CaMiTextSpotter: Character-aware Multi-label Script Identification for End-to-End, Multilingual Scene Text Spotting

Shuyang Feng
fsy_tongji@163.com

Abstract

End-to-end(E2E) scene text spotting has made significant progress in recent years. Existing methods focus mainly on recognizing Latin-alphabet languages and pay less attention to the problem of multilingual recognition.

There are several problems with E2E, multilingual scene text spotting: 1) The character set is too large thus the long-tail effect is obvious. 2) The multilingual mixed line-level samples can not be handled well. 3) In E2E manner, detection, script identification and recognition are not well coupled together. In this paper, we propose a new E2E, multilingual text spotting framework termedCharacter-Aware Multi-script Identification Text Spotter (CaMiTextSpotter). This framework contains a transformer-based detector, character-level multi-label script identification module, and multiple script-related recognition heads. Experiments show that our method can be a good solution to handle the above problems. Our method achieves comparable performance with previous state-of-the-art methods on MLT17 and MLT19 benchmarks.

1. Introduction

In recent years, scene-text related research has been an active research area. Understanding scene text is one of the most direct ways for computational systems to get knowledge in the real world. It usually involves two steps: detection and recognition. Detection is used to locate the text and recognition is used to obtain a transcription of the text. A more concise approach is to design detection and recognition as an end-to-end trainable framework that can bring more benefits in resource utilization, model performance, etc. However, existing end-to-end framework still faces many challenges. There are fewer researchers focus on its dilemmas in multilingual scenarios.

As shown in Figure 1.a, the most naive approach is to train a general OCR model using a relatively large character set that includes common characters in multilingual scenarios. According to the statistics for the ICDAR-MLT19 [24] dataset, [2] found that 23.83% of the characters in the character set appeared only once, which would be a key reason for the poor performance of the current model. Further, [13] introduces a script identification (SI) module. The computational process of the network is guided into a smaller character set for further computation, shown in Figure 1.b. Although it has been experimentally proven to be beneficial, this approach still does not solve the problem of multilingual character mixing, which means characters from multiple character sets are present in one sample. In addition, by further analysis of the MLT data, as detailed in experiment 4.3, we find problem of label ambiguity. First, we aggre-

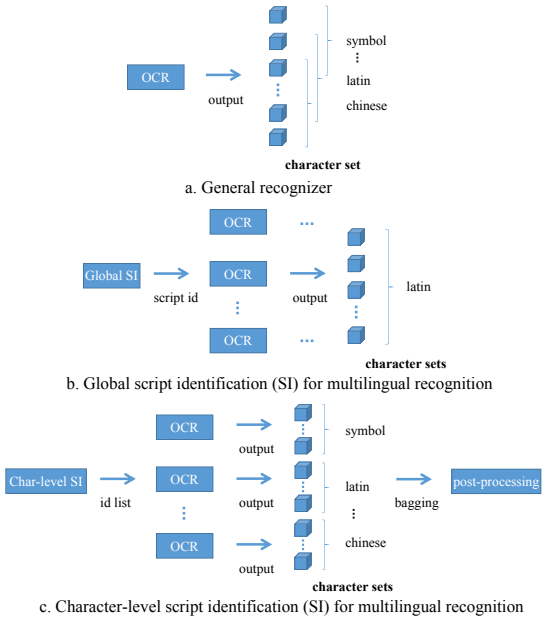

Figure 1. Several solutions for OCR systems for multilingual recognition tasks. Figure a represents that a generic recognizer is trained for all the common characters. Figure b shows that multiple language subsets are divided and a script identification network is introduced to process them separately. Figure c represents, a scheme for character level script identification. The ability of multiple recognition branches is multiplexed to solve samples with mixed languages.

gate the characters of each script to obtain multiple character sets. There is a large number of characters repeatedly contained within the character sets of multiple languages but only one script tag is provided. Besides, the main language is set during ICDAR-MLT19[1] collection, for example, when the main language is Japanese, common Chinese, Japanese, and Korean (CJK) characters will be marked as Japanese. It's a big problem for script identification and obviously introduces noise. For example, if '北野' is labeled as Japanese, obviously the script identification model can also treat it as Chinese. Finally, the strategy of end-to-end optimization of multilingual spotting system is also a problem worth exploring.

To solve the above problems, we propose CaMiTextSpotter, an end-to-end multilingual spotting framework based on SwinTextSpotter [14]. As shown in Figure 1.c, our framework introduces a script identification module, formalizing it as a character-level, multi-label classification task. On the one hand, character-level prediction enables the association of classification results with each character. By sharing the sequence attention with OCR heads, the output of each OCR head can be credibly integrated into the post-processing stage which can solve the problem of multilingual character mixing. On the other hand, multi-label prediction ensures that label ambiguity is reduced during the model training phase. Benefiting from the label generation approach, detailed in Methodology 3.2, the training data does not need to provide script labels. Finally, we propose a module that efficiently couples the three tasks of detection, identification, and OCR, using the finer-grained information provided in the script identification stage (including, character length, alignment attention information, and multi-label prediction) to better serve the end-to-end training.

To summarize, the main contribution of this paper is to propose a novel framework for end-to-end, multilingual text spotting, which unifies script identification and OCR formally and integrates them credibly to solve the problem of multilingual character mixing and reducing label ambiguity. At the same time, an effective end-to-end optimization module is proposed so that the experimental results can be comparable with previous state-of-the-art methods. Our contributions can be summarized as follows:

- We propose the first multilingual end-to-end OCR framework based on the transformer architecture. It formalizes the script identification as a character-level, multi-label prediction task. Also, we propose two constraint terms to help it converge better.
- Our proposed method can effectively solve the language mixed problem. By reusing the attention units of script identification and OCR branches, it makes the post-processing process more credible.
- The experimental results show that our method is able to achieve the Start-Of-The-Art comparable experimental results. Extensive experiments on ICDAR MLT17, MLT19 have demonstrate the effectiveness of our proposed method. Our proposed CaMiTextSpotter algorithm achieves very competitive performance against state-of-the-art methods.

[1] https://rrc.cvc.uab.es/?ch=15&com=downloads

2. Related work

2.1. Text detection and recognition methods

Scene text detection is the first step in scene text reading and it can be classified into two main categories: regression-based methods [12, 18, 28, 35, 42–44] and segmentation-based methods [1, 19, 21, 38, 39]. The regression-based approach represents the detection frame by regressing a number of contour points. The advantage is that it can rely less on post-processing and obtain detection results end-to-end. For example, methods such as EAST [43], TextBoxes++ [18], Wordsup [12], etc. require only one NMS calculation for their post-processing steps. However, the regression-based approach has difficulties in representing curved text, while the segmentation-based approach is a promising solution. TextSnake [21], PSENet [38], DB [19], etc. exploit the skeletal symmetry property of text and gradually solve the problem of curved text detection.

Text recognition is a essential step to achieve image-to-transcription, and there are two paradigms, CTC-based [29, 37] and attention-based [7, 22, 31] approaches. CTC-based methods have earlier solved the problem of end-to-end recognition of text images with variable length. However, CTC-based approaches are unable to learn language models, which is important for languages with dependencies in character order, such as Latin. Therefore, borrowing the idea from NMT [36], attention-based methods use an auto-regressive representation to achieve an implicit alignment of predictions and labels using attention mechanism.

2.2. Script identification

In early years, script identification was considered as a separate task and most methods formalized it as a sequence-to-label task [4, 6, 11, 17, 27, 30, 33, 41]. In order to adopt the characteristics of variable aspect ratio, existing methods [4, 6, 33, 41] usually design module on patch-level. One is to perform discriminative analysis on the patch features extracted by CNN and select language-related salient features to improve the identification results [33, 41]. The other use the attention mechanism, making the model adaptively adjust the response degree among the patches [4, 6].

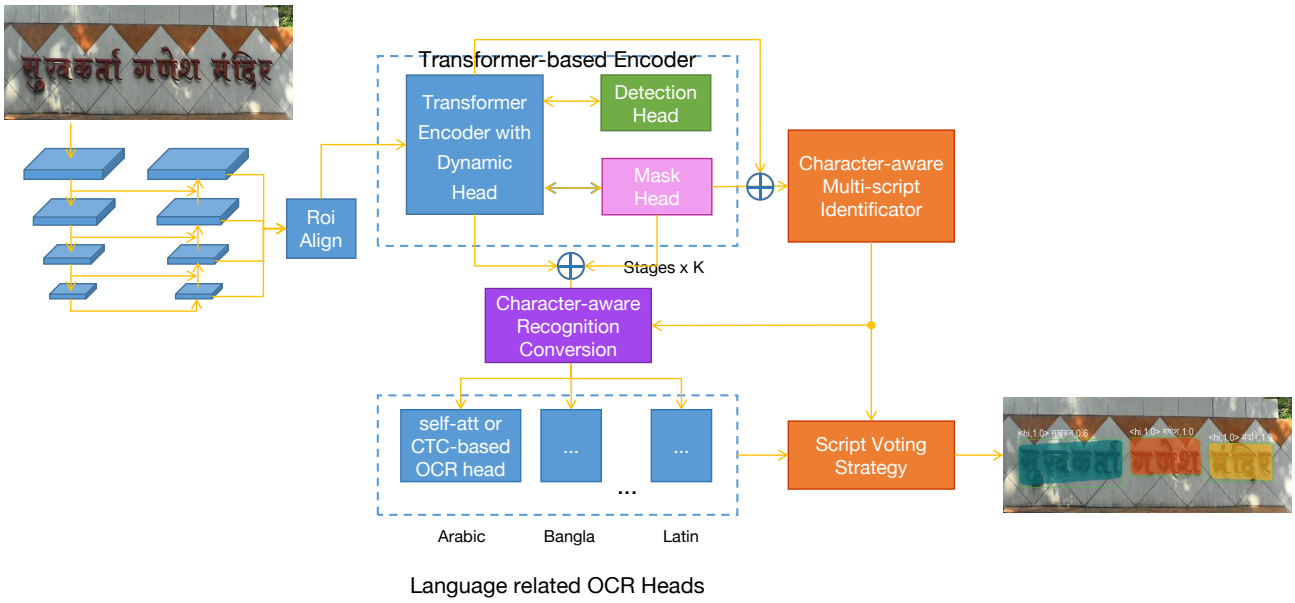

Figure 2. Network architecture. Our model framework consists of three functional components: a Transformer-based detector, a character-level multi-label script classifier, and several language-dependent OCR recognition heads. The character-aware recognition conversion module aims to bridge the gap between the three tasks of detection, identification, and recognition, enabling better collaboration between multiple tasks.

In recent years, some approaches discuss script identification together with text detection and recognition. [3, 5] propose a framework for end-to-end multilingual spotting. These methods do not take script identification into account in the process, only training a generic recognizer for common character sets and then voting on transcriptions to determine the language. However, this practice increases the amount of similar characters in the character set, thereby hampering model performance, such as '1' and 'I'. [10] early introduced script identification into the document analysis pipeline. Later, [13] demonstrated that dividing script subsets in multilingual setting helps to improve the performance of text recognition models.

3. Methodology

3.1. Overview

The proposed CaMiTextSpotter is based on SwinTextSpotter [14]. To reduce the long-tail effect, we borrow the idea from MultiplexedOCR [13], which introduces script identification step to convert general recognition into script-related recognition.

This practice raises the following problems: 1) How to assign the sets of common characters to script-related character sets? 2) Since the script-dependent character set as a prior, how to handle mixed samples across character sets? 3) For an end-to-end trainable framework, how to bridge the gap between the three tasks of detection, script identification, and sequence-to-sequence(Seq2Seq) recognition?

For the first question, we follow the practice of [13] i.e. collecting characters labeled as the same script in MLT17 [25] and MLT19 [24] datasets and a rationality analysis is provided for this practice, which can be seen in Experiment 4.3. For the second problem, we find that there exist some characters that are being included in more than one character set at the same time, such as many characters in Chinese, Japanese, and Korean languages(CJK). There are also cases where a single line of samples is mixed with multiple languages. Therefore, we formalize script identification as a sequence-to-sequence, multi-label classification task, which enables a more fine-grained processing of one line of text. Multilingual mixed samples can be solved by integrating the prediction results of multiple OCR heads at the character level.

3.2. Character-aware Multi-script Identification

In order to reduce the errors introduced by multilingual mixing and characters belonging to multiple character sets, we propose the Character-aware Multi-script Identification(CMI) module. It converts the whole-line global text prediction into character-level, multi-label classification-based prediction, The traditional classification task is formalized as a sequence-to-sequence, multi-label prediction task. This practice has the following advantages: 1) elimi-

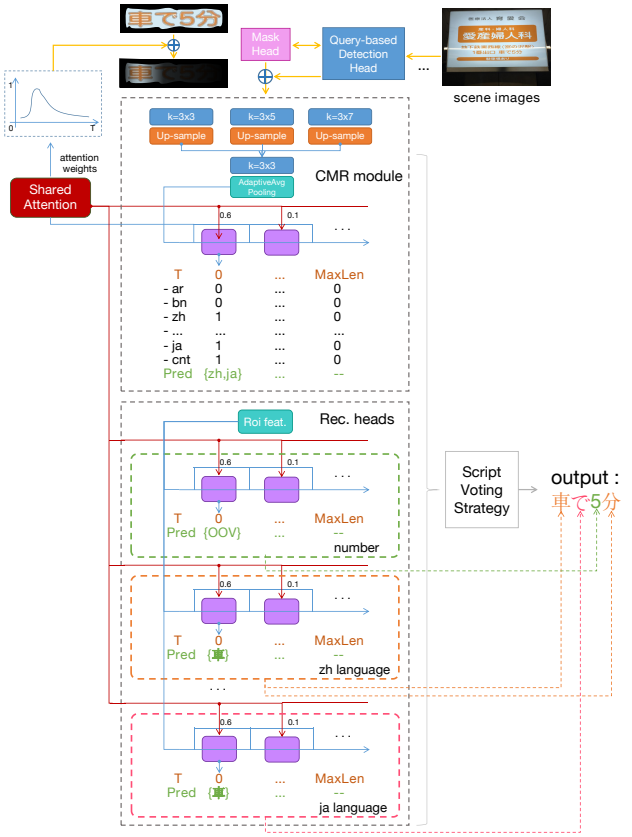

Figure 3. Character-aware Multi-script Recognizer. Based on the results of CMI module, we can re-organize the output at character level, thus solving the mixed script problem. By reusing the attention mechanism in all the Seq2Seq module, we can ensure that the replacement process is more credible.

nates the need for manual language annotation, meanwhile, reduces the ambiguity of line samples annotated with a single language. 2) formally unifies language classification and text recognition, providing more information to act on an end-to-end multilingual recognition system. 3) by reusing the attention vectors of script identification and text recognition, it allows the results of multiple OCR heads' predictions to be formally aligned, thus designing more effective character-level processing strategies to solve the problem of mixed samples.

The details of the CMI module can be seen in Figure 3. After a set of different sizes of convolutions capturing the long-range dependencies, it receives the Roi features obtained from the query-based detector. Subsequently, the features of the three pathways are aligned and then summed up, and after that two layers of convolution and one layer of global average pooling are used to obtain the sequence features. Finally, the Seq2Seq mapping is implemented with reference to ASTER [31]. The difference is that the Seq2Seq model's output is normalized independently instead of Softmax, and the multi-hot prediction is used to indicate whether the current character belongs to the corresponding character set. Considering that the model learning a sparse matrix will increase the optimization difficulty, we add extra constraint terms in the optimization of the CMI module, see 3.4 for details.

Multi-script label generation. Let $G \in \mathbb{R}^{T\times(C+1)}$ denote the label matrix for training CMI module, where T denotes the maximum length of the sequence and C denotes the number of scripts. The added one on the second dimension of the G-matrix indicates characters counting numbers. When 'text' indicates that the text transcription, the $text\,[i]$ means i-th character in it. The function $len\,(text)_i$ is used to count the number of characters in 'text'. We define the set of characters for i-th scripts as $\sigma_i, i \in \{1, C\}$. The function $\phi\,(ch)_i$ indicates whether the character ch belongs to character set σ_i. The G-matrix is generated as follows:

$$\phi\,(ch)_i = \begin{cases} 0,\ if\ ch \notin \sigma_i \\ 1,\ if\ ch \in \sigma_i \end{cases} ,where\ i \in \{1, C\} \quad (1)$$

$$G\,[t, c] = \phi\,(text[t])_c\,,\ t \in \{1, T\}\,,\ c \in \{1, C\} \quad (2)$$

$$G\,[t, c = C+1] = \begin{cases} 0,\ if\ t > len(text) \\ 1,\ if\ t <= len(text) \end{cases} \quad (3)$$

3.3. Bridge the Gap between Identification and Recognition

To better coordinate the identification and recognition, a shared sequence attention mechanism and script voting strategy are designed. In this section, we explain how we use the character-level multi-script feature for further tasks.

Shared Sequence Attention Mechanism.

Algorithm 1 Script voting strategy

Output: $script\_id, ocr$
Input: $rec, score, multi\_label, length, thresh$
  $script\_id \leftarrow argmax(multi\_label.sum(dim = 1))$
  $ocr \leftarrow rec.get(script\_id)$
  $ocr\_score \leftarrow score.get(script\_id)$
  for $(t = 0; t < length; t + 1)$ **do**
    $conf \leftarrow ocr\_score.get(t)$
    if $conf < thresh$ **then**
      $candi\_score \leftarrow score[:, t]$
      $candi\_char \leftarrow rec[:, t]$
      $mask_1 \leftarrow candi\_score > conf$
      $mask_2 \leftarrow multi\_label[:, t]$
      $mask \leftarrow mask_1 \cdot mask_2$
      $replace\_id \leftarrow argmax(mask \cdot candi\_score)$
      $ocr[t] \leftarrow candi\_char[replace\_id])$
    end if
  end for

Script voting strategy. Converting script identification from global prediction to character-level, multi-label prediction enables the introduction of stronger, lower-noise supervised signals. Besides, this fine-grained prediction can also better guide the results of downstream model. Since the process of dividing the multilingual character sets is done before the model training, the problem cannot be solved by any single recognition head alone when there is a sample that contain characters from multiple character sets, i.e. the language mixed problem. Using the fine-grained information obtained from the CMI module help to integrate the multilingual recognition results in the post-processing steps. Therefore, we unify the attention unit for script identification and OCR recognition, it makes the proposed *Script Voting Strategy Algorithm* more credible, which can be seen in Algorithm 1. The purpose of the algorithm is to integrate the output of multiple OCR heads using the results of multi-label identification, where rec denotes the character list, score denotes the confidence map, and multi-label is a bitmap, and their matrix dimensions are all [N, T]. N denotes the number of languages, T denotes the maximum length of the sequence.

3.4. Optimization

The optimization of the proposed model consists of three parts, which are the loss function L_{det} for the query-based detection, L_{script} for the script identification module which contains two constraints, and L_{ocr} for all the OCR branches.

Overall loss. Character-level multi-label constraint. As described in section 3.2, the label matrix $G \in \mathbb{R}^{T\times(C+1)}$ is a sparse matrix. Compared to each time step going through log-softmax, the number constraint between categories is lost. The multi-hot prediction also fails to represent the start and end of sequence, and thus also loses the length constraint. To solve this problem, we first add one row to the second dimension of the G matrix to indicate the character count, and when the count bit becomes zero indicates the termination of the sequence prediction. In addition, we add the following two loss function constraint terms:

$$L_{rela\_seq} = \sum_{c=1}^{C+1} L_{l2}\left(\theta\left(pred[:,c]\right),\ \theta\left(G[:,c]\right)\right) \quad (4)$$

$$L_{rela\_char} = \sum_{t=1}^{T} L_{l2}\left(\theta\left(pred[t,:]\right),\ \theta\left(G[t,:]\right)\right) \quad (5)$$

where $\theta\left(\right)$ denotes the summation function, L_{rela_seq} indicates that the number of characters contained in each script should be consistent with Ground-Truth, and L_{rela_char} indicates that the total number of scripts that contain a character should be consistent with Ground-Truth. The two constraint terms enable the association of multiple independent binary classification tasks to sequence prediction.

$$L_{det} = \lambda_1 \times L_{cls} + \lambda_2 \times L_{l1} + \lambda_3 \times L_{giou} + \lambda_4 \times L_{mask} \quad (6)$$

$$L_{script} = \lambda_5 \times L_{rela\_seq} + \lambda_6 \times L_{rela\_char} + \lambda_7 \times L_{bce} \quad (7)$$

$$L_{ocr} = \lambda_8 \times \sum_{c=1}^{C} -\frac{1}{T}\sum_{t=1}^{T} \log p_c\left(y_t\right) \quad (8)$$

where λ is the hyper-parameter used to balance the loss. In the process of query-based detector optimization, L_{cls} is the focal loss [20], L_{l1} and L_{giou} denotes L_1 loss and generalized IoU loss [26] for regressing the bounding boxes, L_{mask} denotes dice loss [23] for segmenting text regions. L_{script} is used to optimize the CMI module, where L_{rela_seq} and L_{rela_char} are sequence-related constraint terms and L_{bce} is used to learn whether the character ch belongs to the character set σ_c. Finally, L_{ocr} denotes the loss summation of all OCR heads, where p_c denotes the sequence probability of character set σ_c.

4. Experiments

4.1. Datasets

English datasets. *ICDAR 2013 dataset (IC13)* [16] contains 229 images for training and 233 images for testing. It contains primarily on scene text with word-level, horizontal rectangular bounding boxes annotation. *ICDAR 2015 dataset (IC15)* [15] consists of high-resolution images, 1000 for training and 500 for testing. It contains multi-oriented scene text annotated at word-level using quadrangle bounding boxes. *TotalText* [8] has 1255 images for training and 300 images for testing. It contains mainly curved scene text, annotated with polygon points. All the above datasets are focusing primarily on scene text in English.

Bi-lingual (English and Chinese) datasets. *LSVT19* [34] is one of the largest street view OCR datasets currently available. It provides a total of 450,000 images containing Chinese and English scene texts. Among them, 30,000 labeled images are provided as the training set and 20,000 labeled images are provided as the test set. The rest of the data are only provided with weak annotation and are not used in this research. *RCTW17* [32] is a competition on reading Chinese Text in images. It consists of various kinds of images, including street views, posters, menus, indoor scenes, and screenshots. Among them, 8034 images are provided as train images and 4229 images are provided as test images. *ArT19* [9] is a combination of *Total-Text* [8], *SCUT-CTW1500* [40] and *Baidu Curved Scene Text*. Almost a quarter of the text instances in the dataset are arbitrary-shaped and annotated with polygon points. It contains 5603 training and 4563 test images. All the above datasets are focusing primarily on scene text in Chinese and English.

Table 1. Analysis of MLT datasets

	mlt19val	mlt19train	mlt17val	mlt17val
Mo	30.70%	30.75%	28.76%	27.83%
Multi_creole	58.82%	58.34%	60.13%	61.25%
Multi_mixed	10.48%	10.91%	11.11%	10.92%

Table 2. Quantitative recognition results on different types of instances

Method	Mo	Multi_creole	Multi_mixed
Multiplexed	55.89	62.89	44.26
CaMiTextSpotter	56.64	63.21	49.56

Table 3. Ablation study on mlt test data. CL_cls refers to global classification, Seq_cls refers to sequence classification, Sh_att refers to shared attention for identification and recognition. ✓refers use voting strategy for sequence identification. ×refers not using voting strategy.

Method	task 3			task 4		
	F	P	R	F	P	R
Baseline	-	-	-	47.90	60.68	39.57
+ Gl_cls	70.64	87.95	59.03	50.67	62.86	42.44
+ Sh_att + Seq_cls	70.70	88.09	59.04	51.46	63.55	43.23

Table 4. MLT19 task1.

Method	F	P	R	AP
PSENet*	65.83	73.52	59.59	52.73
RRPN*	69.56	77.71	62.95	58.07
CRAFTS*	70.86	81.42	62.73	56.63
CRAFTS	75.50	81.70	**70.10**	-
MaskTextSpotterV3	71.10	83.75	61.76	58.76
Multiplexed	72.66	85.53	63.16	60.46
SwinTextSpotter	71.67	**91.04**	59.09	57.66
CaMiTextSpotter	**75.65**	86.43	67.26	**65.19**

Multi-language dataset. *MLT17* [25] and *MLT19* [24] are the largest multilingual scene text datasets currently available. *MLT17* contains seven scripts in Arabic, Latin, Chinese, Japanese, Korean, Bangla and Symbols. Compared to *MLT17*, *MLT19* adds Hindi scripts and also releases the multilingual synthetic dataset *SynthTextMLT*. *MLT17* contains 7200 training images, 1800 validation images and 9000 test images. *MLT19* contains 10000 training images, 2000 validation images and 10000 test images. *SynthTextMLT* provides ∼273k synthetic data.

4.2. Implementation details

We follow the training strategy, character sets partition, and the OCR heads settings in [13]. First of all, we only train the detection branch of our framework with reference to the practice of [14]. Secondly, we train the model end-to-end on datasets (*MLT* and *SynthTextMLT*) for 260K iterations. The training batch size is 12 with learning rate 1×10^{-4}, which reduces to 1×10^{-5} at 140Kth iteration and 1×10^{-6} at 220Kth iteration. Then we removed the synthetic data and mixed the data from multiple scenes (*TotalText*, *ArT19*, *RCTW17*, *LSVT19*, *IC15*, *IC13*) for mixed training 220K iterations. The training batch size is 12 with learning rate 1×10^{-5}, which reduces to 1×10^{-6} at 160Kth iteration. Finally, model fine-tuned on MLT data (*MLT17* or *MLT19*) for 120k iterations with learning rate 1×10^{-5}, which reduces to 1×10^{-6} at 80Kth iteration.

The whole process is script annotation free, and we find that there is no significant performance difference between training all OCR heads end-to-end and selectively training some heads.

Table 5. MLT17 task3

Method	F	P	R	AP
E2E-MLT	58.69	64.61	53.77	-
CRAFTS	68.31	74.52	63.06	54.56
Multiplexed	69.41	81.81	60.27	56.30
CaMiTextSpotter	69.79	87.67	57.97	55.07

Table 6. MLT19 task3

Method	F	P	R	AP
CRAFTS	68.34	78.52	60.50	53.75
MaskTextSpotterV3	65.19	75.41	57.41	51.98
Multiplexed	69.42	81.72	60.34	56.46
CaMiTextSpotter	70.70	88.09	59.04	56.30

Table 7. MLT19 task4

Method	F	P	R	AP	1-NED
E2E-MLT	26.5	37.4	20.5	7.7	26.4
RRPN+CLTDR	33.8	38.6	30.1	11.6	38.3
CRAFTS	51.7	65.7	42.7	**34.9**	48.3
MaskTextSpotterV3	39.7	**71.8**	27.4	-	-
Multiplexed	48.2	68.0	37.3	-	-
SwinTextSpotter	47.90	60.68	39.57	29.90	48.97
CaMiTextSpotter	**51.46**	63.55	**43.23**	33.13	**52.37**

4.3. Ablation study

Multi-label ambiguity. In MLT19 [24], the author divides ten languages into seven scripts according to the linguistic prior. It adds "symbol" and "mix" classes to further describe instances such as +/(without any other alphabet characters of the language) and others including characters belonging to difficult scripts. However, it labels the data according to the prior knowledge of where the image took. There are many ambiguities and multi-label problems exist. In order to confirm our methods is beneficial for end-

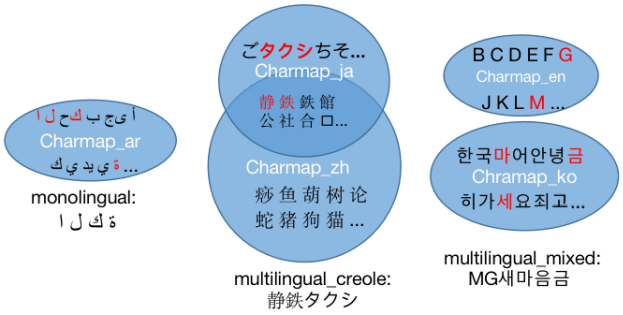

Figure 4. Description of various categories of text

to-end multilingual scene text spotting. We deliver experiments to analyze the dataset, and divide the original data into new classes by identifying how many characters set its belonging to which aims to adapt to multilingual text recognition in real scenarios. Show in Figure 4, we classify instances into there types according to which kind of character sets it belongings to. The character sets are obtained from the annotations, and in order to reduce multi-label problems of every single character which may cause classification or recognition errors, we build the character sets to eliminate co-occurrence problems by reducing the same characters appearing in the multi character sets.

We divide the instances into three types: 1) monolingual (Mo): an instance that only has characters in one character set. 2)multilingual_creole (Multi_creole): the characters in the instance can be found in multiple character sets, but every character can be covered by one set. 3) multilingual_mixed (Multi_mixed): the characters in the instance can be found in multiple character sets, and it can not be covered by one character set. The analysis of the data set is as Table 1. In reality, mixed data is far more than we can think. And for recent methods, it is not trivial to learn this data, especially in the multilingual_mixed case. Benefit from our model for character set classification and recognition, our methods can solve this problem better compared to previous methods, the results are shown in Table 2.

Character-aware Multi-script Recognition.

4.4. Compared with SOTA

5. Conclusion

References

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019. 2

[2] Youngmin Baek, Seung Shin, Jeonghun Baek, Sungrae Park, Junyeop Lee, Daehyun Nam, and Hwalsuk Lee. Character region attention for text spotting. In *European Conference on Computer Vision*, pages 504–521. Springer, 2020. 1

[3] Youngmin Baek, Seung Shin, Jeonghun Baek, Sungrae Park, Junyeop Lee, Daehyun Nam, and Hwalsuk Lee. Character region attention for text spotting. In *European Conference on Computer Vision*, pages 504–521. Springer, 2020. 3

[4] Ankan Kumar Bhunia, Aishik Konwer, Ayan Kumar Bhunia, Abir Bhowmick, Partha P Roy, and Umapada Pal. Script identification in natural scene image and video frames using an attention based convolutional-lstm network. *Pattern Recognition*, 85:172–184, 2019. 2

[5] Michal Bušta, Yash Patel, and Jiri Matas. E2e-mlt-an unconstrained end-to-end method for multi-language scene text. In *Asian conference on computer vision*, pages 127–143. Springer, 2018. 3

[6] Changxu Cheng, Qiuhui Huang, Xiang Bai, Bin Feng, and Wenyu Liu. Patch aggregator for scene text script identification. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1077–1083. IEEE, 2019. 2

[7] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pages 5076–5084, 2017. 2

[8] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017. 5

[9] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 5

[10] Yasuhisa Fujii, Karel Driesen, Jonathan Baccash, Ash Hurst, and Ashok C Popat. Sequence-to-label script identification for multilingual ocr. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 161–168. IEEE, 2017. 3

[11] Lluis Gomez, Anguelos Nicolaou, and Dimosthenis Karatzas. Improving patch-based scene text script identification with ensembles of conjoined networks. *Pattern Recognition*, 67:85–96, 2017. 2

[12] Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, and Errui Ding. Wordsup: Exploiting word annotations for character based text detection. In *Proceedings of the IEEE international conference on computer vision*, pages 4940–4949, 2017. 2

[13] Jing Huang, Guan Pang, Rama Kovvuri, Mandy Toh, Kevin J Liang, Praveen Krishnan, Xi Yin, and Tal Hassner. A multiplexed network for end-to-end, multilingual ocr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4547–4557, 2021. 1, 3, 6

[14] Mingxin Huang, YuLiang liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. Swintextspotter: Scene text spotting via better

synergy between text detection and text recognition. *arXiv preprint arXiv:2203.10209*, 2022. 2, 3, 6

[15] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 5

[16] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493. IEEE, 2013. 5

[17] Prateek Keserwani, Kanjar De, Partha Pratim Roy, and Umapada Pal. Zero shot learning based script identification in the wild. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 987–992. IEEE, 2019. 2

[18] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018. 2

[19] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11474–11481, 2020. 2

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[21] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018. 2

[22] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019. 2

[23] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[24] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1582–1587. IEEE, 2019. 1, 3, 6

[25] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1454–1459. IEEE, 2017. 3, 6

[26] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[27] Gregory Sell, David Etter, Daniel Garcia-Romero, and Alan McCree. Script identification using across-and within-image distribution estimation. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1084–1089. IEEE, 2019. 2

[28] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2550–2558, 2017. 2

[29] Baoguang Shi, Xiang Bai, and Cong 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*, 39(11):2298–2304, 2016. 2

[30] Baoguang Shi, Xiang Bai, and Cong Yao. Script identification in the wild via discriminative convolutional neural network. *Pattern Recognition*, 52:448–458, 2016. 2

[31] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 2, 4

[32] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th iapr international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1429–1434. IEEE, 2017. 5

[33] Baoguang Shi, Cong Yao, Chengquan Zhang, Xiaowei Guo, Feiyue Huang, and Xiang Bai. Automatic script identification in the wild. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 531–535. IEEE, 2015. 2

[34] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019. 5

[35] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision*, pages 56–72. Springer, 2016. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[37] Zhaoyi Wan, Fengming Xie, Yibo Liu, Xiang Bai, and Cong Yao. 2d-ctc for scene text recognition. *arXiv preprint arXiv:1907.09705*, 2019. 2

[38] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2019. 2

[39] Yongchao Xu, Yukang Wang, Wei Zhou, Yongpan Wang, Zhibo Yang, and Xiang Bai. Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing*, 28(11):5566–5579, 2019. 2

[40] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017. 5

[41] Jan Zdenek and Hideki Nakayama. Bag of local convolutional triplets for script identification in scene text. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 369–375. IEEE, 2017. 2

[42] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9699–9708, 2020. 2

[43] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 2

[44] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3131, 2021. 2