OPTIMAL KERNEL FOR REAL-TIME ARBITRARY-SHAPED TEXT DETECTION

Haozhao Ma[1,2], *Chuang Yang*[2], *Yuan Yuan*[2], *Qi Wang*[2*]

[1]School of Software and [2]School of Artificial Intelligence, OPtics and ElectroNics (iOPEN),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China

ABSTRACT

Recently, segmentation-based text detection methods develop rapidly, which achieve competitive accuracy and detection speed. However, these methods are hard to fit text instances accurately, which leads to the decrease of model performance. Meanwhile, the poor perception of the text center by the boundary pixels further affects the detection accuracy. We follow the issues and design an efficient framework for arbitrary-shaped text detection, which is constructed based on Optimal Kernel Representation (OKR) and Pixel Enhancement Module (PEM). Specifically, OKR is proposed to fit texts with optimal kernels. It erodes texts according to the corresponding geometric characteristics, which is simpler and more accurate compared with previous methods. PEM is used to enhance the perception of boundary pixels to the virtual character centers of text, thus improving the cohesion of the whole instance. Particularly, PEM only participates in the training process, which brings no extra computation costs to inference. Ablation experiments show the effectiveness of OKR and PEM. Comparisons on serveral benchmarks verify that our efficient detector is superior to the existing state-of-the-art (SOTA) methods.

Index Terms— Efficient text detector, optimal kernel, pixel enhancement

1. INTRODUCTION

Scene Text Detection (STD)[1, 2, 3, 4, 5, 6] and Text Spotting System (TSS)[7] are playing an increasingly important role in our daily life, due to the applications such as scene perception, video content understanding and so on. Scene Text Detection is particularly important as an upstream task for Optical Character Recognition system.

Existing text detection methods include regression and segmentation-based methods roughly. The former is developed from object detectors, which bring competitive detection speed for them but result in their inability to detect arbitrary-shaped texts. In order to detect arbitrary-shaped texts and avoid text adhesion, the latter proposes erosion-based text representation methods, but the complex postprocessing and low detection efficiency still limit their applications. Therefore, how to construct an efficient network to detect arbitrary-shaped texts with remarkable accuracy and speed remains to be explored.

Considering the above problems, Optimal Kernel Representaion (OKR) is proposed firstly to fit texts with any shapes, it erodes and dilates adaptively based on text geometric characteristics, which is simpler and more accurate compared with others. To enhance the perception of instance centers by boundary pixels, we introduce Pixel Enhancement Module (PEM) as a branch of the network, which imagines serveral virtual character centers of the text and drives pixels of text close to the corresponding center. The experimental results show that PEM is helpful in improving the cohesion of text instances. Importantly, PEM can be removed from the network during inference process. Based on OKR and PEM, an arbitrary-shaped text detction network is constructed and results on multiple benchmarks verify that our method outperforms existing state-of-the-art (SOTA) methods. In summary, the contributions of this work are listed as follows:

- Optimal Kernel Representation (OKR) is introduced to represent text instances. It erodes and dilates adaptively according to the text geometric characteristics, which ensures strong fitting ability for texts with any shapes with low design complexity.
- Pixel Enhancement Module (PEM) is proposed to enhance the perception of instance centers by boundary pixels, which avoids the problem that adjacent instances interfere with the boundary pixels and improves the cohesion of the text instances.
- An accurate and efficient arbitrary-shaped text detection network is constructed, which can achieve high performance in both detection accuracy and inference speed. It provides a new solution for the practical applications of scene text detection.

2. RELATED WORK

Benefiting from the development of deep learning and the improvement of computing power, text detection technique

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902. *Qi Wang is the corresponding author.

has entered a new stage. They can be roughly divided into regression-based and segmentation-based methods.

Regression-based methods are usually developed from object detectors. RRD[8] constructed different heads to obtain rotation-sensitive features and rotation-insensitive features. CornerNet[9] utilized four corner regions to fit text instances and combined them to reconstruct text regions. Textray[10] used multiple contour points to represent the text contour, but it still can't fit highly curved text. However, it is difficult for them to reconstruct arbitrary-shaped text, which limits their applications. **Segmentation-based methods** perform a binary classification in pixel level. PixelLink[11] rebuilded text by means of connected domain analysis. TextSnake[12] decomposed the text instances into several small components, and then combined them to reconstruct text contours. PAN[13] introduced a new postprocessing method close to the clustering methods and achieved good detection results. DB[14] presented a pixel level task named threshold map to enhance the supervision during training stage which can be removed during inference process. OPMP[15] detected text instances through pyramid residual sequence model. However, few of them can meet the requirements of real-time text detection due to the complex backbone and text representation strategies.

3. METHODOLOGY

In this section, the architecture of the proposed framework is introduced firstly. Then, Optimal Kernel Representation (OKR) and Pixel Enhancement Module (PEM) are described in detail. In the end, optimization method is illustrated.

3.1. Overall Architecture

The architecture of our method is illustrated as Fig. 1 which contains three components: Backbone, Pixel Enhancement Module and Optimal Kernel Representation. Given an input image, we first use backbone to extract and fuse the features. For details of backbone, ResNet is adopted to extract the features and the features whose sizes are 1/4, 1/8, 1/16, 1/32 of original image are sent to FPN which can fuse high-dimensional features and low-dimensional features to get the feature pyramid. Then the output features of FPN are concated to get the feature map $F_{concat} \in \mathbb{R}^{(256,H/4,W/4)}$, where H and W are the image height and width respectively. After that, we use a convolution layer to futher fuse features and get features $F_{Conv} \in \mathbb{R}^{(64,H/4,W/4)}$ which pass through different deconvolution layers to predict pixel enhancement features $F_{PEM} \in \mathbb{R}^{(3,H,W)}$ and text kernel $F_{kernel} \in \mathbb{R}^{(1,H,W)}$. Pixel Enhancement Module is only available during training stage to improve cohesion of pixels inside text and it will be removed from the inference process. Optimal Kernel Representation module can consider the size of each predicted text

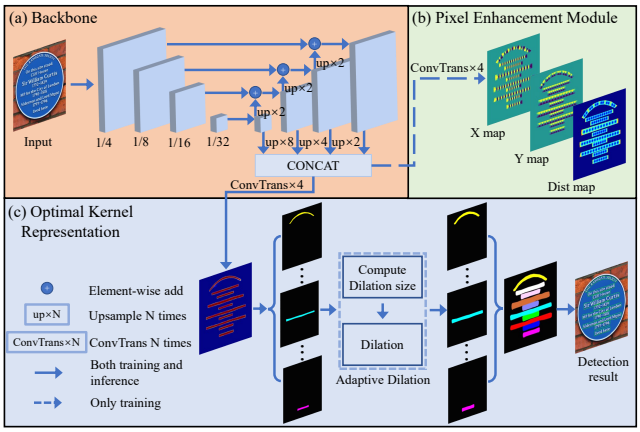


Fig. 1. Overall architecture of our method. Pixel Enhancement Module utilizes distance and direction to improve text cohesion. Optimal Kernel Representation extracts text contours by computing dilation size of each text kernel and dilating each kernel respectively.

kernel and obtain the optimal dilation size for it to reconstruct the original text instance.

3.2. Optimal Kernel Representation

To fit texts with any shapes accurately with simple model, we propose Optimal Kernel Representation (OKR) strategy. Most segmentation-based methods adopt Vatti clipping algorithm, but it is difficult to restore each text instance accurately. Some methods also rebuild text instances by using erosion and dilation kernels with a fixed size, which leads to small texts being missed when the size is too large or adjacent texts can't be separated effectively when the size is too small (as shown in Fig. 2 (FE and FD)).

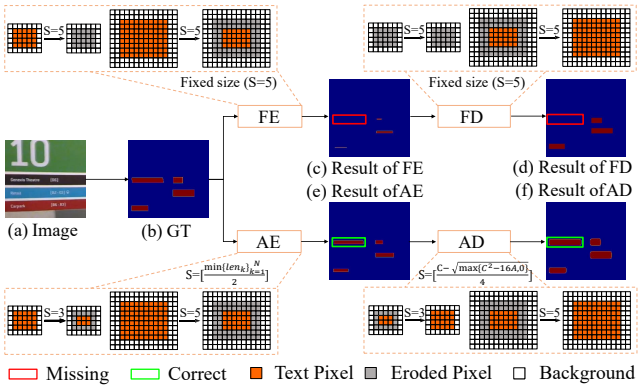


Fig. 2. Illustration of Optimal Kernel Representation strategy. "FE", "FD", "AE", and "AD" represent "Fixed Erosion", "Fixed Dilation", "Adaptive Erosion", and "Adaptive Dilation", respectively.

OKR consists of Adaptive Erosion and Adaptive Dilation processes. They are responsible for generating text kernels and rebuilding texts from kernels respectively. Adaptive Erosion erodes text instances adaptively according to the corresponding geometric characteristics, where the erosion size can be computed by:

$$S = [\frac{\min\{len_k\}_{k=1}^{N}}{2}] \,, \tag{1}$$

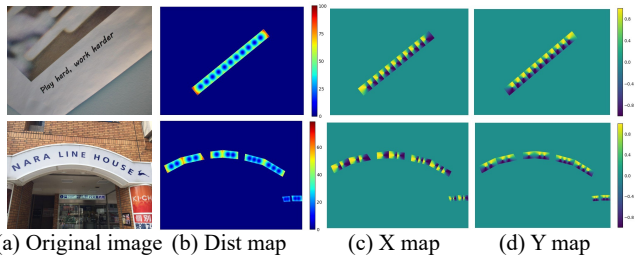
where len_k represents the length of the kth side of the text polygon. $[\cdot]$ is the odd function. Adaptive Dilation adopts different strategies to obtain the dilation size for rebuilding texts. For quadrilateral text, it performs the minimum area rotated rectangle fitting on the predicted binary map, and then takes the shortest side of the rectangle as dilation size. For arbitrary-shaped text, the dilation size can be computed as:

$$S = [\frac{C - \sqrt{\max\{C^2 - 16A, 0\}}}{4}] \,, \tag{2}$$

where C and A respresent the perimeter and area of each irregular text polygon respectively.

3.3. Pixel Enhancement Module

Since the network ability to perceive texts decreases from the center to the boundary, and the adjacent instances interfere with the boundary pixels greatly, both of which bring great obstacles to detection. Inspired by CRAFT[16], Pixel Enhancement Module (PEM) is presented to improve the perception of the text boundary pixels to the center of instance. PEM proposes the concept of virtual character centers, assigns a nearest virtual character center for each pixel in text and drives each pixel to converge to the corresponding center.



(a) Original image (b) Dist map (c) X map (d) Y map

Fig. 3. Structure of virtual character centers. The first row shows virtual character centers of quadrilateral and the second shows the example of polygon.

Virtual character centers consist of three parts: Dist map, X map and Y map. The structure diagram is illustrated as Fig. 3. Dist map respresents the euclidean distance from the pixel to the nearest virtual center, and a two-dimensional vector is introduced to represent the direction of the pixel relative to its virtual center as X and Y maps.

For quadrilaterals, the number of centers is computed as:

$$N(Q) = \max\{\frac{\max\{len_k\}_{k=1}^{4}}{\min\{len_k\}_{k=1}^{4}}, 1\} \,, \tag{3}$$

where Q means quadrilateral and len_k means the length of the kth side of the quadrilateral. Then the virtual character centers are uniformly sampled along the central axis.

For the polygon, it is decomposed into several quadrilaterals along axis firstly, and then sampling centers by Equ. 3.

3.4. Optimization

The proposed method is a multi-task network, including kernel segmentation and virtual character centers prediction. The loss function L used to train network is formulated as:

$$L = L_{ks} + \lambda L_{vcc} \,, \tag{4}$$

where λ is set to 0.05 to balance the loss function.

A weighted binary cross-entropy loss is adopted for L_{ks} which can be computed as:

$$L_{ks} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathrm{w}(i)[y_i \log(x_i) + (1 - y_i)\log(1 - x_i)] \,, \tag{5}$$

where Ω contains all positives and part of negatives due to the unbalance of positives and negatives. It is responsible for sampling the positives and negatives with the ratio $1:3$. $\mathrm{w}: \Omega \rightarrow \mathbb{R}$ is a weight map, which can be computed as:

$$\mathrm{w}(i) = w_0 \cdot \exp(-\frac{(\mathrm{d}_1(i) + \mathrm{d}_2(i))^2}{2\sigma^2}) + 1 \,, \tag{6}$$

where $\mathrm{d}_1(i)$ means the distance between pixel i and the nearest text boundary, $\mathrm{d}_2(i)$ means the distance to the second nearest text boundary. w_0 and σ are hyper-parameters set to 30 and 10 respectively. The formula of L_{vcc} is defined as:

$$L_{vcc} = \log\frac{\max(Dist, \widehat{Dist})}{\min(Dist, \widehat{Dist})} + |Dir - \widehat{Dir}| \,, \tag{7}$$

where $Dist$ and Dir represent the ground-truth for distance and direction, $\widehat{Dist}$ and $\widehat{Dir}$ represent predicted distance and direction from network, respectively.

4. EXPERIMENTS

We do experiments on multiple benchmarks to show the robustness and superiority of our approach. **SynthText**[21] is used to pretrain our model. **ICDAR2015**[17] is a dataset with complex background. It has training and testing images of 1000 and 500, and all text instances are annotated as multi-oriented quadrilaterals. **TotalText**[18] is an irregular (horizontal, curved and multi-oriented) text detection dataset with 1255 images for training and 300 images for testing. **MSRA-TD500**[19] is a multi-language dataset with 700 and 200 images for training and testing, of which 400 training images are introduced from **HUST-TR400**[20].

Table 1. Ablation study for OKR and PEM. "Baseline" adopts the fixed size (S : 9) to erode and reconstruct text instances. The image scale during testing is 736 pixels.

#	Baseline	OKR	PEM	R	P	F	Gain	FPS
1	✓			77.2	88.2	82.4		40.8
2	✓	✓		81.1	86.3	83.6	[+1.2]	37.8
3	✓	✓	✓	80.1	88.3	84.0	[+0.4]	37.8

Table 2. Comparison results with other methods on three datasets, and the values in parentheses indicate the height of image during inference stage.

Methods	Ext.	Recall	Precious	F-measure	FPS
		ICDAR2015			
BiP-Net[1]	✓	82.1	86.9	83.9	24.8
ESTD[2]	×	77.8	85.8	81.5	-
PAN[13]	✓	81.9	84.0	82.9	26.1
TransTT[26]	✓	78.3	89.8	83.7	10
KPN[6]	✓	83.2	84.1	83.6	12.2
Ours (736)	✓	80.1	88.3	**84.0**	**37.8**
		TotalText			
NASK[3]	✓	81.2	83.3	82.2	8.4
CSE[24]	✓	79.1	81.4	80.2	-
PCR[25]	×	80.2	86.1	83.1	-
RFRN[22]	✓	80.9	82.9	81.9	10.8
RSCA[23]	✓	78.5	86.9	82.5	40.3
Ours (800)	✓	80.9	85.8	**83.3**	**40.5**
		MSRA-TD500			
SL[27]	✓	80.8	87.7	84.1	-
OPMP[15]	✓	83.4	86.0	84.7	1.6
SRM[28]	×	80.8	84.2	82.5	21.5
DB[14]	✓	79.2	91.5	84.9	32.0
KPR[5]	✓	85.7	80.6	83.1	18.2
Ours (736)	✓	79.9	90.5	**84.9**	**51.5**

4.1. Ablation Study

To analyze the effectiveness of Optimal Kernel Representation and Pixel Enhancement Module, ablation study is conducted on ICDAR2015 dataset and the detailed experimental results are shown in Table 1. **Optimal Kernel Representation**. We compare the effect of fixed-size erosion-dilation strategy with our OKR. Since the proposed OKR can avoid the disappearance of small instances during preprocessing period and better reconstruct text during postprocessing, the F-measure on ICDAR2015 dataset is improved by 1.2%. **Pixel Enhancement Module**. As shown in Table 1 #3, PEM brings another 0.4% performance improvement, which demonstrates the effectiveness of enhancing the perception of pixels to virtual character centers without introducing additional cost computations.

4.2. Comparisons with previous methods

To show the robustness and superiority of the proposed approach, comparisons with related state-of-the-art (SOTA) methods are performed on three datasets.

As can be seen from Table 2, our framework outperforms others a lot in both detection accuracy and detection efficiency on all three datasets. We achieve 84.0% of F-measure and 37.8 FPS in detection speed on ICDAR2015, which implies that our method is effective for blurry images and can detect multi-oriented text instances successfully. We also verify the detection ability for irregular text on Totaltext, which achieves the highest detection accuracy while running 4 times faster than RFRN[22]. The same conclusion can also be obtained on MSRA-TD500. Our method achieves a good balance in detection accuracy and detection speed.

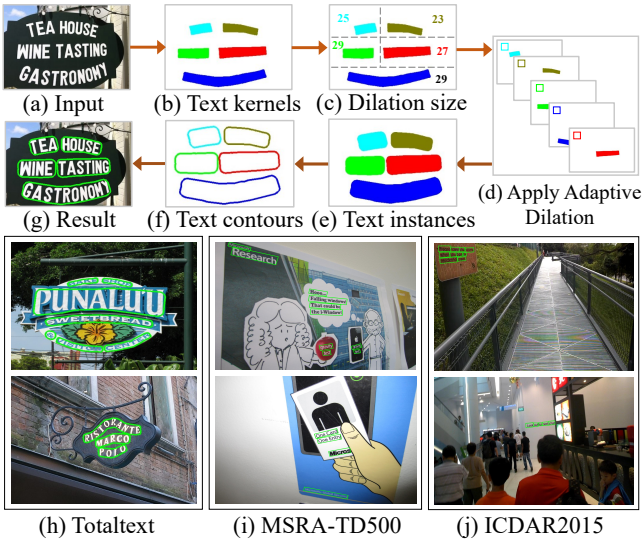


Fig. 4. Visualization of some qualitative detection results. □, □, □, □, □ represent dilation kernels corresponding to the text kernels.

We visualize the inference process in Fig. 4(a)–(g) and the numbers in Fig. 4(c) represent the dilation size calculated by corresponding text kernel according to Equ. 2. Fig. 4(h)–(j) depict quality results with different shapes, which demonstrate the superiority of the proposed framework.

5. CONCLUSION

In this paper, we construct an efficient framework for real-time arbitrary-shaped text detection based on Optimal Kernel Representation (OKR) and Pixel Enhancement Module (PEM). OKR is an efficient text representation strategy to reconstruct text from binary map, which avoids the problems of disappearance of erosion and inaccurate reconstruction. PEM is introduced as auxiliary branch to supervise the network during training stage, which proposes the concept of virtual character centers and promotes the pixels of text instances to approach the center points correspondingly without additional time consumption. With well-designed OKR and PEM, our method achieves remarkable performance in detecting accuracy and detection speed on multiple benchmarks.

6. REFERENCES

[1] C. Yang, M. Chen, Y. Yuan, and Q. Wang, "BiP-Net: Bidirectional Perspective Strategy Based Arbitrary-Shaped Text Detection Network," in *ICASSP*, 2022, pp. 2255-2259.

[2] L. Zhang, Y. Liu, H. Xiao, L. Yang, G. Zhu, S. Shah, M. Bennamoun, and P. Shen, "Efficient scene text detection with textual attention tower," in *ICASSP*, 2020, pp. 4272–4276.

[3] M. Cao and Y. Zou, "All you need is a second look: Towards tighter arbitrary shape text detection," in *ICASSP*, 2020, pp. 2228–2232.

[4] C. Yang, M. Chen, Z. Xiong, Y. Yuan, and Q. Wang, "Cm-net: Concentric mask based arbitrary-shaped text detection," *IEEE Trans. IP*, vol. 31, pp. 2864-2877, 2022.

[5] S. Qin, and C. Lin, "Arbitrary-shaped scene text detection with keypoint-based shape representation," *IJDAR*, vol. 25, no.2, pp. 115-127, 2022.

[6] S. Zhang, X. Zhu, J. Hou, C. Yang, and X. Yin, "Kernel Proposal Network for Arbitrary Shape Text Detection," *IEEE Trans. NNLS*, 2022.

[7] X. Zhang, Y. Su, S. Tripathi, and Z. Tu, "Text Spotting Transformers," in *CVPR*, 2022, pp. 9519-9528.

[8] M. Liao, Z. Zhu, B. Shi, G. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *CVPR*, 2018, pp. 5909-5918.

[9] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *ECCV*, 2018, pp. 734–750.

[10] F. Wang, Y. Chen, F. Wu, and X. Li, "Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection," in *ACMMM*, 2020, pp. 111–119.

[11] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *AAAI*, 2018.

[12] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *ECCV*, 2018, pp. 20-36.

[13] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *ICCV*, 2019, pp. 8440-8449.

[14] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *AAAI*, 2020, pp. 11474-11481.

[15] S. Zhang, Y. Liu, L. Jin, Z. Wei, and C. Shen, "Opmp: An omnidirectional pyramid mask proposal network for arbitrary-shape scene text detection," *IEEE Trans. MM*, vol. 23, pp. 454–467, 2020.

[16] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *CVPR*, 2019, pp. 9365-9374.

[17] D. Karatzas, L. Gomez, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. Chandrasekhar, and S. Lu, "Icdar 2015 competition on robust reading," in *ICDAR*, 2015, pp. 1156–1160.

[18] C. Ch'ng, and C. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *ICDAR*, 2017, pp. 935-942.

[19] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *CVPR*, 2012, pp. 1083–1090.

[20] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Trans. IP*, vol. 23, no. 11, pp. 4737–4749, 2014.

[21] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *CVPR*, 2016, pp. 2315–2324.

[22] G. Deng, Y. Ming, and J. Xue, "RFRN: A recurrent feature refinement network for accurate and efficient scene text detection," *Neurocomputing*, vol. 453, pp. 465-481, 2021.

[23] J. Li, Y. Lin, R. Liu, C. Ho, and H. Shi, "RSCA: Real-time Segmentation-based Context-Aware Scene Text Detection," in *CVPR*, 2021, pp. 2349-2358.

[24] Z. Liu, G. Lin, S. Yang, F. Liu, W. Lin, and W. Goh, "Towards robust curve text detection with conditional spatial expansion," in *CVPR*, 2019, pp. 7269-7278.

[25] P. Dai, S. Zhang, H. Zhang, and X. Cao, "Progressive contour regression for arbitrary-shape scene text detection," in *CVPR*, 2021, pp. 7393-7402.

[26] Z. Raisi, M. Naiel, G. Younes, S. Wardell, and J. Zelek, "Transformer-based text detection in the wild," in *CVPR*, 2021, pp. 3162-3171.

[27] W. Zhang, Y. Qiu, M. Liao, R. Zhang, X. Wei, and X. Bai, "Scene text detection with scribble line," in *ICDAR*, 2021, pp. 79-94.

[28] M. Liang, J. Hou, X. Zhu, C. Yang, J. Qin, and X. Yin, "Multi-orientation scene text detection with scale-guided regression," *Neurocomputing*, vol. 461, pp. 310-318, 2021.