challenges: the performance of layout analysis, the recognition accuracy of words out of vocabulary still far behind the human recognition capability. Unlike the optical character images, scene text image contains rich semantic information and other interference factors other than text as shown in [Figure 1](#) (such as hybrid layout, complex background, noise, color

degradation, etc.), which makes it more difficult for computer to detect and recognize the character in scene text image. To design and evaluate scene text spotting algorithms and systems, the availability of large-scale dataset with character-level annotations is necessary. At present, most of the studies related to scene text spotting are dominantly trained on benchmarks only with word-level annotations, which are not

*Corresponding author: yxu@scut.edu.cn (Yong Xu)

Received: 30 July 2024; **Accepted:** 26 Aug 2024; **Published:** 6 September 2024



Copyright: © The Author(s), 2024. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

accurate enough for scene text spotting. In addition, further development and assessment of scene text recognition model are restricted by existing benchmarks with several issues:

Small-scale. Deep learning has been widely used in scene text spotting during the past few years, and demonstrated state-of-the-art performances. However, most of the existing scene text recognition methods need to train on large scale of scene text images with labels and position annotations. As we all know, it still takes a long time for researchers to capture and preprocess large-scale training data from real-world, which means most of the benchmarks for multilingual scene text spotting are with small scale.

Lack of accurate character-level annotations. For multilingual scene text spotting, character-level annotations with high precision are of significance for several reasons. (i) They ensure the accuracy and reliability of the prediction results of scene text spotting algorithm; (ii) they can adapt to more

flexible scene text images with words out of vocabulary or special layout; (iii) they help us establish the association between each specific character and the recognition result, which is conducive to error tracking of prediction results and model optimization. However, the accuracy of our proposed MLText is higher than most of the existing scene text spotting datasets represented by SynthText [6] and ICDAR 2019 ReCTS [34] have character-level annotations, especially for those images with text in irregular layout, which impact a lot to the performance of character-level scene text spotting.

Insufficient data integrity. Actually, due to the difference of usage frequency between different characters, it is not easy for us to capture all of the character from real world. Most of the publicly available multilingual scene text spotting benchmarks only have limited words and font styles due to the huge workload of manually acquiring large-scale scene text images, which can not satisfy the requirements of Industrial-level scene text spotting.



Figure 1. Multilingual scene text images from SCUT_FORU_DB [35], IIIT5K [19], TotalText [4] and SCUT-CTW1500 [16] with various layout.

With the above motivations, we provide the community a novel benchmark called MLText and a character-level feature based method called CSTDRNet for scene text spotting (STDR) with multi-fold contributions:

1) We propose a novel benchmark for scene text spotting with the following properties:

(i) **Accuracy** All of the text images from MLText are generated automatically based on the labels and related attributes without any manual operations, which ensures the accuracy of the labels. The coordinates of each character are calculated based on minimum area rectangle, which guarantees the accuracy of annotations.

(ii) **Efficiency** We proposed a scene text image batch generation method based on text labels and other related attributes, which can generate a large scale of benchmark for scene text spotting including text labels, character-level and word-level annotations in a short time. According to our experiment, it takes about 0.05s to generate one image.

(iii) **Authenticity** Although MLText is a synthetic dataset, we choose the background from real-world to enhance the authenticity of our proposed dataset. We combine the synthetic text and real scenario to restore the scene text images captured from real-world.

(iv) **Integrity** MLText makes full use of the lexicon from MJSynth [12], SynthText [6], MTHv2 [23] and

SCUT_FORU_DB [35], which consists of about 150k word sequences within more than 71,000 images for English and 17,000 for Chinese. It is well known that the training time will be longer when the scale of training dataset is larger. Therefore, we provide as more style of characters within less images than existing benchmarks. By releasing MLText, we aim to offer a dedicated benchmark for the development and assessment of character-level scene text spotting.

2) We propose a multilingual scene text spotting method based on character-level features, which provides scene text spotting results with clearer directivity. In order to improve the accuracy of scene text spotting, we design a dynamic input scaling module to replace the fixed input scaling module. Adequate experiments are conducted on different test dataset to demonstrate the robustness of our proposed CSTDRNet.

2. Related Work

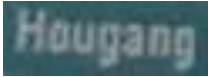
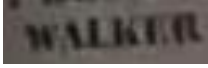
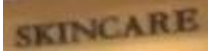
2.1. Multilingual Scene Text Spotting Benchmarks

Scene text spotting benchmarks with character-level annotations provide accurate coordinates for each individual

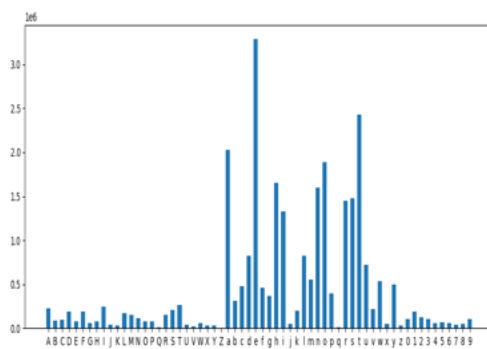
character, which is beneficial for increasing the accuracy and interpretation. However, the bounding boxes are usually manually annotated based on LabelImg or other tools, which may lead to some mistakes and bring heavy workload. Currently, the popular scene text benchmarks including SynthText [6], ICDAR 2013 [14], ICDAR 2015 [13], ICDAR 2017 MLT [20], ICDAR 2017 RCTW [29], ICDAR 2019 MLT [21], ICDAR 2019 ReCTS [34], Total-Text [4], IIIT5K-Words [19], SCUT-CTW1500 [16], SVT [31], SVHN [22], CUTE80 [28], MSRA-TD500 [5] and SCUT_FORU_DB [35]. The properties comparison of benchmarks mentioned above are shown as Table 2. As we can see, only SynthText, ICDAR 2019 ReCTS, IIIT5K-Words, SVHN and SCUT_FORU_DB have character-level annotations among the benchmarks mentioned above, while SVHN only contains digits and MSRA-TD500 only have word-level coordinate annotations without labels. In addition, most of the benchmarks mentioned above are preprocessed by human, which means it may bring extra errors from manual operation as shown in Table 1.

Table 1. Part of the incorrect labels from ICDAR 2015 test set.

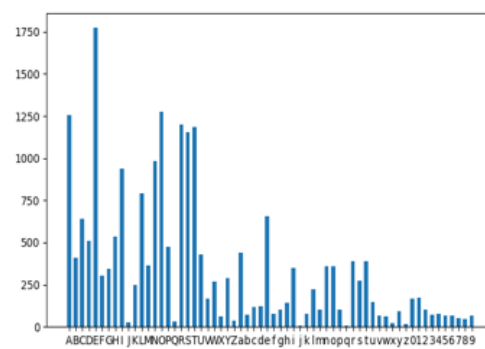
	Origin label	Correct label
	VEICHLES	VEHICLES

	Origin label	Correct label
	Haugang	Hougang
	SINCLARE	SKINCARE
	WALKIN	WALKER

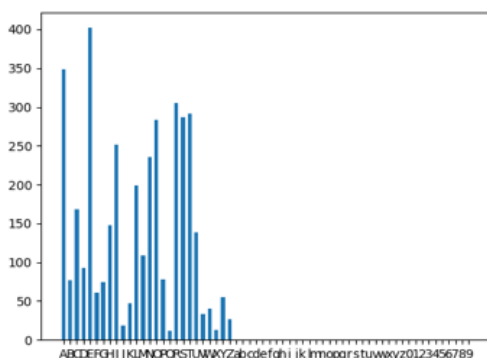
Moreover, most of the above datasets are obtained from the real world, considering the difference in the frequency of each character in the real scenario, which means that it is difficult for them to cover all character categories in a specific language. Due to the large number of Chinese character categories which is not convenient for visual display, we selected a total of 62-character categories including uppercase and lowercase English letters and digits from some existing scene text benchmarks and compared them with MLText generated by ourselves, the specific results are shown in Figure 2. As we can see, the character frequency difference of the existing public scene text dataset is very large, and there is a serious problem of category imbalance. Some commonly used characters even do not appear in existing benchmarks, which has a great impact on the performance of scene text spotting models.



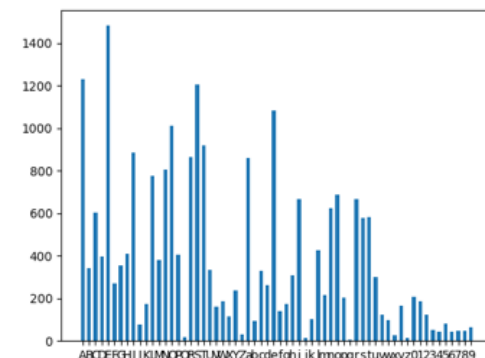
(a) SynthText



(b) SCUT_FORU_DU



(c) SVT



(d) IC15

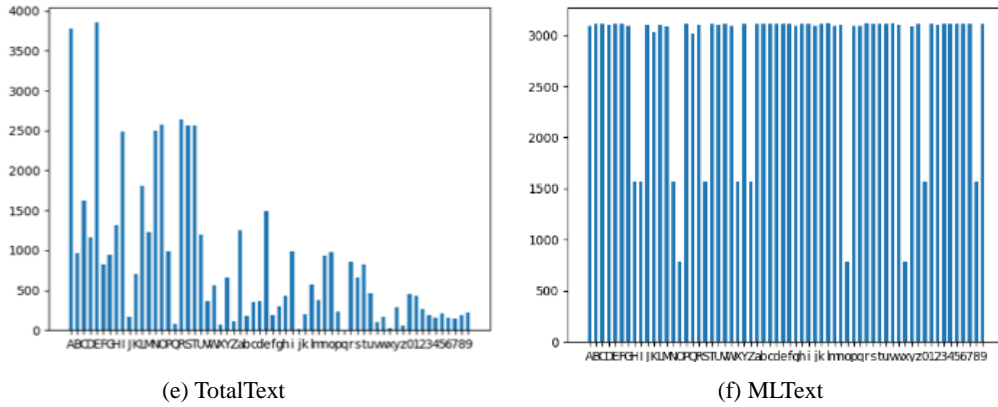


Figure 2. The character frequency of the existing public scene text image dataset is compared with MLText.

2.2. Multilingual Scene Text Spotting

Most of the existing researches focus on training one model using datasets from a specific language. However, they do not unveil the capability of identifying script and recognizing text instances varying from multiple languages in a single unified network. Only a few works investigated end-to-end multilingual text spotters and achieved exciting results. Among them, E2E-MLT [3] proposes a framework for multilingual text spotting by handling characters from all languages in the same

recognizer, the same as CRAFTS [1]. In comparison, Multiplexed TextSpotter [8] and Grouped TextSpotter [9] leverage several adapted recognizers to deal with different languages respectively, with Language Prediction Network (LPN) routing the detected text to an appropriate recognizer. However, all of the above methods rely on RoI and sequence-to-sequence recognizers. Besides, in [8, 9] hand-crafted LPNs are required for script identification and multi-head routing. From a different paradigm, we demonstrate that our simple baseline, i.e., DeepSolo++ [33] with a simple routing scheme for multilingual recognition.

Table 2. Comparison of MLText with the most popular scene text spotting benchmarks.

Benchmark	Categories	Text Type	Train images	Test images	Word-level annotations	Character-level annotations
SynthText [6]	69	Synthetic	858,768	×	7,266,866	46,372,500
ICDAR 2013 [14]	78	Real	229	233	1,944	×
ICDAR 2015 [13]	86	Real	500	1,000	17,116	×
ICDAR 2017 MLT [20]	3,885	Real	7,200	1,800	107,547	×
ICDAR 2017 RCTW [29]	3,499	Real	11,154	1,000	65,248	×
ICDAR 2019 MLT [21]	4,205	Real	9,000	1,000	111,998	×
ICDAR 2019 ReCTS [34]	4,137	Real	20,000	5,000	146,952	440,027
TotalText [4]	71	Real	1,255	300	13,708	×
IIIT5K [19]	62	Real	2,000	3,000	5,000	18,268
SCUT-CTW1500 [16]	×	Real	1,000	500	10,760	×
SVT [31]	26	Real	100	250	904	×
CUTE80 [28]	37	Real	288	×	288	×
SCUT_FORU_DB [35]	72 (ENG)	Real	1,162 (ENG)	×	7,136 (ENG)	22,555 (ENG)
	2,116 (CHN)		1,861 (CHN)	355 (CHN)	4,338 (CHN)	21,393 (CHN)
MLText (Ours)	6,810	Hybrid	71,868 (ENG)	3,000 (ENG)	143,736 (ENG)	320,060 (ENG)
			17,655 (CHN)	4,666 (CHN)	×	365,692 (CHN)

3. Methods

3.1. Image Collection

The images in MLText are composed of images from public available benchmarks and images generated by ourselves. All of these images have character-level as well as word-level position annotations, which also provides accurate text labels. Figure 3 shows the generation process of our proposed MLText. Existing datasets for scene text spotting tend to focus on images captured from real-world, while MLText balances the two text types to achieve a more real-world and diverse dataset with various font style and scenarios. In addition, rather than focusing on text lines, the proposed MLText includes a large amount of stylish text. Sharing the language setting from those representative text segmentation datasets, the proposed MLText mainly focuses on English (i.e., case sensitive alphabet, digits, and special tokens) and Chinese.

In order to better meet the requirements of scene text spotting, and make up for the problems of insufficient data integrity, low image resolution and heavy workload of manual annotation in the existing public datasets. We proposed a batch generation method of arbitrary character images based on text labels. This method can generate arbitrary character image data according to the preset font style, font size, font color, labels and other related attributes, which greatly reduces the workload of manual annotation in scene text image.

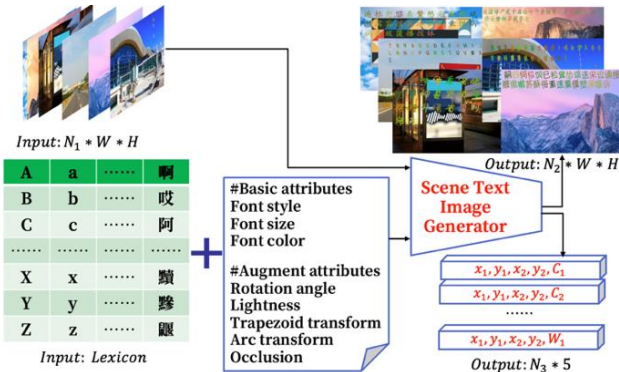


Figure 3. Generation process of our proposed scene text spotting benchmark MLText.

We have fully considered the relevant attributes involved in the existing scene text image datasets, and found that most of the scene text images can be composed of eight types of attribute parameter changes, such as font style, font size, character-level rotation, word-level rotation, trapezoidal change, arc change, light effect and occlusion, etc., which provides the inspiration of the design of our dataset generation method. We proposed a scene text image data generation method based on character-level features, and design eight

functions to set the values of eight types of properties by adjusting parameters. The values range of each parameter are shown as Table 3.

Table 3. Values range of each parameter of our proposed MLText.

Property type	Values
Font style	eg. Arial.ttf, simkai.ttf
Font size	$\{x x \in N^+\}$
Word-level rotation	$\{angle -\pi \leq angle \leq \pi\}$
Character-level rotation	$\{angle -\pi \leq angle \leq \pi\}$
Trapezoidal transform	$\{k 0 \leq k \leq 1\}$ $\{d d \in \{left, right, up, down\}\}$
Arc transform	$\{k 0 \leq k \leq 1\}$ $\{d d \in \{left, right, up, down\}\}$
Occlusion area	$\{x_1, y_1, x_2, y_2 0 \leq x_1 < x_2 \leq w, 0 \leq y_1 < y_2 \leq h\}$
Light central point	$\{x_1, y_1 0 \leq x_1 \leq w, 0 \leq y_1 \leq h\}$
Lightness	$\{x 0 \leq x \leq 100\}$

3.2. Character-Level Scene Text Spotting Network

After analyzing the previous studies about scene text spotting, we found that most of these methods perform well on images with horizontally arranged text but worse on vertically arranged text. Furthermore, these methods detect and recognize text based on word-level features, which make the prediction speed very slow and cannot establish the association between the recognition result and each individual character from the original image.

Inspired by the real-time object detection models [27, 25, 26, 2] proposed recent years, we proposed a Character-level Scene text spotting Network called CMSTR. The full framework of CMSTR is shown as Figure 4, which comprises a Dynamic Input Scaling module (DIS) and a VR model based on CSPDarkNet [30], SPP [7] and PAN [15]. Compared with fixed scaling method, dynamic input scaling method can keep the information from the original image as more as possible, which can help to increase the accuracy of scene text spotting.

3.3. Design Principle

Scene text spotting are different from natural object detection and recognition, which motivate specific designs for scene text spotting. As for natural object detection and recognition, we usually pay more attention to the colors, textures or other stable features. However, in scene text

spotting task, each individual character may have various colors, textures and font size, even though they are the same character, which means we need to focus on the structure of character itself. In order to retain the valid information in the original picture to the greatest extent, we replace the fixed resolution scaling module with a dynamic input scaling module.

The loss function of our proposed CSTDRNet consists of three parts, which can be written as follow:

$$\mathcal{L} = \mathcal{L}_{obj} + \mathcal{L}_{bbox} + \mathcal{L}_{cls} \quad (1)$$

$$\mathcal{L}_{obj} = \sigma BCELoss(x) \quad (2)$$

$$\mathcal{L}_{bbox} = 1 - \left(\frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \right) \quad (3)$$

$$\mathcal{L}_{cls} = \sigma BCELoss(x) \quad (4)$$

4. Experimental Evaluation

4.1. Experiment Settings

The experimental environment of scene text recognizers described in this paper is as follows: the memory is 32GB, the operating system is windows 11, the CPU is Intel i9-12900k, and the graphics card is NVIDIA Geforce RTX

3090 with 24GB memory. In order to ensure the consistency of experimental parameters, all of the evaluated recognizers in this paper use the same hyper parameters, such as learning rate, batch size and optimization function. Our model recognizes 9,139 types of characters for English, digits and Chinese, including “0-9”, “a-z”, “A-Z” and other widely-used Chinese characters. In order to reduce the category redundancy, we merge those letters with similar appearance in upper and lower cases into one category, such as ‘C’ and ‘c’.

During the training period, we use SGD [24] to optimize our training with the initial learning rate 0.01. The momentum is set to 0.937. The model is totally trained for about 100 epochs. We train our model with image batch size of 128. Data augmentation is also applied in the training period, including random mosaic, noisy and blur.

4.2. Experimental Results

4.2.1. Ablation Study

In order to verify the effectiveness of the dynamic input scaling module proposed in this chapter, we conducted ablation experiments on some mainstream scene text spotting datasets, and compared the test results before and after using dynamic scaling. The experimental results are shown in Table 4.

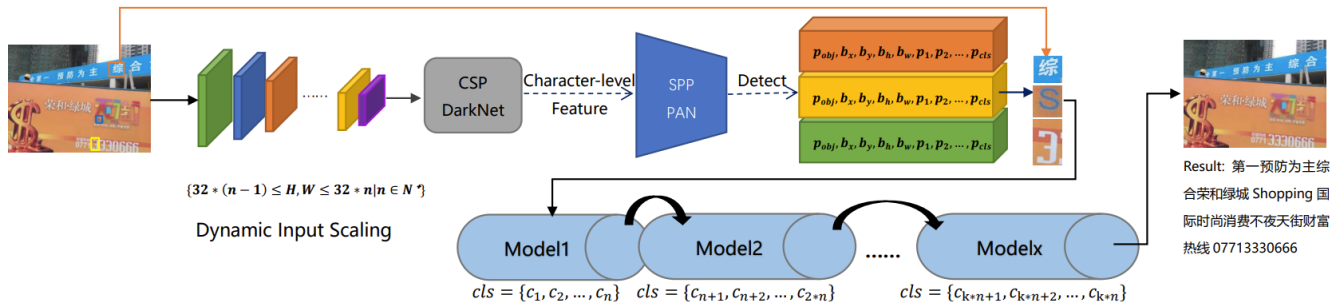


Figure 4. Architecture of our proposed character-level scene text spotting model CSTDRNet. In this framework, we generate three bounding boxes for each pixel, p_{obj} denotes the probability of object, b_x and b_y denote the centroid point's coordinate of each individual bounding box, b_h and b_w denote the height and width of each individual bounding box, p_{cls} denotes the probability of each category.

Table 4. Ablation experimental results of dynamic input scaling module.

#Training Set	#Test Set	w/DIS	P	R	AP50	AP
ICDAR 2019 ReCTS	SCUT_FORU_DB (ENG)	√	68.1	72.1	71.6	53.9
			68.4	71.9	71.9	53.2
	IIIT5K-Words	√	74.9	32.4	54.3	23.8
			76.7	37.9	57.6	25.4

#Training Set	#Test Set	w/DIS	P	R	AP50	AP
MLText	SCUT_FORU_DB (ENG)	√	82.3	72.8	78.8	57.1
			82.8	72.0	79.0	56.6
	IIIT5K-Words	√	79.5	29.5	55.1	25.4
			79.6	33	57.2	26.3

As we can see, the scene text spotting results have higher integrity with dynamic input scaling module, which means most of the characters in the image can be recognized correctly. As a result, dynamic input scaling module can select the most appropriate size for each individual image, which leads to less information redundancy or loss.

4.2.2. Character-Level Scene Text Spotting

In order to prove the capability of our proposed benchmark for multilingual scene text spotting, we conduct

character-level scene text spotting on two publicly available datasets: IIIT5K and SCUT_FORU_DB. The experimental results are shown as Table 5. As we can see, compared with training on SynthText and ICDAR 2019 ReCTS, our proposed CSTDRNet performs better on character-level scene text spotting training on our proposed benchmark MLText. In addition, our proposed benchmark and method focus on character-level features of scene text images, which increases the interpretability and robustness of multilingual scene text spotting.

Table 5. Comparison of test results on IIIT5K and SCUT_FORU_DB with different training datasets based on CSTDRNet.

#Training Set	#Test Set	P	R	AP50	AP
SynthText	IIIT5K-Words	68.2	46.6	48.5	15.1
	SCUT_FORU_DB (ENG)	58	70.1	71.4	33.8
	SCUT_FORU_DB (CHN)	-	-	-	-
ICDAR 2019 ReCTS	IIIT5K-Words	69.8	42.5	44.9	17.1
	SCUT_FORU_DB (ENG)	71.8	71.9	71.3	49.8
	SCUT_FORU_DB (CHN)	68.7	75.9	79.1	63.3
MLText (Ours)	IIIT5K-Words	75.3	70.1	64.6	58.5
	SCUT_FORU_DB (ENG)	82.8	72	79.1	56.4
	SCUT_FORU_DB (CHN)	70.9	76.8	81.4	66

4.2.3. Word-Level Scene Text Spotting

In order to better reveal the generalization ability of our proposed benchmark and method, we compared with some of the SOTA methods test on two popular multilingual scene text spotting datasets: TotalText and ICDAR 2019 ReCTS. The experimental results are shown as Table 6. As we can see, our proposed scene text spotting method CSTDRNet has higher accuracy and efficiency than the state-of-the-art

methods, which means our proposed benchmark and method have higher generalization ability. Although our proposed method cannot achieve the highest recall among the state-of-the-art methods, but we achieve highest precision with less redundant output, faster inference and smaller model size. Furthermore, our proposed CSTDRNet can achieve outstanding performance with only one time training on MLText, while most of existing methods need to fine-tune on each downstream test dataset to achieve better performance.

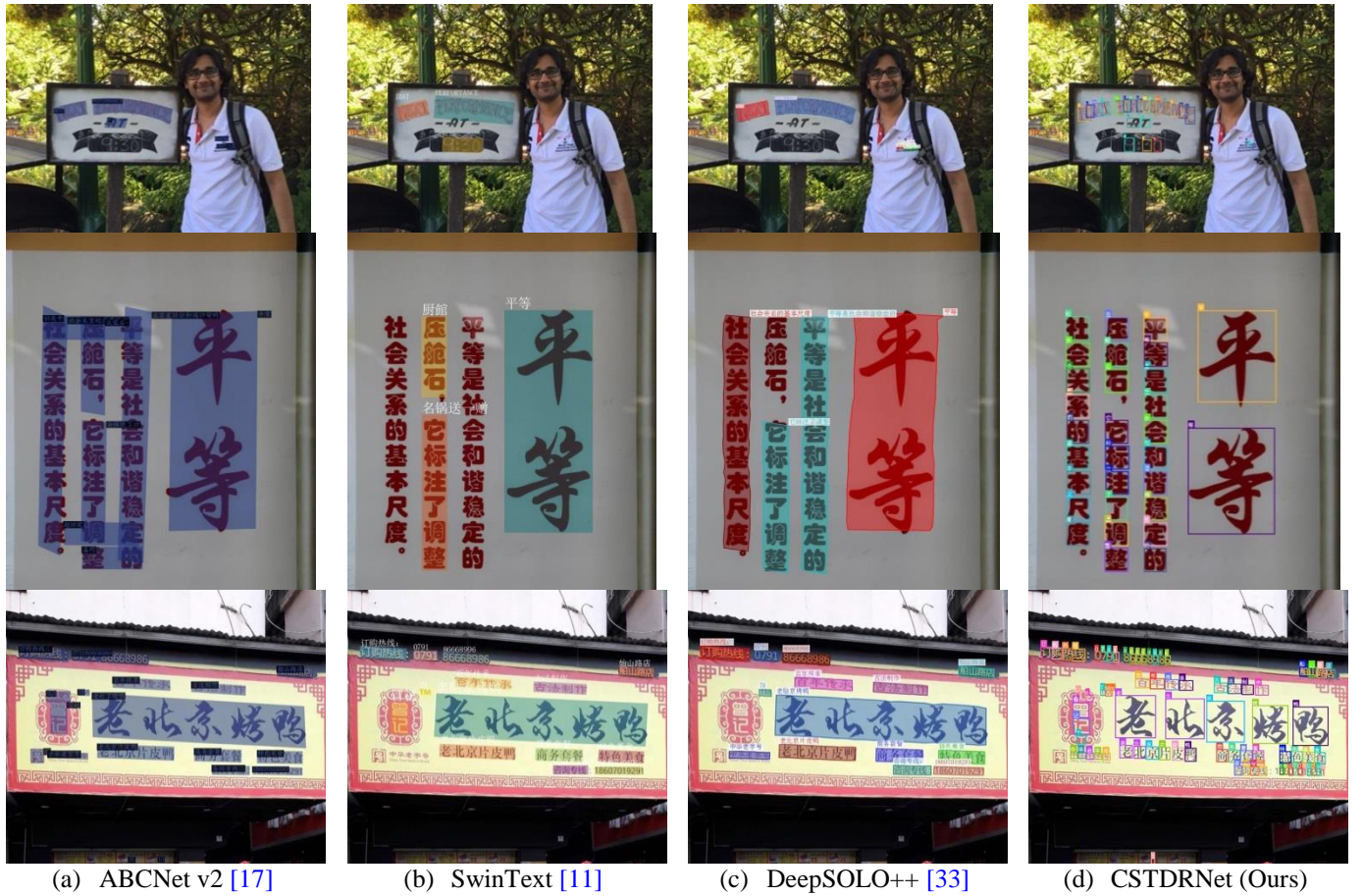


Figure 5. Scene text spotting results on images from TotalText, ICDAR 2019 ReCTS and SCUT_FORU_DB.

In order to better reveal the difference between our proposed method and the existing methods, the visualize results of scene text detection and recognition on some of the publicly available datasets are shown as Figure 5.

Table 6. Scene text spotting results on TotalText and ICDAR 2019 ReCTS. *P*, *R*, *H* represent precision, recall and hmean, respectively.

Methods	Datasets						FPS	Params (*10 ⁶)
	TotalText			ICDAR 2019 ReCTS				
	P	R	H	P	R	H		
SANHL_v1 [18]	-	-	-	91.98	93.86	92.91	1.15	500.11
AE TextSpotter [32]	-	-	-	93.38	89.98	91.65	0.89	352.97
ABCNetv2 [17]	70.2	82.1	75.7	92.89	87.91	90.33	1.58	50.84
SwinText [11]	-	-	88.0	94.1	87.1	90.4	0.13	176.77
ESTextSpotter [10]	92.0	88.1	90.0	91.3	94.1	92.7	4.3	49.84
DeepSolo++ [33]	92.9	87.4	90.0	92.55	89.01	90.74	1.11	64.4
CSTDRNet (Ours)	93.1	86.7	90.5	93.1	94.3	93.9	6.67	48.85

5. Conclusions

In this paper, we propose a novel benchmark for multilingual scene text spotting called MLText and a scene text spotting method based on character-level feature called CSTDRNet, which provides accurate character-level results based on dynamic input scaling module. The experimental results show that the dataset and method proposed by us can adapt to multilingual scene text spotting tasks in various scenarios, and provide a completely different method for scene text detection and recognition, with higher efficiency, accuracy and interpretability.

As for future work, we will try to consider how to integrate the advantages of word level recognition methods and character level recognition methods, fully integrating semantic information with the methods proposed by us to further improve model accuracy. As for those recognition results provided by our character-level scene text recognizer, we can use the word-level scene text recognizer to recheck the accuracy. To meet the requirements of scene text recognition in more scenarios, we will also try to design more complex glyph transforms.

Author Contributions

Siliang Ma: Conceptualization, Resources, Methodology, Writing

Yong Xu: Supervision

Funding

This work is supported in part by the National Natural Science Foundation of China (grant numbers 61602184, 61872151, 61672241, U1611461).

Data Availability Statement

The data is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Baek Y, Shin S, Baek J, et al. Character region attention for text spotting [C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. Springer International Publishing, 2020: 504–521. https://doi.org/10.1007/978-3-030-58526-6_30
- [2] Bochkovskiy A. Yolov4: Optimal speed and accuracy of object detection [J]. arxiv preprint arxiv:2004.10934, 2020.
- [3] Bušta M, Patel Y, Matas J. E2e-mlt-an unconstrained end-to-end method for multi-language scene text [C]// Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14. Springer International Publishing, 2019: 127–143. https://doi.org/10.1007/978-3-030-21074-8_11
- [4] Ch'ng C K, Chan C S, Liu C L. Total-text: toward orientation robustness in scene text detection [J]. International Journal on Document Analysis and Recognition (IJDAR), 2020, 23(1): 31–52. <https://doi.org/10.1007/s10032-019-00334-z>
- [5] Yao C, Bai X, Liu W, et al. Detecting texts of arbitrary orientations in natural images [C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 1083–1090. <https://doi.org/10.1109/CVPR.2012.6247787>
- [6] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2315–2324. <https://doi.org/10.1109/CVPR.2016.254>
- [7] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904–1916. <https://doi.org/10.1109/TPAMI.2015.2389824>
- [8] Huang J, Pang G, Kovvuri R, et al. A multiplexed network for end-to-end, multilingual OCR [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 4547–4557. <https://doi.org/10.1109/CVPR46437.2021.00452>
- [9] Huang J, Liang K J, Kovvuri R, et al. Task grouping for multilingual text recognition [C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 297–313. https://doi.org/10.1007/978-3-031-25069-9_20
- [10] Huang M, Zhang J, Peng D, et al. Estextspotter: Towards better scene text spotting with explicit synergy in transformer [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 19495–19505. <https://doi.org/10.1109/ICCV51070.2023.01786>
- [11] Huang M, Liu Y, Peng Z, et al. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition [C]//proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 4593–4603. <https://doi.org/10.1109/CVPR52688.2022.00455>
- [12] Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic data and artificial neural networks for natural scene text recognition [J]. arxiv preprint arxiv:1406.2227, 2014.
- [13] Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. ICDAR 2015 competition on robust reading [C]// 2015 13th international conference on document analysis and recognition (ICDAR). IEEE, 2015: 1156–1160. <https://doi.org/10.1109/ICDAR.2015.7333942>
- [14] Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 robust reading competition [C]// 2013 12th international conference on document analysis and recognition. IEEE, 2013: 1484–1493. <https://doi.org/10.1109/ICDAR.2013.221>

- [15] Li H, Xiong P, An J, et al. Pyramid attention network for semantic segmentation [J]. arxiv preprint arxiv:1805.10180, 2018.
- [16] Liu Y, Jin L, Zhang S, et al. Curved scene text detection via transverse and longitudinal sequence connection [J]. Pattern Recognition, 2019, 90: 337-345.
<https://doi.org/10.1016/j.patcog.2019.02.002>
- [17] Liu Y, Shen C, Jin L, et al. Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 8048-8064. <https://doi.org/10.1109/TPAMI.2021.3107437>
- [18] Liu Y, Zhang S, Jin L, et al. Omnidirectional scene text detection with sequential-free box discretization [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019: 3052-3058.
- [19] Mishra A, Alahari K, Jawahar C V. Scene text recognition using higher order language priors [C]//BMVC-British machine vision conference. BMVA, 2012.
<http://dx.doi.org/10.5244/C.26.127>
- [20] Nayef N, Yin F, Bizid I, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt [C]// 2017 14th IAPR international conference on document analysis and recognition (ICDAR). IEEE, 2017, 1: 1454-1459. <https://doi.org/10.1109/ICDAR.2017.237>
- [21] Nayef N, Patel Y, Busta M, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019 [C]//2019 International conference on document analysis and recognition (ICDAR). IEEE, 2019: 1582-1587. <https://doi.org/10.1109/ICDAR.2019.00254>
- [22] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning [C]// NIPS workshop on deep learning and unsupervised feature learning. 2011, 2011(2): 4.
- [23] Peng D, Jin L, Liu Y, et al. Pagenet: Towards end-to-end weakly supervised page-level handwritten Chinese text recognition [J]. International Journal of Computer Vision, 2022, 130(11): 2623-2645. <https://doi.org/10.1007/s11263-022-01654-0>
- [24] Qian Q, Jin R, Yi J, et al. Efficient distance metric learning by adaptive sampling and mini-batch stochastic gradient descent (SGD) [J]. Machine Learning, 2015, 99: 353-372.
<https://doi.org/10.1007/s10994-014-5456-x>
- [25] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
<https://doi.org/10.1109/CVPR.2017.690>
- [26] Redmon J. Yolov3: An incremental improvement [J]. arxiv preprint arxiv:1804.02767, 2018.
- [27] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
- [28] Risnumawan A, Shivakumara P, Chan C S, et al. A robust arbitrary text detection system for natural scene images [J]. Expert Systems with Applications, 2014, 41(18): 8027-8048.
<https://doi.org/10.1016/j.eswa.2014.07.008>
- [29] Shi B, Yao C, Liao M, et al. Icdar2017 competition on reading chinese text in the wild (rctw-17) [C]// 2017 14th iapr international conference on document analysis and recognition (ICDAR). IEEE, 2017, 1: 1429-1434.
<https://doi.org/10.1109/ICDAR.2017.233>
- [30] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.
<https://doi.org/10.1109/CVPRW50498.2020.00203>
- [31] Wang K, Babenko B, Belongie S. End-to-end scene text recognition [C]// 2011 International conference on computer vision. IEEE, 2011: 1457-1464. <https://doi.org/10.1109/ICCV.2011.6126402>
- [32] Wang W, Liu X, Ji X, et al. Ae textspotter: Learning visual and linguistic representation for ambiguous text spotting [C]//Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer International Publishing, 2020: 457-473.
https://doi.org/10.1007/978-3-030-58568-6_27
- [33] Ye M, Zhang J, Zhao S, et al. DeepSolo++: Let Transformer Decoder with Explicit Points Solo for Multilingual Text Spotting [J]. arxiv preprint arxiv:2305.19957, 2023.
- [34] Rui Zhang et al. "ICDAR 2019 Robust Reading Challenge on Reading Chinese Text on Signboard". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019.
- [35] Zhang S, Lin M, Chen T, et al. Character proposal network for robust text extraction [C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 2633-2637. <https://doi.org/10.1109/ICASSP.2016.7472154>

Biography



Siliang Ma is a Ph.D student at the School of Computer Science and Engineering at South China University of Technology. He received his B.E. degree from South China Agricultural University in 2019. He is the President of CCF South China University of Technology Student Chapter. His research interests include deep learning, pattern recognition, and text image processing.



Yong Xu received the B.S., M.S., and Ph.D. degrees in mathematics from Nanjing University, Nanjing, China, in 1993, 1996, and 1999, respectively. He was a Postdoctoral Research Fellow of computer science with the South China University of Technology, Guangzhou, China, from 1999 to 2001, where he became a Faculty Member and is currently a Professor with the School of Computer Science and Engineering. He is the vice president of South China University of Technology and the Dean of Guangdong Big Data Analysis and Processing Engineering & Technology Research Center. He is also a member of the Peng Cheng Laboratory. His current research interests include computer vision, pattern recognition, image processing, and big data. He is a senior member of the IEEE Computer Society and the ACM. He has received the New Century Excellent Talent Program of MOE Award.